Design and analysis of environmental monitoring programs

Søren Lophaven

Kongens Lyngby 2004
IMM-PHD-2004-138

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

This thesis describes statistical methods for modelling space-time phenomena. The methods were applied to data from the Danish marine monitoring program in the Kattegat, measured in the five-year period 1993-1997. The proposed model approaches are characterised as relatively simple methods, which can handle missing data values and utilize the spatial and temporal correlation in data. Modelling results can be used to improve reporting on the state of the marine environment in the Kattegat.

The thesis also focus on design of monitoring networks, from which geostatistics can be successfully applied. Existing design methods are reviewed, and based on these a new Bayesian geostatistical design approach is suggested. This focus on constructing monitoring networks which are efficient for computing spatial predictions, while taking the uncertainties of the parameters in the geostatistical model into account. Thus, it serves as a compromise between existing methods.

The space-time model approaches and geostatistical design methods used in this thesis are generally applicable, i.e. with minor modifications they could equally well be applied within areas such as soil and air pollution.

Resumé

Denne PhD afhandling beskriver statistiske metoder til modellering af fænomener i tid og rum. Metoderne er anvendt på data fra det danske marine overvåningsprogram i Kattegat, der er målt i perioden 1993-1997. De foreslåede modeller er karakteriseret ved at være forholdsvis simple metoder, der kan håndtere manglende observationer, samt udnytte den spatielle og tidslige korrelation i data. Model resultaterne kan anvendes til at forbedre viden og afrapportering af miljøtilstanden i Kattegat.

PhD afhandlingen omhandler ligeledes design af måleprogrammer, således at effektiv modellering ved brug af geostatistik muliggøres. Der gives en oversigt over eksisterende design metoder, og på baggrund af disse er foreslået en ny Bayesiansk design metode. Denne fokuserer på at konstruere måleprogrammer, der kan anvendes til effektiv beregning af spatielle prediktioner ved brug af en geostatistisk model, og under hensyntagen til usikkerhederne i modellens parametre. Den Bayesiansk design metode kombinerer således eksisterende design metoder.

De statistiske metoder til modellering af fænomener i tid og rum, samt de geostatistiske design metoder, der anvendes i PhD afhandlingen, er generelle metoder, og kan med små ændringer anvendes indenfor andre miljøområder, som f.eks. jord- og luftforurening.

Preface

This thesis was prepared at Informatics and Mathematical Modelling (IMM), the Technical University of Denmark (DTU) in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The thesis deals with different aspects of mathematical modelling of space-time phenomena and geostatistical design of monitoring networks using a dataset containing measurements of nutrients, biomass, salinity and temperature in the Kattegat.

The thesis consists of a summary report and a collection of six research papers written during the period 2001–2004, and elsewhere published.

Kongens Lyngby, October 2004

Søren Lophaven

Papers included in the thesis

Appendix A Lophaven, S., Carstensen, J., and Rootzén, H. (2002). Methods for estimating the semivariogram. In *Symposium i Anvendt Statistik*, p. 128-144. Århus, Denmark, January 21-23, 2002.

Appendix B Lophaven, S., Carstensen, J., and Rootzén, H. (2004). Space-time modeling of environmental monitoring data. *Environmental and Ecological Statistics* vol. 11, p. 237-256.

Appendix C Lophaven, S., Carstensen, J., and Rootzén, H. (2003). Space-time modeling of dissolved inorganic nitrogen. In *Bulletin of the International Statistical Institute 54th Session*, Contributed Papers, vol. LX(1), pp. 749-750, International Statistical Institute, Berlin, Germany, August 13-20, 2003.

Appendix D Lophaven, S., Carstensen, J., and Rootzén, H. (2004). Stochastic modelling of dissolved inorganic nitrogen. *Ecological Modelling*. Submitted.

Appendix E Lophaven, S., Carstensen, J., and Rootzén, H. (2004). Computing spatial designs in R. *Computers & Geosciences*. Submitted.

Appendix F Lophaven, S., Carstensen, J., and Rootzén, H. (2004). A Bayesian geostatistical approach to optimal design of monitoring networks. *Environmental and Ecological Statistics*. Submitted.

Acknowledgements

I wish to thank all who have contributed to this research. First of all, I owe great thanks to my supervisors Associate Professor Helle Rootzén, Informatics and Mathematical Modelling (IMM), the Technical University of Denmark (DTU), and Senior Research Scientist Jacob Carstensen, Department of Marine Ecology, the National Environmental Research Institute of Denmark (NERI) for their help and guidance, and for the many inspiring discussions we have had during the last three years.

I also wish to thank colleagues at IMM for a stimulating statistical environment and participants of the STAMP[1] research group for interesting and valuable discussions.

Finally, I am very grateful to Professor Peter Diggle and his colleagues at Department of Mathematics and Statistics, Lancaster University for their hospitality during my stay there.

[1] Statistical Analysis and Modelling of Phytoplankton Dynamics

Contents

Chapter 1

Introduction

Environmental monitoring is often conducted over time at a number of fixed sampling sites. In the past, most resources for monitoring programs have been allocated to collecting data, while less emphasis has been put into designing monitoring programs, in-depth statistical analysis of the data and development of methods for extracting meaningful information. Ward et al. (1986) referred to this as the "Data-rich but Information-poor" syndrome.

This thesis applies statistical methods for designing an environmental monitoring program, and analysing data from it. The dataset used as an example consists of measurements of salinity, temperature, nutrient concentrations, chlorophyll-a and phytoplankton from the Kattegat in the period 1993-1997. The Kattegat is the estuaries between Denmark and Sweden limited to the north by the North Sea. The dataset is a part of the Danish marine monitoring program, which was established in connection with the adoption of the Action Plan on the Aquatic Environment in 1987.

Although a quite comprehensive dataset exists, the monitoring program in the Kattegat comprise a puzzle where at many sampling sites only a few measurements have been carried out during the five-year period. Hence, many sampling sites may not provide any detailed information when considered individually, while combining data from several sampling sites could potentially increase the information content significantly.

The overall objective of this thesis is to investigate how the reporting on the state of the marine environment in the Kattegat can be improved by means of statistical methods for data analysis and design of monitoring programs. An improvement in the reporting on the state of the marine environment would increase the knowledge on the processes in the marine environment and provide a better assessment of effects of anthropogenic stresses. Statistical methods have been developed with this specific application in mind. However, the generality of the methods is discussed throughout the thesis.

1.1 Outline of the thesis

This thesis consists of six research papers (Papers A-F) and a summary report (Chapters 1-5), for which an overview is given below. For the summary report the content of the individual chapters is shortly described, while the presentation of the research papers aims at shortly describing the objectives, the methodology and the main results of the individual papers.

1.1.1 The summary report

Chapter 2 gives an overview of the Danish marine monitoring program, including a description of the hydrography of the Kattegat basin from where the data used in this thesis were obtained. Furthermore, a short description of how the measurements were carried out and what information is contained in the data is given. The description is supported by statistical descriptive analyses and aims at providing readers without a chemical or biological background information about the monitoring data analysed.

Chapter 3 gives an overview of the statistical theory applied in the research papers, including a description of geostatistics (section 3.1), space-time statistics (section 3.2) and geostatistical design methods (section 3.3). The primary focus is on geostatistics, because this gives the background theory of space-time statistics and geostatistical design methods. The theory of these two topics can be seen as extensions of the geostatistical theory. Rather than trying to give a very comprehensive description, the chapter aims at shortly describing the principles of the theory, and moreover refer to the relevant literature for further reading. This should enable readers without a geostatistical background to better understand the applications in the research papers. The description is supported by statistical analyses using marine monitoring data from the Kattegat area.

The applied statistical methods and the results obtained in the thesis are discussed in chapter 4, while chapter 5 concludes on the obtained results.

1.1.2 The research papers

The six research papers can be separated into three groups: 1) Paper A, which is about parameter estimation in geostatistics 2) Paper B, C and D, which deal with space-time modelling of nutrients in the Kattegat 3) Paper E and F, which are about geostatistical design methods. The first group (Paper A) deals with classical geostatistics, and in that sense it can be seen as a background paper. The second and third group deals with two different extensions of classical geostatistics, as illustrated in Figure 1.1.

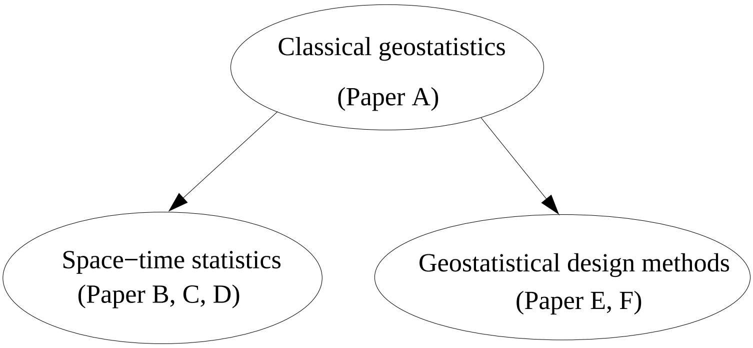

Figure 1.1: *Overview of the research papers included in the thesis.*

When applying geostatistics the spatial correlation in data has to be determined. Spatial correlation is usually modelled by a parametric correlation or covariance function. Various methods for estimating the parameters in such functions have been proposed, and paper A compares the efficiency of eight of these. When assuming that data are normally distributed likelihood-based methods, such as maximum likelihood (ML) and restricted maximum likelihood (REML) can be used. Another way, which is traditionally used within geostatistics, is to estimate the parameters by least squares fitting of the sample (experimental) semivariogram with a valid semivariogram model. The comparison includes ML and REML as well as six estimation approaches based on the sample semivariogram. The comparison is made from a simulation study, where values of a Gaussian random field are simulated in a number of spatial locations. Different types of covariance functions, parameters of these, grid structures and sample sizes are investigated. The comparison show that when data are normally distributed maximum likelihood estimation results in the smallest variances, while the smallest biases, on the other hand, were found when restricted maximum likelihood is used to estimate the parameters. The results also show that spatial predictions computed by kriging, is relatively insensitive to the choice of estimation method.

Paper B aims at developing a statistical approach, which can be used to predict nutrient concentrations with a temporal resolution of one week at the locations of monitoring stations in the Kattegat. The model is formulated as a sum of a mean field and a residual component. The mean field is discretised so that variations between monitoring stations and time intervals are described by means of indices for each week and monitoring station. Different modelling approaches are tested and compared by means of cross validation, which shows that the inclusion of a spatial covariance structure to individual weeks gives predictions which are more efficient than assuming uncorrelated residuals.

Paper C also aims at developing a statistical approach, which can be used to predict nutrient concentrations with a temporal resolution of one week at the locations of monitoring stations in the Kattegat. Paper C use paper B as a starting point, and focus on improving the weaknesses of this approach, e.g. the high number of parameters and the lack of temporal correlation in the model. A station-effect is still included in the mean field while the week-effect in the model in paper B is substituted by the sum of a year-effect and two sine-functions. Both spatial and temporal correlation are included in the residual component, and the spatial covariance structure is modelled by means of a separable space-time model. Results are presented for two monitoring stations and show that the predictions fit observations quite well.

Paper D extends the model presented in paper C to non-sampling locations by geostatistical modelling of the station effect. Both spatial and temporal correlation are included in the residual component, and the spatial covariance structure is modelled by means of a separable space-time model. The approach can be applied to compute maps of the spatial distribution of monitoring data with a weekly resolution. Results are presented for dissolved inorganic nitrogen in four different weeks, representing the four seasons, and at three different locations, and agree very well with our prior belief about the spatial distribution and temporal dynamics of this variable.

Paper E describes, implements and tests some of the geostatistical design methods found in the literature. These methods focus on either designs which are optimal from a spatial prediction point of view, or designs optimal for estimation of the parameters in a geostatistical model, and hence they could be separated into two groups. The problem of the first group of methods is that these assume that the model parameters are known, which in reality they are not. The problem of the second group is that the final goal of most geostatistical applications is to compute efficient predictions rather than estimating model parameters. The paper also suggests how the two groups of design methods could be combined.

Paper F describes and applies a geostatistical approach for reducing the number of sampling stations in the Kattegat area. The choice of design approach is based on the review of geostatistical design methods given in paper E. The idea of the applied approach is to find the design which produces the most efficient spatial predictions whilst making proper allowance for the effects of parameter uncertainty. The design criterion to be optimised is formulated based on the variance of the predictive distribution computed by Bayesian kriging, and the application showed that the number of monitoring stations can be reduced from 31 to 14 with only a marginal increase in this criterion.

Chapter 2

The Danish marine monitoring program

The first real marine monitoring program in Denmark was the Belt Sea project from 1974 to 1978. In 1979 a national monitoring program was established including the monitoring set out by HELCOM's first monitoring program. In the beginning the national monitoring program only covered the Belt Sea and Arkona Basin while Kattegat, Skagerrak and the North Sea were included at a later stage. In 1987 the first Danish Action Plan for the Aquatic Environment was passed through the Danish parliament, resulting in a new more extensive monitoring program of the Danish waters starting in 1989. This program was revised in 1993 to incorporate more intensive monitoring stations. A new revision started in 1998 including new measurement variables such as harmful substances and biological effects, however, with reductions in other measurement variables.

The monitoring in Danish waters is carried out by National Environmental Research Institute (NERI) and 15 counties. Monitoring at some HELCOM stations is co-ordinated with Norway, Sweden and Germany. The majority of stations are located in tributaries and coastal areas, where the monitoring is conducted by the local authorities (counties).

2.1 Area of study - Kattegat

The Kattegat is a shallow transition area between the saline North Sea/Skagerrak and the brackish Baltic Sea (Figure 2.1) with a surface area, volume and average depth of 22,290 km^2, 533 km^3 and 23.9 m, respectively (Gustafsson, 2000). The general circulation is dominated by north-flowing surface water with a salinity gradient from 15-30 psu and south-flowing deep water with salinities around 30-34 psu. It is considered to be almost permanently stratified with a halocline located at approximately 15 m depth (Andersson and Rydberg, 1988).

The Kattegat is characterised by a coastal shelf <20 m in the western part and a trench in the eastern part, where the outflow from the Baltic Sea dominates. Major tributaries to the Kattegat are scattered along the Jutland and Swedish coast and include Limfjorden, Randers Fjord, Rönne Å , Lagan, Nissan, Ätran, Viskan and Göta river, which occasionally spills into Kattegat with winds from northerly directions. The Kattegat-Skagerrak front in the northern part, where surface salinities rapidly changes with 5-10 psu (Jakobsen, 1997), is another important feature leading to increased primary production (Richardson, 1985).

Water exchange between the Baltic Sea and the North Sea through Kattegat is closely coupled to the wind conditions. Strong westerly winds forces Skagerrak water into Kattegat building up a southward surface current, while easterly wind forces Baltic Sea water through Øresund and the Great Belt. Changing wind conditions give rise to alternating flow patterns in Kattegat, however, on an annual basis there is a net flow from the Baltic Sea of 470 km^3, which is equal to the freshwater input to the Baltic Sea. When the wind is westerly, surface water is initially blown away from the east coast of Jutland and sea level rises along the west coast of Sweden. Replacement of surface water along the Jutland coast with nutrient-rich bottom water (upwelling) is an important process for bringing nutrients to the surface layer (Kiørboe, 1996).

The Jutland Coastal Current is another important source of nutrients to the Kattegat. It derives from the German Bight bringing nutrient-rich water along the west coast of Jutland and occasionally spills into Kattegat. The episodic character of the Jutland Coastal Current is governed by wind conditions where strong south-westerly wind forces the current into Kattegat, where it enters at intermediate depths of 10-25 m and can mix with both surface and bottom waters. It is estimated that 10-20% of the bottom water derives from the German Bight with the Jutland Coastal Current (Christensen, 1998).

Bioassay studies have shown that primary production is nitrogen limited (Granéli, 1987; Granéli et al., 1990). The external loading of total nitrogen was on average

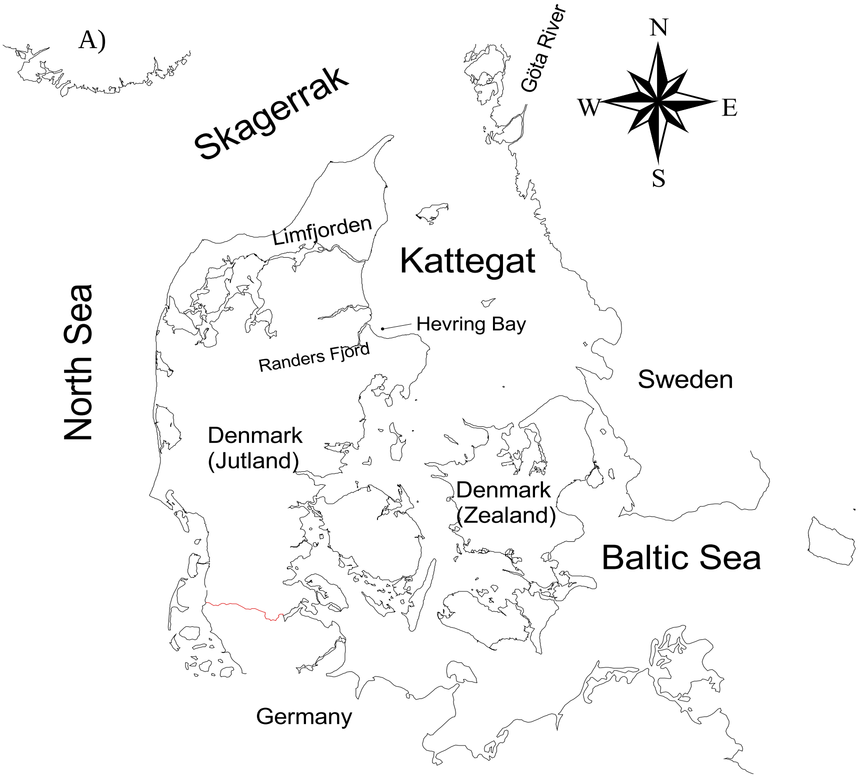

Figure 2.1: *The Kattegat area as a transitional sea between the North Sea and the Baltic Sea.*

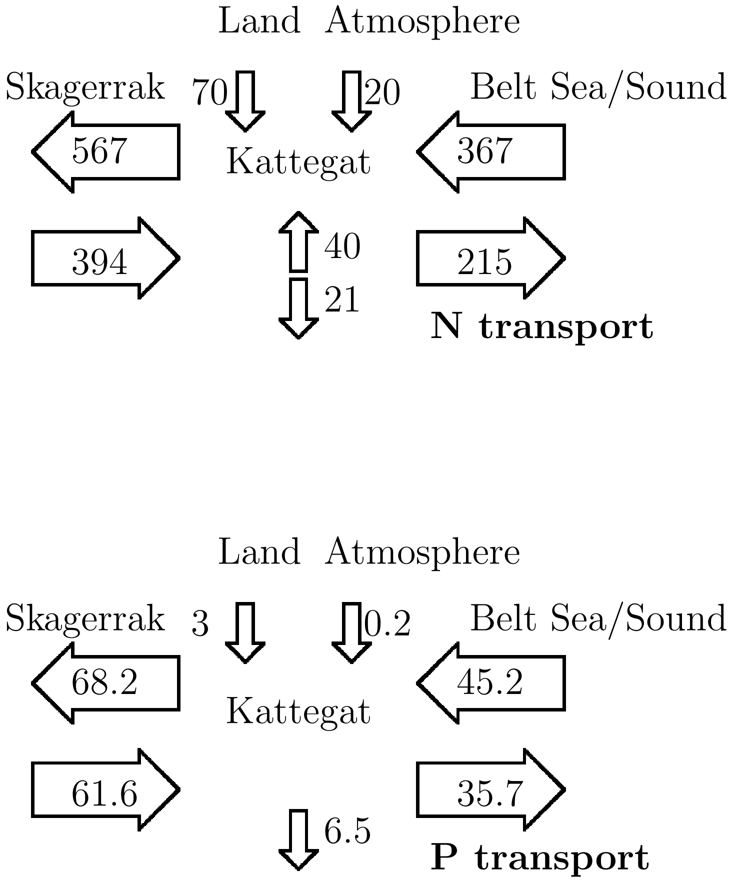

Figure 2.2: *Nitrogen and phosphorus budgets for the Kattegat in 1000 tons per year (Christensen, 1998).*

69000 ton N/year (1989-1997) with large inter-annual and seasonal variations.

During the last four decades the Kattegat has been seriously affected by eutrophication, and frequent oxygen depletions of bottom waters have been recorded (Andersson, 1996). The nitrogen load has increased fourfold during the period 1930-80 (Edler, 1984) and doubled in the period 1950-80 (Ærtebjerg, 1986). The gross nutrient budgets are dominated by the water exchange with the North Sea and the Baltic Sea (Figure 2.2). However, if we consider the net fluxes the load from land and atmosphere becomes important. A total of 90-95% of the annual primary production is degraded in the water mass and at the sediment-water interface (Anton et al., 1993).

2.2 Description of data

The majority of monitoring data is sampled by traditional shipboard surveys from research vessels, which has been constructed for this specific purpose, visiting a number of predefined stations (Figure 2.3). New emerging technologies include continuous measurements from moored buoys, ships-of-opportunity, remote sensing etc. Although several of these technologies make good promises for the future of monitoring, they are still in a premature stage to substitute the traditional shipboard sampling. The focus of this work have therefore been directed towards the information contained in traditional monitoring data.

The water column is normally sampled at discrete depths for analysis. Water samples at discrete depths are pooled to constitute an integrated sample over a depth interval. Samples representing discrete depths or depths intervals are analysed to measure the hydrochemical and biological composition. The data used in this thesis are surface concentrations, measured in the five-year period 1993-1997. Surface concentrations represent the upper 10 meters of the water column. The different variables in the dataset are briefly described below. The description is accompanied by statistical descriptive analyses, i.e. Figure 2.4 shows for each variable in the dataset the total number of observations, the number of monitoring stations where samples have been taken, and the number of weeks where samples have been taken, Figure 2.5 shows the temporal dynamics of the variables in the dataset at individual monitoring stations, Figures 2.6 and 2.7 show histograms of the variables in the dataset, while Figure 2.8 shows the relationship between the variables in the dataset.

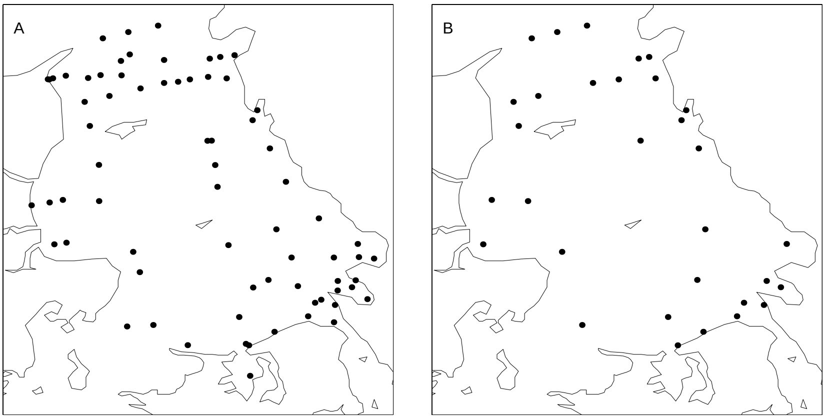

Figure 2.3: *A) Locations of all the monitoring stations included in the dataset. B) Locations of monitoring stations which are intensively sampled nowadays.*

2.2.1 Salinity (SALI) and temperature (TEMP)

Salinity is given in psu (practical salinity unit). It is chosen so that a water mass with approximately 35g salt/kg has a salinity of 35 psu. In the Kattegat the salinity varies from approximately 30 psu in the northern part to approximately 20 psu in the south. Above 30 psu is termed as oceanic environment while freshwater is characterised by less than 0.5.

Today salinity is estimated with the help of conductivity, which is measured with a CTD probe (CTD=Conductivity, Temperature, Depth). The probe measures conductivity, temperature and pressure simultaneously while it is continuously lowered from the surface to the bottom. Salinity is then calculated from these continuous observations.

Advanced thermometers were developed for oceanic conditions in the beginning of this century. However, temperature may also be measured with the help of a thermistor. This is lowered into the water and the temperature is reported at predefined depths.

2.2.2 Dissolved inorganic nitrogen (DIN)

DIN is the inorganic form of nitrogen that is used as nitrogen source by the primary producers. In fact, in most marine waters DIN is assumed to control the size of the primary production. High supplies of DIN can cause a severe algae bloom. The concentration of DIN in the euphotic zone is mainly controlled by the biological activity. Other factors that have an impact is the mixing of surface and deep waters, load from land and atmosphere. Normally the concentration in the surface layer is relatively low (0-2 μmol l^{-1}) during the production season, i.e. March-September, while the DIN concentration in the surface waters increases during winter. DIN consists of several nitrogen compounds (mainly ammonia, nitrite and nitrate), which are all more or less readily available for primary production.

Ammonia is formed when proteins and other nitrogen rich organic compounds are reduced. In undisturbed waters the ammonia is oxidised, by nitrifying bacteria, to nitrate. In oxygenated marine waters the typical ammonia concentration level is 0-0.5 μmol l^{-1}. Ammonia is measured with photometric or ion selective electrodes.

Nitrite is the inorganic nitrogen form that constitutes the middle step in the microbial oxidation/reduction processes between ammonium and nitrate. In conditions of low oxygen supply, nitrate may be reduced to nitrite and in an anaerobic situation nitrite and nitrate can be reduced by denitrifying bacteria to nitrogen gas.

Nitrate is analysed photometrically where nitrate is reduced to nitrite before the analysis. The nitrite concentration is then decided with the help of a colour reagent. In addition, nitrate may also be measured using an ion selective electrode or with ion chromatography.

2.2.3 Dissolved inorganic phosphorus (DIP)

Phosphorus is a necessity for all living organisms but phosphate (equivalent to DIP) is the only form of phosphorus that plants can assimilate. It is released either directly from living organisms or by dead organic material. The latter occurs mainly in the deep waters where the phosphate concentrations are high.

The ratio of phosphate/phosphorus is hard to measure as it varies greatly with season, type of water area and effects from the surroundings. Usually, however, this relation is 1/3 or larger. During the production season the phosphate concentrations in the surface waters are very low - sometimes even down to the limit of detection. Despite this, several algae can reach a high growth rate. This is partly a consequence of its high recycling rate.

At oxic conditions, in the sediments, phosphate is adsorbed onto clay particles and iron oxyhydroxides. By doing this ten percent of the deep water phosphorus leaves the water mass. At conditions with low oxygen the iron will be reduced (from Fe^{3+} to Fe^{2+}) and the adsorbed phosphate will once again be available in the water mass. Phosphate is analysed photometrically, usually on an unfiltered sample.

2.2.4 Silicate (DSi)

Dissolved and particulate Si are supplied by rivers to the marine environment. Dissolved silicate (DSi) comes from the weathering of soil and bedrock. Particulate silicate comes in two forms: 1) mineral silicates, which are considered unreactive on biological time scales, and 2) biogenic silica (BSi), an amorphorus form of Si biogenically precipitated by a variety of organisms, but mostly from diatoms in aquatic systems.

DSi is an important nutrient element and can influence the growth of diatoms in aquatic environments and DSi:DIN ratios suggest that DSi will be more limiting than DIN (dissolved inorganic nitrogen) concentrations for diatoms in the future (Rahm et al., 1996). DSi is also analysed photometrically.

2.2.5 Chlorophyll a (CHLA)

Chlorophyll concentration is a surrogate measurement of the phytoplankton biomass in a water sample. It varies with light, temperature, and availability of nutrients and grazing pressure. Chlorophyll is measured in the upper layer, just below the surface. The sample is filtered and then dissolved, and the colour of the solution is decided with the help of spectrophotometer or a flourometer.

2.2.6 Phytoplankton biomass (PB)

Phytoplankton biomass by species is not only important as a quantitative measure of the amount of algae in the water column. These measurements provide invaluable information on the phytoplankton composition, which is important for understanding the function of the algae community. In particular, identification of harmful algae blooms is of great interest to the public. Phytoplankton biomass is normally determined from samples representing an integrated depth interval (e.g. 0-10 m).

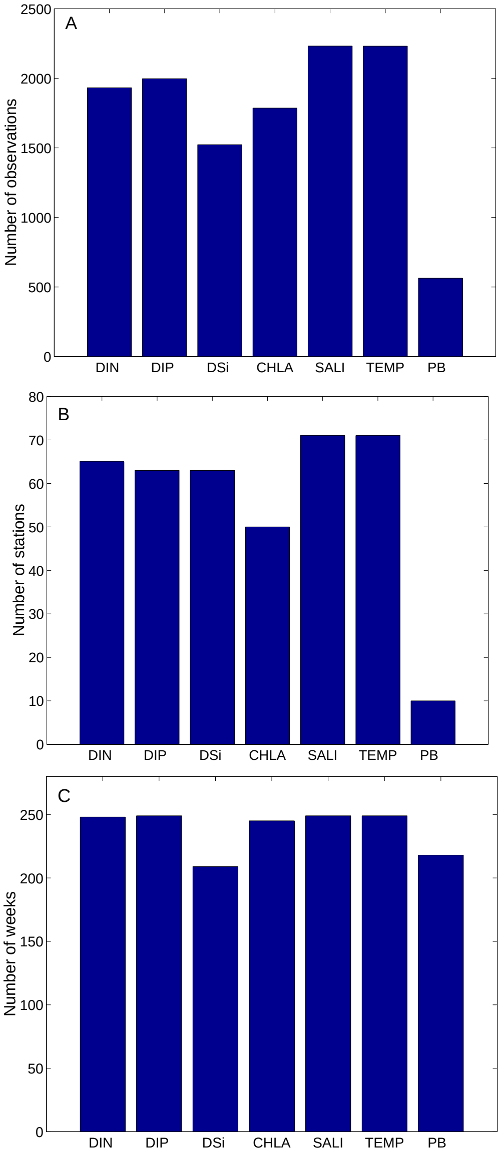

Figure 2.4: *A) The total number of observations. B) The number of monitoring stations where samples have been taken. C) The number of weeks where samples have been taken.*

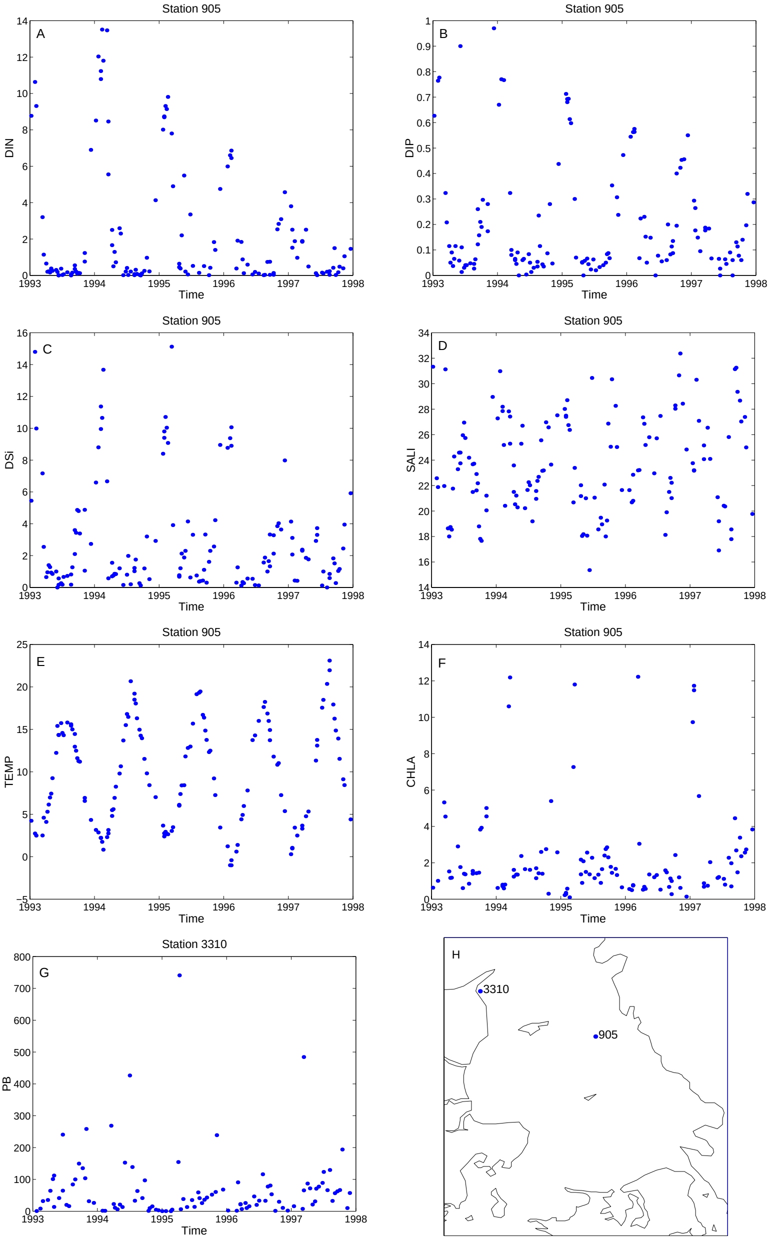

Figure 2.5: *Temporal dynamics of nutrients (DIN, DIP and DSi), salinity (SALI), temperature (TEMP) and chlorophyll a (CHLA) at monitoring station 905. The dynamics of phytoplankton biomass (PB) is shown for station 3310.*

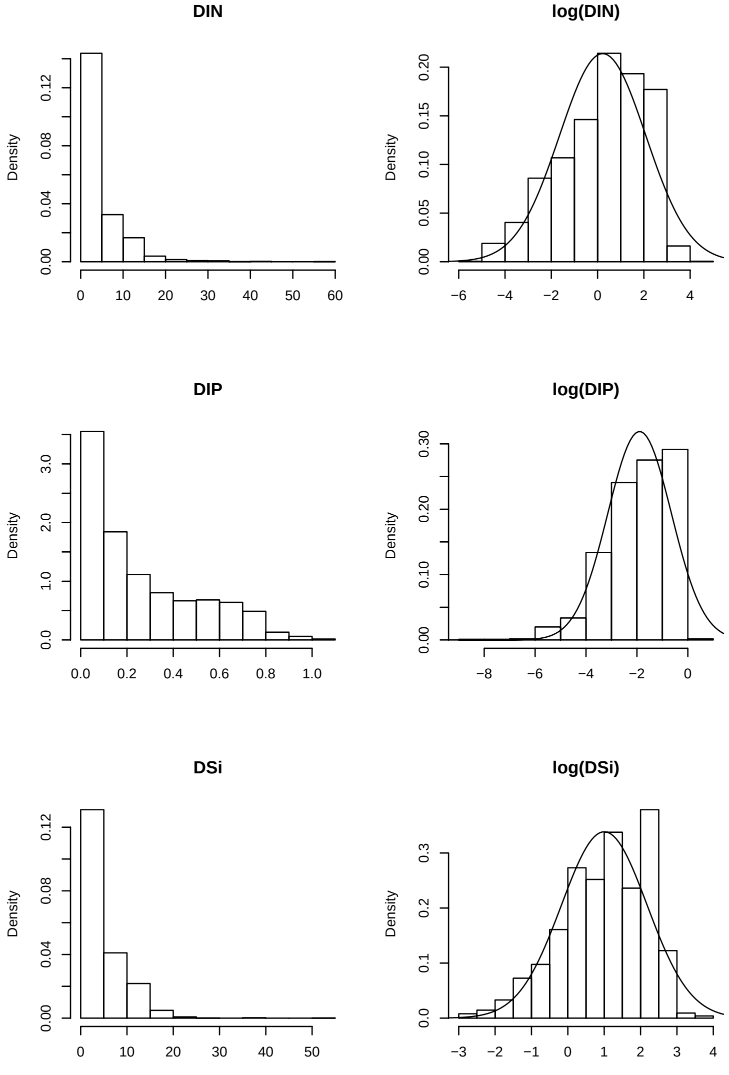

Figure 2.6: *Histograms of nutrients (DIN, DIP and DSi) and of log-transformed nutrient concentrations.*

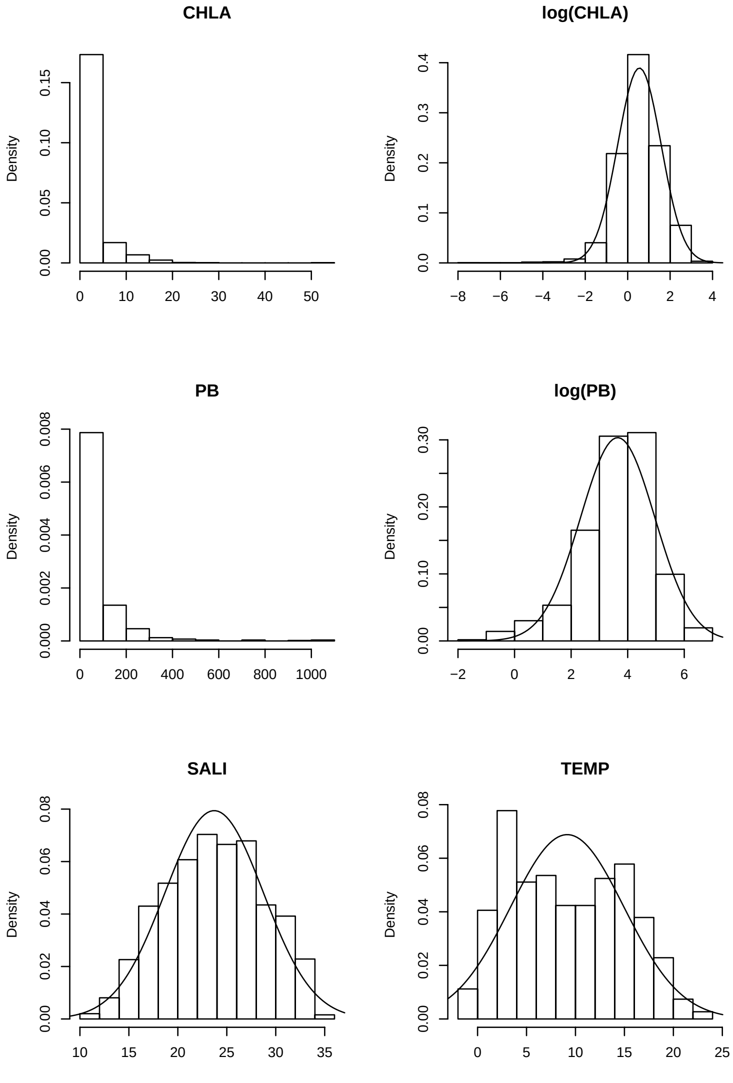

Figure 2.7: *Histograms of chlorophyll a (CHLA), log-transformed chlorophyll a, phytoplankton biomass (PB), log-transformed phytoplankton biomass, salinity (SALI) and temperature (TEMP).*

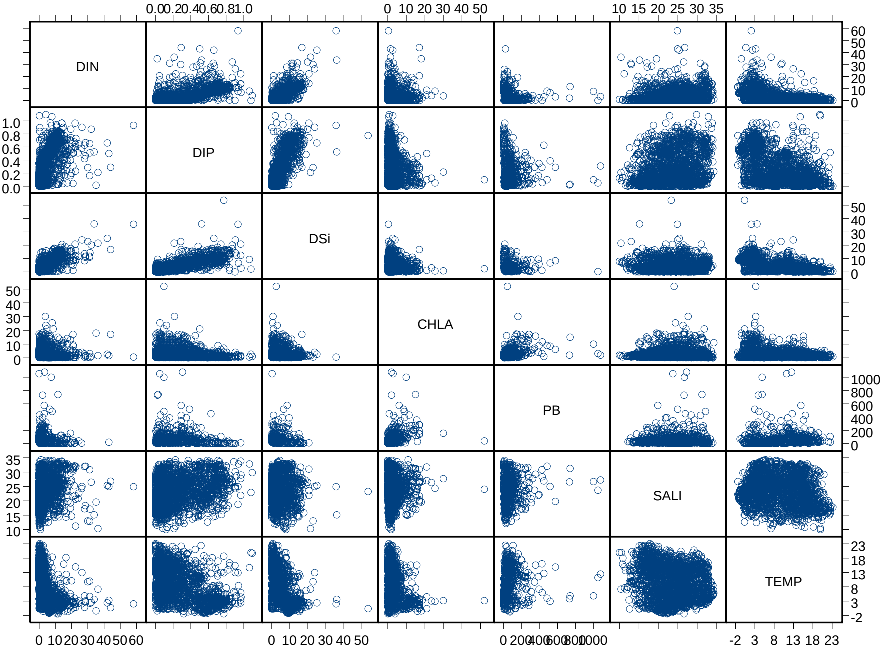

Figure 2.8: *Matrix plot showing the relationship between nutrients (DIN, DIP and DSi), salinity (SALI), temperature (TEMP), chlorophyll a (CHLA) and phytoplankton biomass (PB).*

Chapter 3

Methodology

This chapter describes the theory of geostatistics (section 3.1), space-time statistics (section 3.2) and geostatistical design methods (section 3.3). The chapter aims at shortly describing the principles of the theory applied in the research papers, rather than trying to give a very comprehensive description. The primary focus is on geostatistics, because this gives the background theory of space-time statistics and geostatistical design methods.

3.1 Geostatistics

Geostatistics involves the analysis and prediction of continuous spatial phenomena (Cressie, 1993), such as metal grades, porosities, pollutant concentrations etc. The prefix geo- comes from geology, since geostatistics has its origins in mining. Nowadays, geostatistics is just a name associated with a class of techniques used to analyse and predict values of a spatially distributed variable, which are implicitly assumed to be spatially correlated.

A lot of books have been written on the subject of geostatistics, ranging from books which mainly focus on the applications of geostatistics (Isaaks and Srivastava, 1989), to books which treat geostatistics from a more mathematical and statistical point of view (Cressie, 1993; Stein, 1999; Diggle et al., 2003). Other good and recent books on geostatistics are Wackernagel (2003); Chilès and Delfiner (1999); Kitanidis (1997). In this thesis the intension has been to provide a statistical description of geostatistics, because it makes it easier to see the link between geostatistics and other areas of statistical theory. The description below is mainly inspired by the work of Diggle et al. (2003); Stein (1999).

Given data y_i, $i = 1, ..., n$ at spatial locations x_i it is assumed that data follow the model

$$Y_i = S(x_i) + Z_i, \qquad i = 1, ..., n \tag{3.1}$$

where $S(x)$ is a stationary Gaussian process with expectation $\mathrm{E}[S(x)] = \mu = F\beta$, where F is a $n \times p$ matrix of covariates, and β is the parameter vector. Furthermore, $S(x)$ has variance $\mathrm{Var}[S(x)] = \sigma^2$ and correlation function $\rho(u) = \mathrm{Corr}[S(x_i), S(x_j)]$, where $u = \parallel x_i - x_j \parallel$ is the spatial distance between x_i and x_j, and $Z_i \sim N(0, \tau^2)$.

3.1.1 The spatial correlation function

From the above description it is seen that the correlation function $\rho(u)$ only depends on the distances between observations, and eventually on the direction (see section 3.1.4). The correlation is usually modelled by some parametric function, which has to be valid in the sense that it has to be positive definite, see Schlather (1999) and research paper A for an introduction. A widely used and valid family of models is the spherical family given by

$$\rho(u; \phi) = \begin{cases} 1 - \frac{3}{2}\frac{u}{\phi} + \frac{1}{2}\frac{u^3}{\phi^3} & : \quad 0 \leq u \leq \phi \\ 0 & : \quad u > \phi \end{cases} \tag{3.2}$$

where $\phi > 0$ is a parameter describing the correlation. As seen this family of correlation models only have one parameter which makes it less flexible compared to the other widely used families. Another disadvantage of the spherical family is that it can cause difficulties when using likelihood-based methods for estimating the unknown parameters (Stein, 1999; Diggle et al., 2003). Another family of models is the powered exponential family, which is defined for $\phi > 0$ and $0 < \kappa \leq 2$ and given by

$$\rho(u; (\phi, \kappa)) = \exp\Big(- \big(\frac{u}{\phi}\big)^{\kappa} \Big) \tag{3.3}$$

The inclusion of an additional parameter κ makes this family more flexible than the spherical. For $\kappa = 1$ we have the exponential correlation model, while

$\kappa = 2$ leads to the Gaussian model. It is well known that applying the latter may cause the correlation matrix to be ill-conditioned (Ababou et al., 1994). Estimation of the parameters in the above described models can be based on likelihood methods (maximum likelihood or restricted maximum likelihood) or on the sample (or experimental) semivariogram. A thorough description and comparison of these methods is given in research paper A. At this point it should be noted that the approach suggested by Cressie (1985), which is included in the comparison in paper A, has received some criticism, see e.g. Zhang et al. (1995).

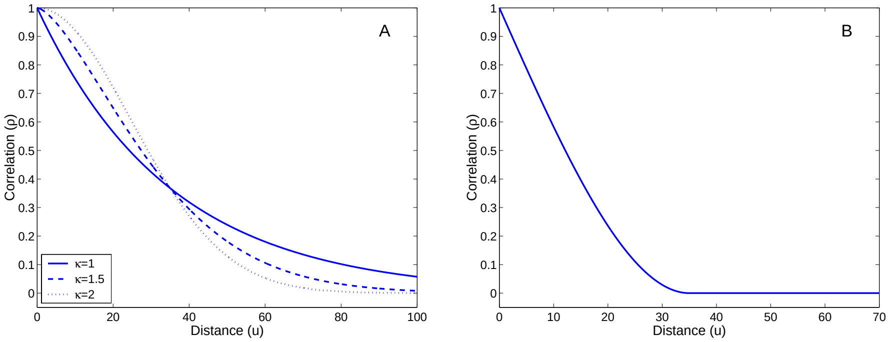

Figure 3.1: *A) The powered exponential correlation model with ϕ=35 and κ=1, 1.5 and 2. B) The spherical correlation model with ϕ=35.*

3.1.2 Prediction

In geostatistics predictions are usually computed via kriging. The kriging predictor is the predictor that minimizes $\mathrm{E}[(\hat{S}(x) - S(x))^2]$. It can be shown that the kriging predictor for $T = S(x_0)$ is

$$\hat{T} = \mu + \sigma^2 r^T (\tau^2 I + \sigma^2 R)^{-1} (y - \mu I) \tag{3.4}$$

with prediction variance

$$\mathrm{Var}[T|y] = \sigma^2 - \sigma^2 r^T (\tau^2 I + \sigma^2 R)^{-1} \sigma^2 r \tag{3.5}$$

where R is a symmetric $n \times n$ matrix with elements $\rho(\parallel x_i - x_j \parallel)$ and r is a $n \times 1$ vector with elements $\rho(\parallel x_0 - x_i \parallel)$. The case where $\mathrm{E}[S(x)] = \mu = F\beta$ is referred to as universal kriging, while the case $\mathrm{E}[S(x)] = \mu = \beta$ is called ordinary kriging, and is probably the most frequently used kind of kriging. The case where μ is constant and known, i.e. not estimated from data, is called simple kriging. Predictions via some form of kriging have been computed in all six research paper.

The kriging predictor compromises between the mean μ, i.e. what we believe is the truth, and the observed data y_i. This is very similar to the way the Kalman filter works. The compromise depends on the location in which we want to compute a prediction, the n data-locations, the values of the model parameters as well as the chosen model.

3.1.3 Simulation

As described above predictions computed via kriging (3.4) minimise the error variance. However, there is no guarantee that the predictions have the same covariance structure as the original data. Simulation allows us to come up with a number of realizations (typically some hundreds or thousands) of maps, each of which has approximately the same covariance structure as of the original data. Theoretically, the average of a large number of simulated maps would look like the kriged map. Simulating a Gaussian random field in n points with zero mean and a $n \times n$ covariance matrix Σ involves calculating the "square root" $\Sigma^{1/2}$ of Σ such that $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^T$. This can be done using Cholesky decomposition which requires that the covariance matrix is positive definite. Afterwards a simulated Gaussian random field $S = \Sigma^{1/2}Z$ is computed, where the elements of the vector Z are simulated independently from a standard normal distribution, i.e. $Z_i \sim N(0,1)$, $i = 1, \cdots, n$. Simulation of Gaussian random fields is used in research paper A and F.

3.1.4 Directional effects

Directional effects can be induced by environmental conditions such as wind, soil formation etc. As a consequence the spatial correlation may vary not only with the distance between observations but also with direction. Such directional effects can be identified by estimation of directional sample semivariograms, and is referred to as anisotropy (Zimmerman, 1993). Given the angle of anisotropy ψ_A and the anisotropy ratio ψ_R, anisotropy can be handled by rotating and stretching the original spatial coordinates (x_1, x_2)

$$(x_1^{'}, x_2^{'}) = (x_1, x_2)\begin{pmatrix} \cos(\psi_A) & -\sin(\psi_A) \\ \sin(\psi_A) & \cos(\psi_A) \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & \psi_R^{-1} \end{pmatrix} \tag{3.6}$$

After transforming the original coordinates (x_1, x_2), the spatial correlation function can be modelled as a function of the distance in the transformed space

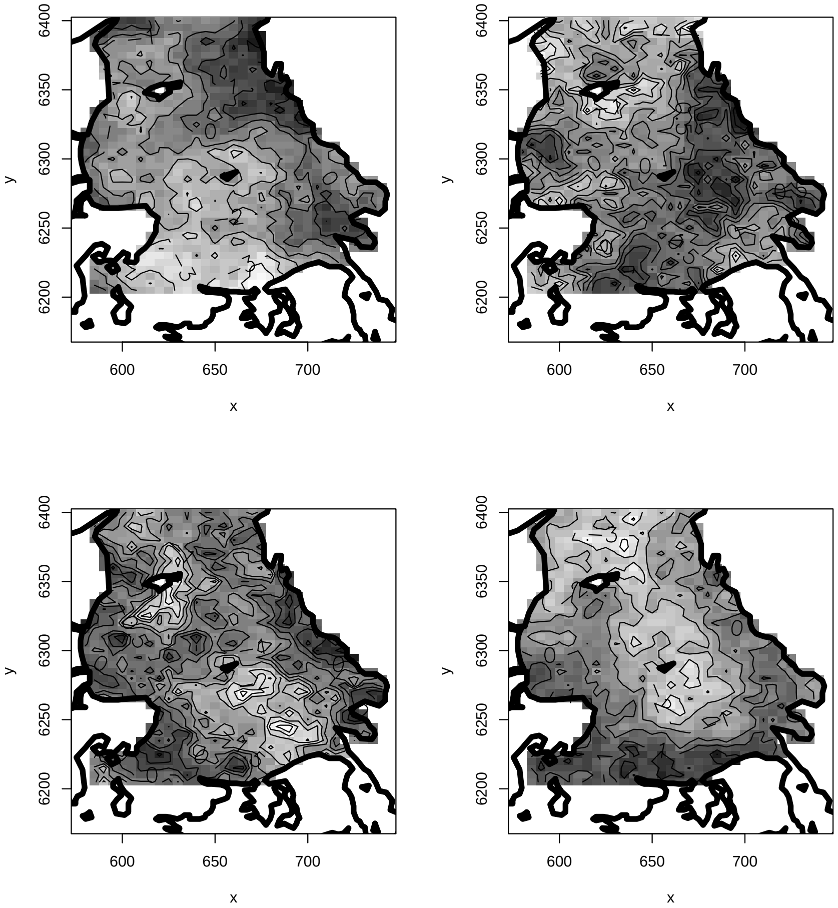

Figure 3.2: *Four simulations in the Kattegat of Gaussian random fields with zero mean and exponential covariance structure with σ^2=1.5 and ϕ=35.*

$(x_1^{'}, x_2^{'})$.

3.1.5 Multivariate geostatistics

When more than one variable have been measured in the study area, it is sometimes desirable also to include the spatial cross-correlation between variables in the prediction of the variable of primary interest. Cokriging is a multivariate extension, which can handle this situation (Wackernagel, 2003; Kitanidis, 1997). Usually the linear model of coregionalization is used to determine the geostatistical model, including the covariance structures of the variables involved, as well as their pairwise cross-covariance structures. If the number of variables involved are denoted by p and the number of data points by n, then predictions are computed by

$$\hat{T}_m = \Gamma_m^T y_m \tag{3.7}$$

where $\hat{T}_m$ is a $p \times 1$ vector of predictions at location x_0, y_m is a $np \times 1$ data vector, and Γ_m is found by solving

$$K_m \Gamma_m^T = K_{0m} \tag{3.8}$$

with

$$K_m = \begin{pmatrix} K & II \\ II^T & 0 \end{pmatrix} \tag{3.9}$$

where II is the $np \times p$ matrix formed of n identity matrices of size $p \times p$, 0 is a $p \times p$ matrix of zeros, and K is the $(np \times np)$ matrix of point to point covariances for the p variables. Thus, K is formed of $n \times n$ submatrices K_{ij} of size $p \times p$ giving the covariances between point i and j for the p variables. Furthermore, K_{0m} in (3.8) is

$$K_{0m} = \begin{pmatrix} K_0 \\ I \end{pmatrix} \tag{3.10}$$

where K_0 is the $np \times p$ matrix of covariances between a sampling point and x_0, i.e. K_0 is formed of n matrices K_{0i} of size $p \times p$ for the p variables, and I is a $p \times p$ identity matrix.

3.1.6 Bayesian geostatistics

In the Bayesian geostatistical approach model parameters, i.e. the parameters describing the mean field as well as the covariance structure, are treated as random variables. This means that parameter uncertainty is incorporated in the computed predictions, which makes the approach attractive to apply for constructing or comparing monitoring networks. Bayesian methods for geostatistical analysis were proposed independently by Kitanidis (1986); Le and Zidek (1992); Handcock and Stein (1993). A recent and thorough description is given in Banerjee et al. (2004). Diggle et al. (1998) extended the Bayesian methodology to embrace generalized linear models (McCullagh and Nelder, 1989).

In general given data y and prior distributions $pr(\theta) = pr(\beta, \sigma^2, \phi, \tau^2)$ of the parameters the posterior distributions are found using the relation

$$\begin{aligned} pr(\beta, \sigma^2, \phi, \tau^2 | y) \quad \propto \quad & pr(\beta, \sigma^2, \phi, \tau^2) |\tau^2 I + \sigma^2 R|^{\frac{1}{2}} \\ & \exp\left(\frac{1}{2}(y - F\beta)'(\tau^2 I + \sigma^2 R)^{-1}(y - F\beta)\right) \end{aligned} \tag{3.11}$$

and the predictive distribution $pr(y_0|y)$ by

$$pr(y_0|y) = \int pr(y_0|y, \theta) pr(\theta|y) \mathrm{d}\theta \tag{3.12}$$

Classical geostatistical methods estimate the parameters, and then use these to perform predictions as if the estimates were the truth (plug-in prediction). Predictions computed by the Bayesian approach can be interpreted as a weighted average of a number of classical predictions, with weights given by the posterior distributions of the parameters. Furthermore, it is seen in (3.11) that posterior distributions are computed by weighting the likelihood function with the prior distributions.

The prior distributions, or just the prior, of the model parameters can be flat or non-informative, improper or conjugate. Flat priors have a minimal effect on the posterior distributions, improper priors, are priors which are not probability distributions, and conjugate priors are priors which give posterior distributions on closed form, i.e. distributions that can be expressed analytically. Thus, within the general framework, various types of the Bayesian geostatistical approach can be applied, according to how the prior distributions of the model parameters are specified. Classical geostatistical methods like ordinary and universal kriging can be seen as special cases of the Bayesian approach, obtained by specifying an improper uniform prior for β, while assuming that σ^2, ϕ and τ^2 are well-known. In the following we first consider the situation where the correlation parameter ϕ is fixed, τ^2=0, and β and σ^2 are random, and afterwards the more general situation with ϕ random as well.

Fixed correlation parameter ϕ

In the applications of Bayesian geostatistics in this thesis a conjugate prior for $(\beta, \sigma^2|\phi)$, which is the product of normal and scaled-inverse-χ^2 densities, was used. This is given by

$$[\beta, \sigma^2|\phi] \sim N(m_b, \sigma^2 V_b)\chi^2_{ScI}(n_\sigma, S^2_\sigma) \tag{3.13}$$

With the above prior specification the posterior distribution of the random parameters is

$$[\beta, \sigma^2|y, \phi] \sim N(\tilde{\beta}, V_{\tilde{\beta}})\chi^2_{ScI}(n_\sigma + n, S^2) \tag{3.14}$$

where

$$\begin{aligned} V_{\tilde{\beta}} &= (V_b^{-1} + F^T R^{-1} F)^{-1} \\ \tilde{\beta} &= V_{\tilde{\beta}}(V_b^{-1} m_b + F^T R^{-1} y) \\ S^2 &= \frac{n_\sigma S^2_\sigma + m_b^T V_b^{-1} m_b + y^T R^{-1} y - \tilde{\beta}^T V_{\tilde{\beta}}^{-1} \tilde{\beta}}{n_\sigma + n} \end{aligned} \tag{3.15}$$

Having specified the posterior distribution the predictive distribution is given by

$$pr(y_0|y, \phi) = \int \int pr(y_0|y, \beta, \sigma^2, \phi) pr(\beta, \sigma^2|y, \phi) d\beta d\sigma^2 \tag{3.16}$$

where $[y_0|y, \beta, \sigma^2, \phi]$ is a multivariate normal distribution with mean and variance given by (3.4) and (3.5). For the prior distributions considered here analytical solutions can be obtained, i.e. the predictive distribution is a multivariate t-distribution defined by

$$\begin{aligned} [Y_0|y, \phi] &\sim t_{n_\sigma + n}(\mu^*, S^2 \Sigma^*) \\ \mathrm{E}(Y_0|y, \phi) &= \mu^* \\ \mathrm{Var}(Y_0|y, \phi) &= \frac{n_\sigma + n}{n_\sigma + n - 2} S^2 \Sigma^* \end{aligned} \tag{3.17}$$

where S^2 is given by (3.15) and

$$\begin{aligned} \mu^* &= (F_0 - r^T R^{-1} F) V_{\tilde{\beta}} V_b^{-1} m_b + [r^T R^{-1} + (F_0 - r^T R^{-1} F) V_{\tilde{\beta}} F^T R^{-1}] y \\ \Sigma^* &= R_0 - r^T R^{-1} r + (F_0 - r^T R^{-1} F)(V_b^{-1} + V_{\hat{\beta}}^{-1})^{-1}((F_0 - r^T R^{-1} F))^T \end{aligned} \tag{3.18}$$

Random correlation parameter ϕ

In the more general case where the correlation parameter is considered as random, no conjugate prior exist for ϕ. Instead, a discrete prior $\pi(\phi)$ is used, i.e.

a reasonable range of discrete values is selected, and a discrete uniform prior is assigned in a set of values spanning the chosen range. In this case the posterior of the parameters is given by

$$pr(\beta, \sigma^2, \phi|y) = pr(\beta, \sigma^2|y, \phi)pr(\phi|y) \tag{3.19}$$

with $[\beta, \sigma^2|y, \phi]$ given by (3.14) and

$$pr(\phi|y) = \pi(\phi)|V_{\tilde{\beta}}|^{1/2}|R|^{-1/2}(S^2)^{-\frac{n+n_\sigma}{2}} \tag{3.20}$$

where $V_{\tilde{\beta}}$ and S^2 are given by (3.15). Given this the posterior distribution the predictive distribution is given by

$$pr(y_0|y, \phi) = \int pr(y_0|y, \phi)pr(\phi|y)d\phi \tag{3.21}$$

In the situation where ϕ is also treated as a random variable, inference is done by Monte Carlo simulations, where samples are taken from the posterior and predictive distributions and used for inference and predictions. An algorithm for computing the posterior distribution when β, σ^2 and ϕ are random, while τ^2 is fixed is:

1. Choose a range of values for the distribution $(\phi|y)$ which is sensible for the given data, and assign a discrete uniform prior for ϕ on a set of values spanning the chosen range.
2. Compute the posterior probabilities on this discrete support set using (3.20). This defines the discrete posterior distribution $\tilde{pr}(\phi|y)$
3. Sample a value of ϕ from this discrete distribution $\tilde{pr}(\phi|y)$.
4. Attach the sampled value of ϕ to the distribution $pr(\beta, \sigma^2|y, \phi)$ given by (3.14), and sample from this distribution.
5. Repeat steps 3 and 4 as many times as required/desired. The resulting sample is a sample from the joint posterior distribution of the parameters $pr(\beta, \sigma^2, \phi|y)$.

After having obtained the posterior distribution, the predictive distribution is computed by the following algorithm:

1. Choose a range of values for the distribution $(\phi|y)$ which is sensible for the given data, and assign a discrete uniform prior for ϕ on a set of values spanning the chosen range.

2. Compute the posterior probabilities on this discrete support set using (3.20). This defines the discrete posterior distribution $\tilde{pr}(\phi|y)$
3. Sample a value of ϕ from this discrete distribution $\tilde{pr}(\phi|y)$.
4. Attach the sampled value of ϕ to the distribution $pr(y_0|y, \phi)$ given by (3.16), and sample from it to obtain realisations of the predictive distribution.
5. Repeat steps 3 and 4 as many times as required/desired. The resulting sample is a sample from the predictive distribution.

If both ϕ and τ^2 are random, a range of values are chosen for the distribution $(\phi, \tau^2|y)$ in step 1 in both algorithms.

3.2 Space-time statistics

Modelling of space-time phenomena is the subject of paper B, C and D. It is relatively straight forward to extend the geostatistical model (section 3.1) to the space-time domain. However, one major issue is how to specify and model the space-time covariance structure. One approach could be simply to consider time as another dimension. This would require an appropriate metric in space-time, and consequently the technique of separability is used instead, which means that distances in space and time are computed separately. A short description of space-time covariance structures and space-time modelling approaches is given in paper B, while separable models are applied in paper C and D for modelling nutrient concentrations in the Kattegat. It is important to note that not all space-time covariance structures formed by a spatial and a temporal component can be used, e.g. the sum of a spatial covariance and a temporal covariance will not in general or at least for some designs be positive definite, i.e. the kriging equations (3.4) can not be solved. On the other hand the product of two covariances will be a valid covariance, while the product of two semivariograms is not a semivariogram (De Cesare et al., 2001a). Another group of space-time covariance models is the nonseparable models. These are not applied in this thesis, and consequently not treated in this section. The reader is referred to De Iaco et al. (2002); Gneiting (2001); Cressie and Huang (1999); Brown et al. (2000) for a description of such models.

In this section all experimental space-time semivariograms (Figures 3.3, 3.4 and 3.5) and cross-semivariograms (Figures 3.6, 3.7 and 3.8) which can be computed from the Kattegat monitoring data are shown. The computations are based on log-transformed values of DIN, DIP, DSi, chlorophyll a and phytoplankton biomass, while original data values for salinity and temperature are used. Below a short description of the space-time covariance and semivariogram is given.

It is assumed that data values are realisations of a second order stationary space-time random field

$$Y = \{Y(x,t), \quad x \in D, \quad t \in T\} \tag{3.22}$$

with expected value $\mathrm{E}[Y(x,t)] = 0$. In this case the space-time covariance

$$\mathrm{C}_{xt}(u_x, u_t) = \mathrm{Cov}[Y(x+u_x, t+u_t), Y(x,t)] \tag{3.23}$$

and semivariogram

$$\begin{aligned} \gamma_{xt}(u_x, u_t) &= \frac{\mathrm{Var}[Y(x+u_x, t+u_t) - Y(x,t)]}{2} \\ &= \frac{\mathrm{E}[(Y(x+u_x, t+u_t) - Y(x,t))^2]}{2} \end{aligned} \tag{3.24}$$

depend solely on the lag vector (u_x, u_t), i.e. they do not depend on the spatial location or the time. To estimate the experimental space-time semivariogram from data the expectation in 3.24 is replaced by an average, yielding

$$\hat{\gamma}_{xt}(u_x, u_t) = \frac{1}{2N(u_x, u_t)} \sum (y(x+u_x, t+u_t) - y(x,t))^2 \tag{3.25}$$

where $N(u_x, u_t)$ is the number of datapairs with spatial distance u_x and temporal distance u_t (see Figure 3.3). Similarly, the experimental semivariogram for no spatial separation distance (Figure 3.5) or no temporal separation distance (Figure 3.4) can be computed as

$$\begin{aligned} \hat{\gamma}_{xt}(0, u_t) &= \frac{1}{2N(0, u_t)} \sum (y(x, t+u_t) - y(x,t))^2 \\ \hat{\gamma}_{xt}(u_x, 0) &= \frac{1}{2N(u_x, 0)} \sum (y(x+u_x, t) - y(x,t))^2 \end{aligned} \tag{3.26}$$

The corresponding experimental cross-semivariogram describing the space-time variability between two variables Y_1 and Y_2 are estimated by

$$\begin{aligned}
&\hat{\gamma}_{12,xt}(u_x,u_t) = \\
&\quad \frac{1}{2N(u_x,u_t)} \sum (y_1(x+u_x,t+u_t)-y_1(x,t))(y_2(x+u_x,t+u_t)-y_2(x,t)) \\
&\hat{\gamma}_{12,xt}(0,u_t) = \\
&\quad \frac{1}{2N(0,u_t)} \sum (y_1(x,t+u_t)-y_1(x,t))(y_2(x,t+u_t)-y_2(x,t)) \\
&\hat{\gamma}_{12,xt}(u_x,0) = \\
&\quad \frac{1}{2N(u_x,0)} \sum (y_1(x+u_x,t)-y_1(x,t))(y_2(x+u_x,t)-y_2(x,t)) \qquad (3.27)
\end{aligned}$$

Plots of experimental space-time cross-semivariogram for all combinations of two variables are shown in Figures 3.6, 3.7 and 3.8.

3.3 Geostatistical design methods

Environmental monitoring of air, soil and water pollution is a challenge to the modern society. Ultimately, it is desirable to sample at all possible locations within a specific area of interest, but in practice the design of a monitoring program is limited by economic and operational constraints. In such cases the limited number of locations where samples are to be taken has to be determined. Existing monitoring networks have mainly been established on heuristic rules such as appointing a sampling location to be representative for a larger area, rather than defining statistical optimality criteria and determine the location of sampling stations on this basis.

One major problem of using a statistical optimality criterion is that a single criterion encompassing all the different pollutants and biological effects in the monitoring network cannot be established, due to contrasting definitions of optimality. Moreover, the general objectives of a monitoring network is often formulated in rather broad terms making the translation into a more stringent mathematical optimality formulation quite difficult.

A part of this thesis has focused on how to design a monitoring program when the objective is to model the spatial distribution of the studied phenomenon. In this situation the optimality criterion is usually based on geostatistics, and the applied methods are referred to as geostatistical design methods. These can be separated into two groups focusing on either parameter estimation or

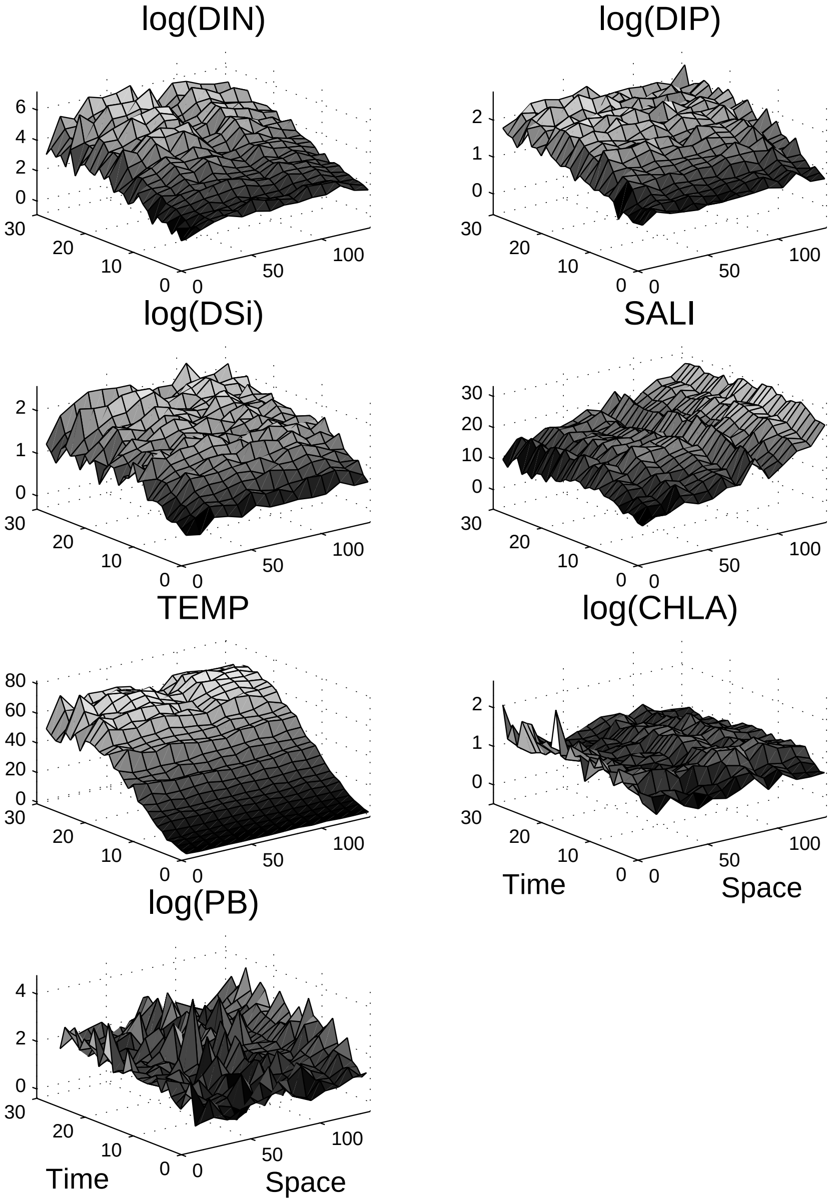

Figure 3.3: *Space-time semivariograms for the monitoring data in the Kattegat.*

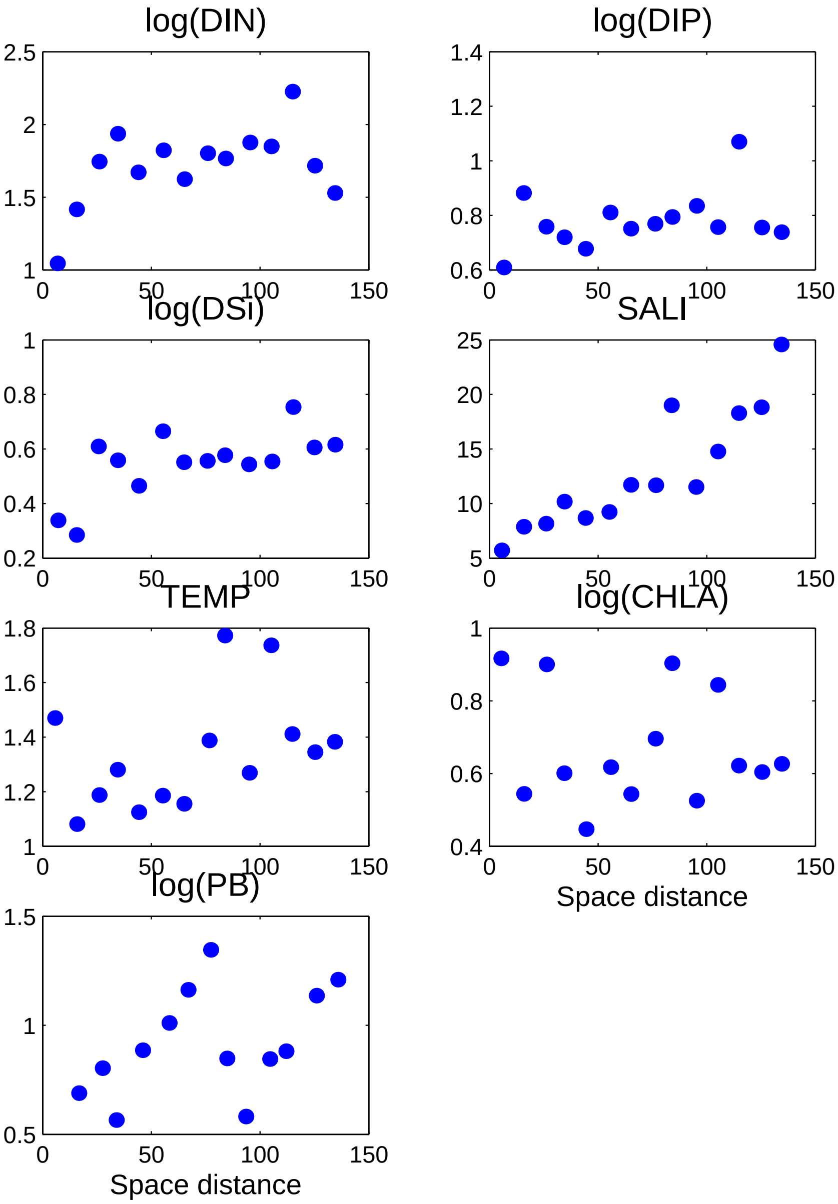

Figure 3.4: *Spatial semivariograms for the monitoring data in the Kattegat.*

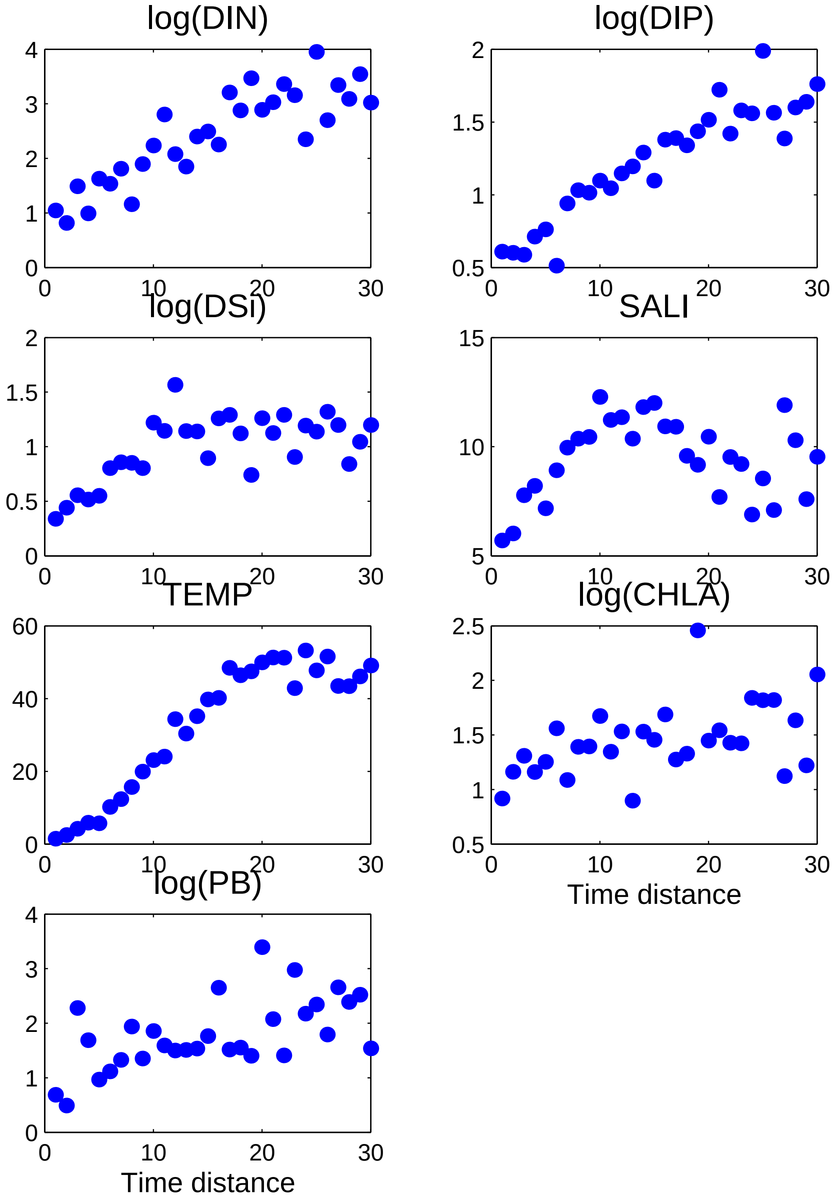

Figure 3.5: *Temporal semivariograms for the monitoring data in the Kattegat.*

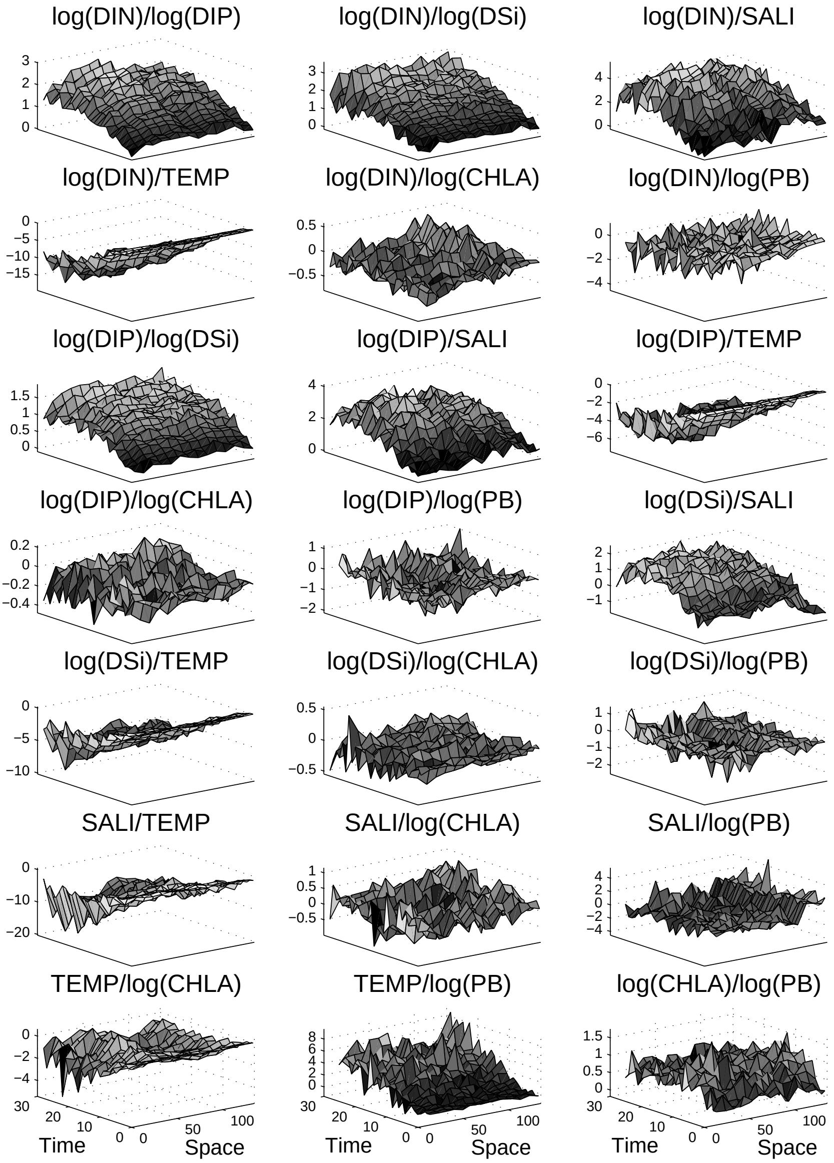

Figure 3.6: *Space-time cross-semivariograms for the monitoring data in the Kattegat.*

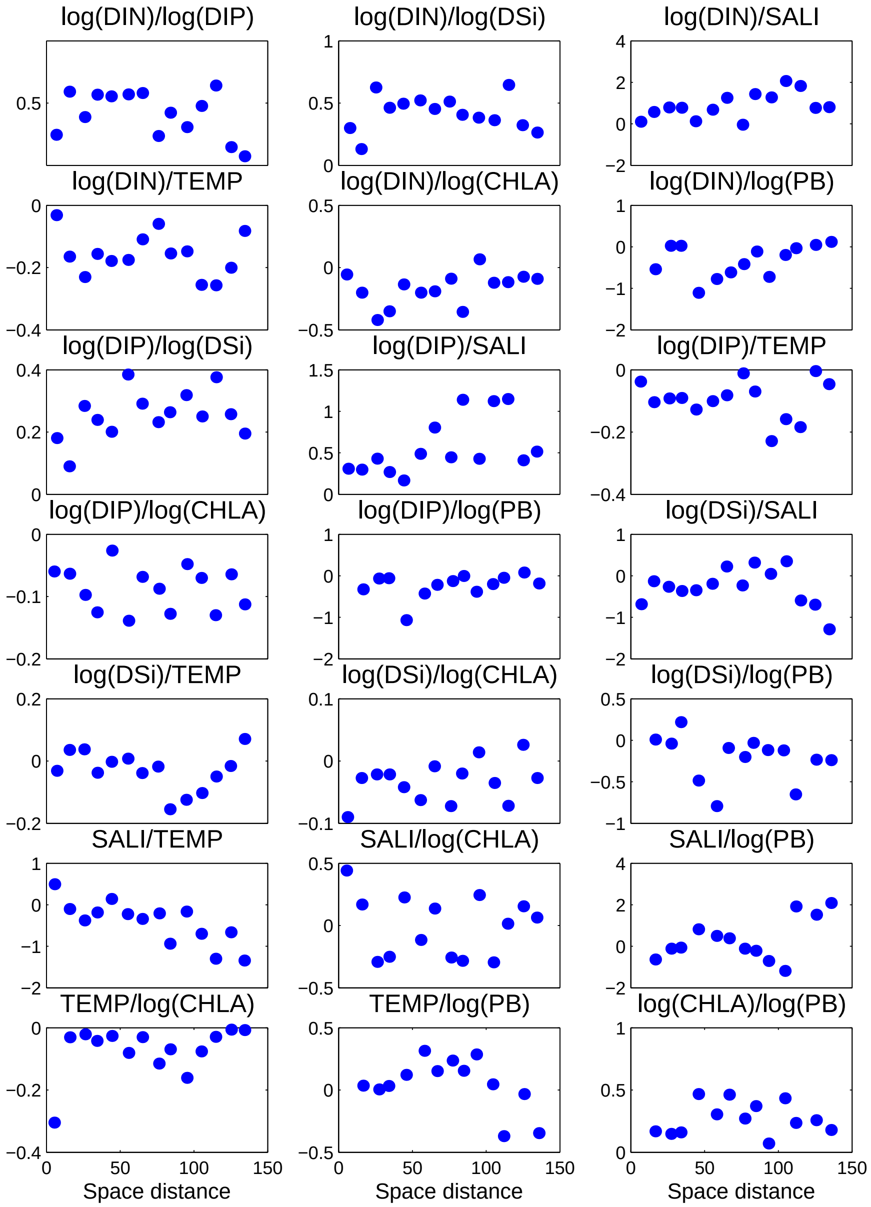

Figure 3.7: *Spatial cross-semivariograms for the monitoring data in the Kattegat.*

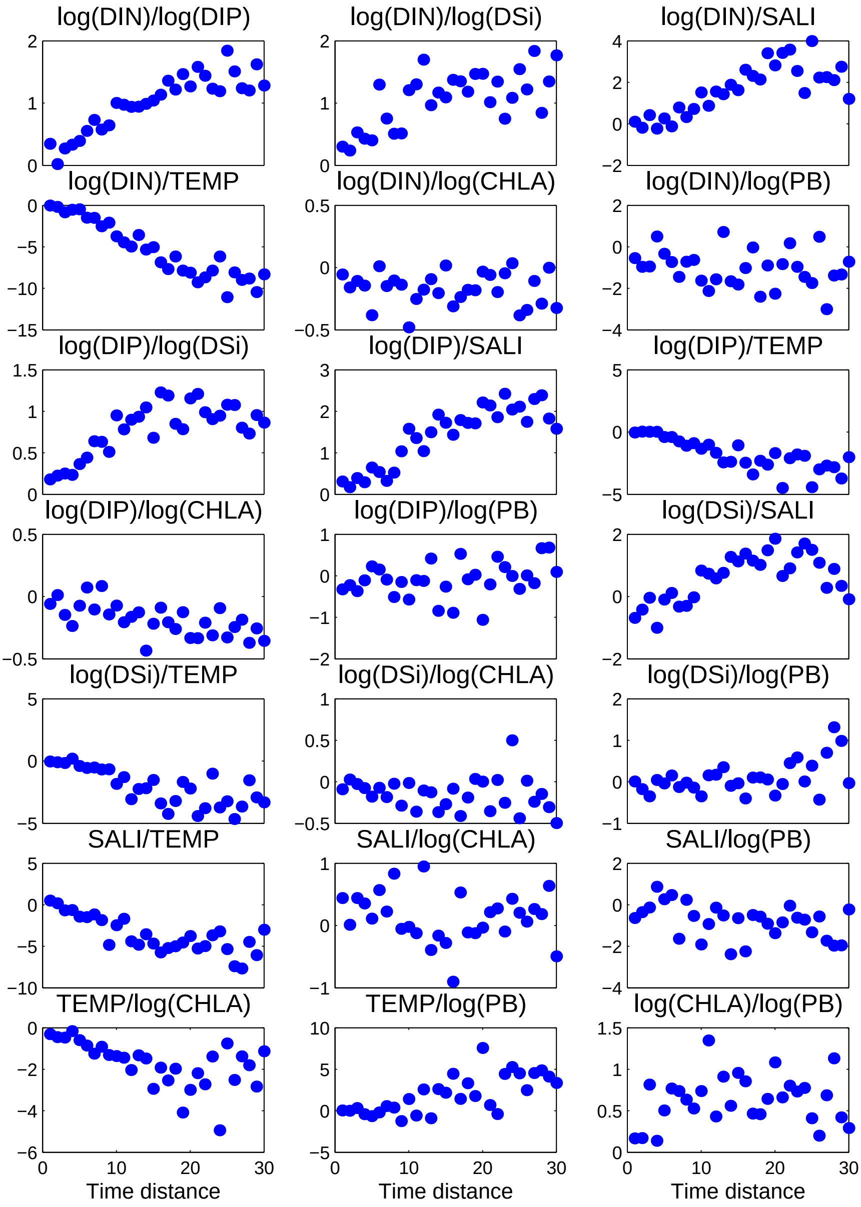

Figure 3.8: *Temporal cross-semivariograms for the monitoring data in the Kattegat.*

spatial prediction. A review of classical geostatistical design methods is given in paper E. This paper also describes how the two groups of methods can be combined. Such a combination is applied in paper F for reducing the number of monitoring stations in the Kattegat. In practice the design situation could be that a monitoring program has to be constructed from scratch, or that an existing monitoring program is to be revised, which is the situation with the monitoring program in the Kattegat. Diggle and Lophaven (2004) denoted these two design situations as prospective and retrospective, respectively. They suggest two classes of prospective designs (Figure 3.9). The $(k \times k, m, \alpha)$ lattice plus close pairs design consist of locations in a regular $k \times k$ lattice at spacing Δ together with a further m points, each of which is located uniformly at random within a disc of radius $\delta = \alpha\Delta$ whose centre is at a randomly selected lattice location. The $(k \times k, m, r \times r)$ lattice plus in-fill design consists of locations in a regular $k \times k$ lattice at spacing Δ together with further locations in a more finely spaced lattice within m randomly chosen cells of the primary lattice. Each in-filled lattice cell consists of an $r \times r$ lattice and therefore involves $r^2 - 4$ additional locations. The lattice plus in-fill design was used in Diggle et al. (1998).

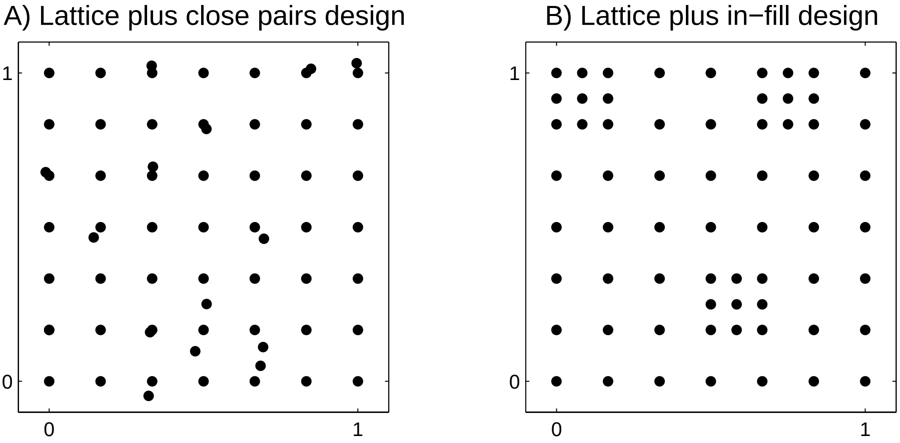

Figure 3.9: *Examples of a* $(7 \times 7, 15, 0.5)$ *lattice plus close pairs design, and a* $(7 \times 7, 3, 3 \times 3)$ *lattice plus in-fill design.*

A comparison of the efficiency of the suggested designs showed that the lattice plus close pairs design is a good design from a prediction point of view, whereas the performance of the lattice plus in-fill design is only slightly better than that of the regular lattice.

Chapter 4

Discussion

This chapter gives a discussion and an overview of the statistical methods applied in this thesis and the results obtained. Thus, questions like are the applied methods sufficient, which other statistical methods could be used, and what are the implications of the obtained results will be addressed. Section 4.1 focus on space-time modelling first from a methodological point of view, and afterwards in relation to the Kattegat. Section 4.2 focus on design of monitoring programs and is organised in a similar way. The statistical methods applied to the Kattegat dataset in this thesis are general, and with minor modifications they could equally well be applied within areas such as soil and air pollution.

4.1 Space-time modelling

Environmental monitoring datasets are very often sampled over time on a distinct number of locations in the area of interest. Usually sampling is done quite frequently, whereas the number of locations is relatively small. However, the Kattegat dataset, which is used in this thesis, is quite different from this general pattern. This consist of measurements from a lot of rarely sampled monitoring stations, compared to the temporal dynamics of the variables being measured. Consequently, the methodological starting point of space-time modelling in this thesis has been geostatistics rather than time series analysis. In this thesis a temporal resolution of one week has been used, and this resulted in a huge amount of missing data values for combinations of monitoring station and week. Hence, an important requirement of the space-time models being used, is that they are able to handle missing values.

4.1.1 Methodology

Different space-time modelling approaches are described and applied in papers B, C and D. The models are on the general decomposed form

$$Y(x,t) = m(x,t) + \epsilon(x,t) \tag{4.1}$$

where Y is the response variable, m is the deterministic mean component and ϵ the stochastic residual component. In the general setting all of these are functions of space $x = (x_1, x_2)$ and time t. In paper B the mean component is described by different levels of the two factors station and week, while the residual component accounts for spatial variation in individual weeks, i.e. when estimating parameters in the general linear model the covariance matrix is a kind of block-diagonal matrix. Due to the station-effect in the mean component it is not straight forward to apply the model to non-sampling locations in the Kattegat. Thus, paper B focus on modelling time series at the locations of monitoring stations by including information from surrounding stations in terms of the week-effect and the covariance matrix. One problem with this approach is that the number of parameters in the model is high, i.e. more than 300 parameters are included, which is primarily due to the week-effect. It is well-known that models with many parameters might cause over-fitting, i.e. it is not just the signal of interest that is modelled, but also some random fluctuations. In fact, some of the Figures in paper B indicate over-fitting, e.g. Figure B.6. Another methodological problem in paper B is that temporal correlation in data is not included.

These two weaknesses caused the development of the model approach in papers C and D, in which the station-effect is still included in the mean component while the week-effect is substituted by the sum of a year-effect and two sine-functions. Both spatial and temporal correlation are included in the residual component, and the spatial covariance structure is modelled by means of a separable space-time model. As shown in paper D the model can be extended to non-sampling locations by geostatistical modelling of the station-effect.

The basic form of the space-time models in this thesis is the same, i.e. given by (4.1). Parameters describing the residual component are modelled by means of the semivariogram, and prediction is based on classical geostatistics. Another type of space-time model, which recently have been widely used is the space-time

Kalman filter (Huang and Cressie, 1996; Mardia et al., 1998; Brown et al., 2001). This can be seen as an extension of traditional time series models, whereas the models in this thesis are extensions of classical geostatistics. The idea is to treat data as spatially correlated time series in discrete time, and write the model in state space form. For a given time t and n locations we have

$$\begin{array}{rrcll} \text{Observation:} & Y_t & = & IA_t + (X_t\beta) + \sigma_z^2 Z_t, & Z_t \sim N(0,1) \\ \text{State:} & A_t & = & \phi A_{t-1} + H\eta_t, & \eta_t \sim N(0,1) \end{array} \tag{4.2}$$

where Y_t is a $n \times 1$ vector of observations, A_t is a $n \times 1$ vector of states, ϕ is a parameter which in Brown et al. (2001) is a scalar, I is the identity matrix, H is an $n \times n$ matrix of spatial interactions, β is a $p \times 1$ vector of parameters, and X_t is a $n \times p$ matrix, where p is the number of parameters. Thus, each row in X_t correspond to a location. In this setting spatial correlation is included by making the H-matrix non-diagonal. Furthermore, parameter estimation is based on maximum likelihood rather than the semivariogram, and predictions $\mathrm{E}[A_t|Y_1,\ldots,Y_t]$ are computed by the Kalman filter rather than kriging. This approach is interesting because it can handle missing values, it is relatively simple to implement, and computations are expected to be fast. It would be an interesting study to compare the performance of the space-time Kalman filter with the space-time models used in this thesis.

4.1.2 The Kattegat

In relation to the Kattegat, this thesis have aimed at developing model approaches which are simple and utilize a high amount of data, by incorporating correlation in the models. Space-time models with different residual components were tested and compared in paper B. The comparison was done by means of cross validation, and showed that better predictions of DIN were obtained when introducing spatial correlation in the residual component. The best of the tested methods seem to fit observations quite well. In paper D a model approach, which could be applied to any location in space and point in time, was presented. Results were presented for DIN as time series for three different locations, and as maps for four different weeks. The results obtained can be interpreted biologically and physically. Thus, for improving reporting the state of the marine environment in the Kattegat, these models have proved very useful. As mentioned elsewhere, modelling results could also be applied as forcing functions for deterministic, hydrodynamic ecosystem models, and thereby advance the knowledge of the biochemical processes in the marine environment, and reduce uncertainties of regional nutrient and carbon budgets. In this thesis the performance of the applied models is neither measured in terms of a quantitative reduction in the uncertainties of nutrient and carbon budgets, nor as a quantitative advance in the knowledge of biochemical processes. Hence, the coupling to the marine processes is not made directly, and the mentioned applications of the obtained modelling results should therefore only be seen as suggestions.

4.2 Design of monitoring programs

When designing environmental monitoring networks it is desirable to formulate an optimality criterion based on the objective of monitoring. Many monitoring programs, and in particular the Danish marine monitoring program, do not have one single objective, or the objective can only be formulated in rather broad terms like ”reporting the state of the environment”. Thus, it is a general problem in environmental monitoring that the question ”why are we monitoring ?” is not clearly specified (Ward et al., 1986).

The problem of where to take samples is highly relevant in connection to the Danish marine monitoring program in the Kattegat, because the monitoring network is characterised by having a lot of monitoring stations at which only few samples are taken. The most frequently sampled stations in the Kattegat are sampled biweekly, which is due to the fact that monitoring is conducted by traditional shipboard sampling. Compared to the highly fluctuating concentrations of for example DIN over time, the temporal resolution of data is not sufficient. Hence, a natural thing to do when revising the Danish marine monitoring program would be to reduce the number of monitoring stations.

4.2.1 Methodology

The thesis deals with this problem from a statistical point of view, focusing on designing monitoring networks based on which geostatistics can be successfully applied. Thus, the idea of using statistics for design of monitoring networks is to go beyond heuristic rules, such as appointing a sampling location to be representative for a larger area. The results obtained should not be seen as the final answer to the design problem, but should be used as a decision tool in cooperation with other experts.

Prior to this thesis a number of studies have addressed this problem (see paper E). These focus on designing monitoring networks which are optimal for estimating the parameters in a geostatistical model or for computing spatial predictions. The resulting networks are quite different, i.e. those optimal for estimation consist of clustered sampling sites, while those optimal for computing spatial predictions consist of sampling sites where distances between neighboring points are large. However, it is well known that an overall efficient design is obtained by combining these two conflicting design criteria, see Müller (2001), who also show some rather ad hoc ways of doing this (see paper E). One of the main contributions of this thesis is that a statistically more correct way of combining design criteria is through the Bayesian paradigm (see section 3.1.6 and paper F). The idea has been to formulate a design criterion based on the variance of the predictive distribution, in which parameter uncertainties are included.

There are many good things to say about the Bayesian design approach. First of all, it is very flexible and general, and therefore it can be applied to a great variety of situations. For example, differences in the mean value, can be accounted for by including a polynomial trend. Anisotropy can be included, and the approach can in principle be extended to designs for non-Gaussian data, e.g. counts, either through data-transformation or through models formulated as a combination of generalized linear and geostatistical models (Diggle et al., 1998; Christensen, 2002). On the other hand, the Bayesian design approach leaves some open questions which need further investigation. For example, a necessary consequence of working within the Bayesian paradigm is that the choice of design depends on the choice of prior. This dependence remain unexplored, but it is expected that when a relatively informative prior is used, there is less need to choose a design which provides information on the parameters, and consequently less clustered designs are optimal, and vice versa. The Bayesian approach is computational intensive but can easily be applied to situations where points are to be deleted from an existing monitoring network. For example Figure E.5 in paper E was computed in only a few minutes. For design situations where points are to be added to an existing network, or where a new network is to be constructed the time for computation increases dramatically, see Diggle and Lophaven (2004) who due to this fact only compared a limited number of different designs without claiming that they were the optimal solutions. Work by Rue (2001); Rue and Tjelmeland (2002) is leading to substantial gains in the speed of computations for geostatistical models. Another area which needs further investigation is how to compute designs in the multivariate case. One solution to this could be to apply one of the design approaches in paper E to the principal components.

4.2.2 The Kattegat

In relation to the Kattegat the Bayesian design approach was applied to reduce the number of monitoring stations from the network of 31 stations, which are currently intensively monitored (paper F). This design situation is complicated by the fact that data are measured in both space and time. Before applying the design approach the seasonal variation was removed to give residuals with zero mean. Based on these the spatial semivariogram $\gamma_{xt}(u_x, u_t) = \gamma_{xt}(u_x, 0)$ was estimated. Thus, the design approach was based on a model describing the spatial variation, from which data can be simulated, rather than actual data values. The approach switches in 1000 runs between simulating data values in the design points and predicting values and variances in 103 points covering the Kattegat area. A design criterion to be minimised was formulated as a function of the prediction variances. Although this approach is computationally intensive, i.e. it takes a few days to compute the map in Figure F.6, it is better than the other methods described in paper E, because it addresses the problem of parameter uncertainty. As described in paper F the selection of monitoring stations to be removed is, at least in the beginning of the selection process, to some degree random. Thus, the main result of this design application is not which monitoring stations to delete from the network, but more the pattern of the final monitoring network, i.e the approach ensures that the final network consist of some monitoring stations close together in clusters, and some allocated for prediction anywhere in the Kattegat. Another important result of paper F is the number of stations which can be removed (Figure F.6). The paper concluded that the current network can be reduced to 14 monitoring stations with only a marginal increase in the design criterion, i.e. the prediction variance. Thus, one way of revising the current monitoring network in the Kattegat could be to reduce the number of monitoring stations to approximately 14, and monitor these frequently to obtain an increased knowledge about the temporal dynamics. This strategy could be supplied by intensive spatial sampling in some periods of the year.

In the above design situation the answer to the question ”why are we monitoring dissolved inorganic nitrogen in the Kattegat ?” is assumed to be ”because we want to determine the spatial distribution of it ?”, which is of course a strong simplification of the true answer. On the other hand, this is one of the answers, and it is probably an intermediate answer to objectives like improving reporting the state of the marine environment in the Kattegat, determining regional nutrient and carbon budgets and increasing the knowledge of the chemical and biological processes in the marine environment.

Another issue of monitoring the marine environment is that traditional shipboard sampling, which is associated with large costs for personnel and equipment, might in the future be substituted by new technologies such as continuous measurements from moored buoys, ships-of-opportunity and remote sensing. However, there will still be a need for traditional sampling in calibrating these methods.

Chapter 5

Conclusion

This thesis describes methods for modelling space-time phenomena. The methods are applied to data from the Danish marine monitoring program in the Kattegat, measured in the five-year period 1993-1997. The proposed model approaches are characterised as relatively simple methods, which can handle missing data values and utilize the spatial and temporal correlation in data. The modelling results improve the reporting on the state of the marine environment in the Kattegat.

The thesis also describes methods for designing monitoring networks based on which geostatistics can be successfully applied. There has been a need to combine existing design methods, which has motivated the development of a Bayesian geostatistical design approach. This approach focus on constructing monitoring networks which are efficient for computing spatial predictions, while taking the uncertainties of the parameters in the geostatistical model into account. When applied to the Kattegat the approach shows that the current monitoring network can be reduced to 14 stations, with only a minor increase in the prediction variances.

Finally, further applications of the space-time model approaches, and recommendations on design of monitoring programs are given. Space-time modelling results could serve as a surrogate model for a more advanced and eventually computationally intensive model, i.e. it could be used to correct the advanced model or to speed up computations in this. The Bayesian design approach suggest that the number of monitoring stations in the Kattegat could be reduced substantially, which should be accompanied by an increase in the sampling frequency at the remaining stations, and intensive spatial sampling in some periods of the year. The space-time model approaches and geostatistical design methods used in this thesis are generally applicable, i.e. with minor modifications they could equally well be applied within areas such as soil and air pollution.

Appendix A

Methods for estimating the semivariogram

The paper is published in:

Lophaven, S., Carstensen, J., and Rootzén, H. (2002). Methods for estimating the semivariogram. In *Symposium i Anvendt Statistik*, p. 128-144. Århus, Denmark, January 21-23, 2002.

Methods for estimating the semivariogram

Søren Lophaven[1], Jacob Carstensen[2], and Helle Rootzén[1]

[1] Informatics and Mathematical Modelling, Technical University of Denmark
[2] Department of Marine Ecology, National Environmental Research Institute of Denmark

Abstract

Modelling spatial variability, typically in terms of the semivariogram, is of great interest when the objective is to compute spatial predictions of parameters measured in space. Such parameters could be rainfall, temperature or concentrations of polluting agents in aquatic environments. In the existing literature various methods for modelling the semivariogram have been proposed, while only a few studies have been made on comparing different approaches. In this paper we compare eight approaches for modelling the semivariogram, i.e. six approaches based on least squares estimation of an experimental semivariogram, as well as maximum likelihood and restricted maximum likelihood estimation. The comparison is made by simulating spatial data with a known covariance structure, and comparing the "true" parameters with those computed. The comparison showed that maximum likelihood and restricted maximum likelihood performed better than the least squares approaches. We also modelled the performance as a function of the sill/nugget effect - ratio. This showed that the advantage of using maximum likelihood is much greater when this ratio is high, while for small ratios the improvement of estimation is insignificant. Taking into account the complexity of the maximum likelihood approaches, we recommend only to use these methods when the sill/nugget effect - ratio is high. We also applied maximum likelihood and least squares estimation to a real dataset, containing measurements of salinity at 71 sampling stations in the Kattegat basin. This showed that the calculation of spatial predictions is insensitive to the choice of estimation method, but also that the uncertainties of predictions were reduced when applying maximum likelihood.

KEY WORDS: *Experimental semivariogram, semivariogram model, least squares estimation, maximum likelihood, restricted maximum likelihood.*

A.1 Introduction

Many of the various parameters measured to describe the environment are functions of space. Examples of such parameters could be measurements of rainfall, temperature and sunshine at climatological stations, or concentrations of polluting agents in a lake or in the sea, at a number of sampling stations. It is of great interest to know the magnitude of environmental parameters at any location in space, however, it is usually impossible, both financially and operationally, to sample at all possible locations in the area. Instead samples are taken at a finite number of locations, and mathematical or statistical models are applied to compute the spatial distribution.

Kriging is the most widely used statistical approach for describing the spatial distribution of a parameter measured in space. Different variants of kriging, e.g. ordinary kriging, universal kriging and cokriging exist. The common idea of these approaches is to weight the point observations in a way that minimizes the squared prediction error, and the computation of the weights is therefore based on the spatial varibility of data.

A lot of studies have been made on applying and comparing different variants of kriging, see e.g. Brus et al. (1996); Laslett (1994); Asli and Marcotte (1995); Hosseini et al. (1993), but there seems to be a lack of studies concentrating on estimating the spatial variability, which for kriging is expressed in the form of a semivariogram. A comprehensive work within this area was done by Zimmerman and Zimmerman (1991). They used Monte Carlo simulations for comparison, and found that the nonparametric least squares methods, see section A.3.1, for estimating the semivariogram perform as well or nearly as well as more computationally demanding parametric methods like maximum likelihood and restricted maximum likelihood, see section A.4. Zimmerman and Zimmerman (1991) only considered regular grid structures, however, we believe it is very relevant to include irregularly grids, because environmental data are almost always irregularly sampled in space. Furthermore, Zimmerman and Zimmerman (1991) only compared the semivariogram estimators for different values of one of the model parameter. In our study all three model parameter, i.e. the sill, range and nugget effect, are varied. Swallow and Monahan (1984) focuse on comparison of maximum likelihood and restricted maximum likelihood for estimation of variance components, while McGilchrist (1989) applied and compared the same estimators in regression models. Both studies found that the restricted maximum likelihood estimator can have a significantly smaller bias.

In this paper eight approaches will be tested and compared. Some of these are the same as those considered by Zimmerman and Zimmerman (1991), while others were not included in their comparison. The different approaches can be separated into two groups, one containing maximum likelihood estimators and the other containing least squares estimation of an experimental semivariogram. This means that the second group consists of estimation methods on two levels, i.e. first the experimental semivariogram is estimated and afterwards the parameters of a proposed semivariogram model are estimated by least squares fitting of the experimental semivariogram. Three different methods for calculating the experimental semivariogram are considered, see section A.2, and each of these can be fitted by ordinary or weighted least squares, as described in section A.3.1. This gives six nonparametric approaches for estimating the semivariogram. The last two approaches are maximum likelihood and restricted maximum likelihood, see section A.4, i.e. a total of eight approaches are considered. As indicated in Figure A.1, the two maximum likelihood methods are not based on an experimental semivariogram. Spatial data for the estimation is generated as a Gaussian random field, with a given sample size, grid structure, and covariance structure, as shown in Figure A.1. The comparison is made by computing 100 realizations of a random field and applying the eight approaches to each realization. Afterwards the mean and variance of the estimations can be calculated, as well as the bias, which is found as the difference between the mean and "true" value of the Gaussian random field.

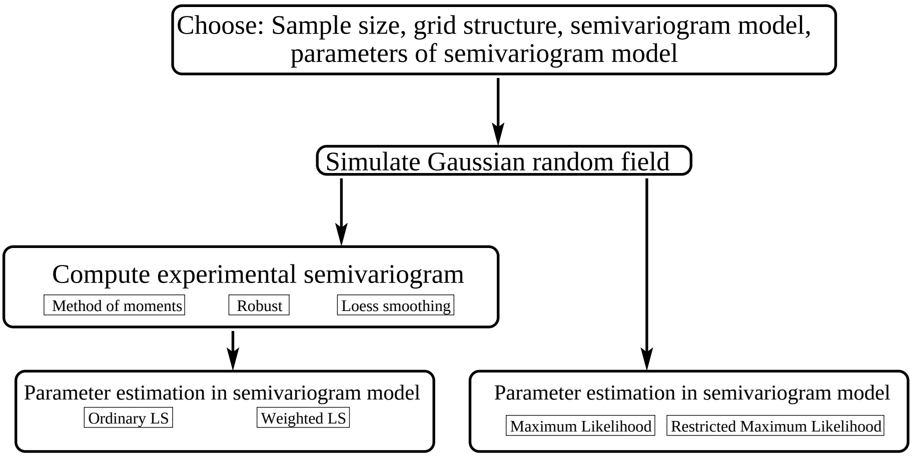
Choose: Sample size, grid structure, semivariogram model, parameters of semivariogram model
Simulate Gaussian random field
Compute experimental semivariogram
Method of moments
Robust
Loess smoothing
Parameter estimation in semivariogram model
Ordinary LS
Weighted LS
Parameter estimation in semivariogram model
Maximum Likelihood
Restricted Maximum Likelihood

Figure A.1: *The principle of estimating the semivariogram of a Gaussian random field, given a given sample size, grid structure, and covariance structure.*

A.2 Estimating the experimental semivariogram

The spatial patterns of a region can be characterized quantitatively by the semivariogram, which is defined as

$$\gamma(d) = \frac{1}{2}\mathrm{Var}[Z(s+d) - Z(s)] = \frac{1}{2}E[(Z(s+d) - Z(s))^2] \qquad \text{(A.1)}$$

where Z is a stochastic variable and d is the distance between observations measured in space. The typical form of the semivariogram is illustrated in Figure A.2. The parameters of the semivariogram model are also illustrated in the Figure. The nugget effect describes the fact that measurements taken at locations infinitely close to each other are different. This is caused by measurement error and micro-variability. The range describes the distance beyond which data are assumed to be uncorrelated. The semivariance corresponding to the range is the total variability, and this valus minus the nugget effect is called the sill. The semivariogram is the most commonly used function for describing spatial variability, but the covariance- or correlation function could also be applied. The main reason for using the semivariogram is that it does not depend on the mean value of Z.

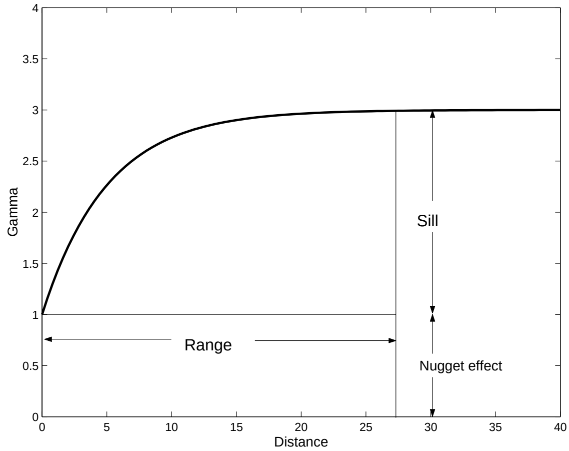

Figure A.2: *Example of a typical semivariogram. The horizontal axis plots the distance between pairs of data, while the vertical axis plots the semivariance.*

A.2.1 The method of moments estimator

The simplest and most commonly used estimator of the semivariogram is the method of moments, which is given by

$$\hat{\gamma}(d) = \frac{1}{2N(d)} \sum_{l=1}^{N(d)} [z(s_l + d) - z(s_l)]^2 \tag{A.2}$$

where $\hat{\gamma}$ is called the experimental semivariogram, and $N(d)$ is the number of data pairs with a separation distance d. With n observations, $z(s_l)$ $l = 1, \cdots, n$, of Z, the number of pairs becomes $\frac{n(n-1)}{2}$. When data are irregularly located in space, which is usually the case in environmental studies, (A.2) takes the form

$$\hat{\gamma}(d_k) = \frac{1}{2N(k)} \sum_{l=1}^{N(k)} [z(s_1^l) - z(s_2^l)]^2 \tag{A.3}$$

In this case the distances are grouped into k intervals (lags), characterized by the midpoint of the interval, d_k, and a distance tolerance, ϵ. We have

$$|s_1^l - s_2^l| \in [d_k - \epsilon, d_k + \epsilon] \tag{A.4}$$

The method of moments has some disadvantages, which were described by Cressie and Hawkins (1980). The first objection is that the method is not robust to outlying values of Z. Secondly, if Z is normally distributed, then the distribution of $(Z(s_1) - Z(s_2))^2$ is of the form $2\gamma(d)\chi_1^2$, and it is well-known that the distribution of $X \in \chi_1^2$ is highly skewed.

A.2.2 The robust estimator

To overcome the problems of the method of moments mentioned above, Cressie and Hawkins (1980) suggested to compute more robust estimations by using power transformations to transform the problem to one of estimating a center of symmetry. It was found that the distribution of $X^{1/4}$ is nearly symmetric, if Z is normally distributed, see Figure A.3. This means that sample averages of $|Z(s_1) - Z(s_2)|^{1/2}$ are better behaved than those of $(Z(s_1) - Z(s_2))^2$, and leads to the suggestion of using

$$\hat{\gamma}(d_k) = \frac{\frac{1}{2}\left(\frac{1}{N(k)} \sum_{l=1}^{N(k)} |z(s_1^l) - z(s_2^l)|^{1/2}\right)^4}{\left(0.457 + 0.494/N(k)\right)} \tag{A.5}$$

as a robust estimator of the experimental semivariogram. It can be shown that the denominator in (A.5) is a bias correction (Cressie and Hawkins, 1980).

A.2.3 Smoothing the semivariogram cloud

In addition to the computation of the experimental semivariogram, the so-called semivariogram cloud should be computed and investigated. In the semivariogram cloud, one point is plotted for each pair of data, instead of combining data pairs into one value, representing the semivariance for a single lag. The

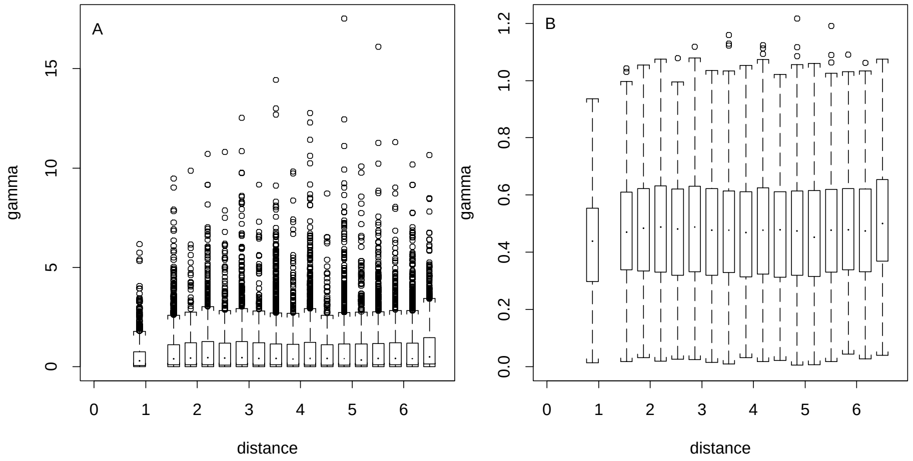

Figure A.3: *Boxplots for data pairs. Data are generated as a Gaussian random field with spherical covariance structure. The parameters are: Range=2, sill=1, nugget effect=0. A)* $0.5(z(s_1) - z(s_2))^2$. *B)* $0.5|z(s_1) - z(s_2)|^{1/2}$. *Note the difference in scale.*

distance, d_{ij}, between two data points, s_i and s_j, is plotted on the x-axis, while an estimate of $\text{Var}(Z(s_i) - Z(s_j))$ corresponding to d_{ij}, is plotted on the y-axis. The semivariogram cloud can be used to identify outliers in data. Two estimates of $\text{Var}(Z(s_i) - Z(s_j))$ can be considered, these are:

Method of moments: $0.5(z(s_i) - z(s_j))^2$

Robust: $0.5|z(s_i) - z(s_j)|^{1/2}$

Semivariogram clouds computed using the two different estimates are shown in Figure A.4. The idea is to use locally weighted regression (LOESS), for smoothing the semivariogram cloud, and use the smoothed curve as an estimate of the experimental semivariogram. In this study LOESS will only be used to smooth the method of moments semivariogram cloud, while the robust semivariogram cloud is not considered. LOESS is a nonparametric estimation method, see e.g. Cleveland (1979, 1988), where the relationship between a dependent variable, γ_i, and an independent variable, d_i, is

$$\gamma_i = g(d_i) + \epsilon_i \tag{A.6}$$

where g is the regression function and ϵ_i are independent normally distributed variables with a mean 0 and a variance σ_ϵ^2. In LOESS data within a neighbourhood around a point d can be approximated by fitting a regression function,

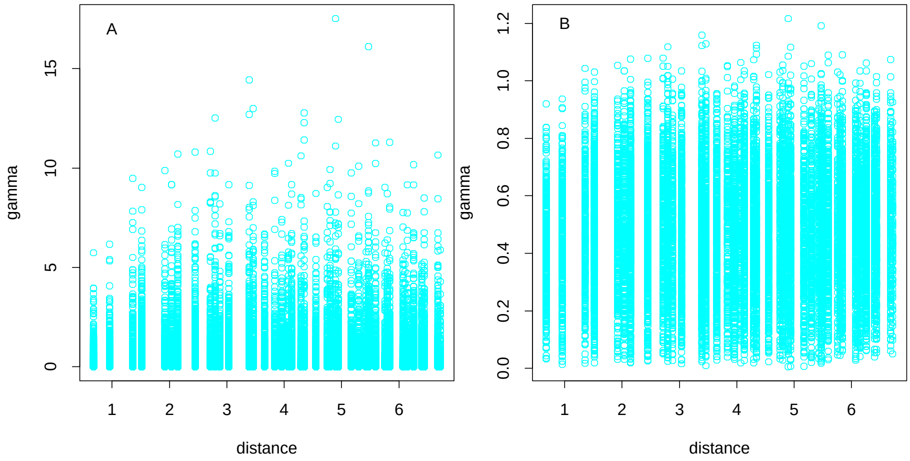

Figure A.4: *Semivariogram cloud for data pairs. Data are generated as a Gaussian random field with spherical covariance structure. The parameters are: Range=2, sill=1, nugget effect=0. A)* $0.5(z(s_1) - z(s_2))^2$*. B)* $0.5|z(s_1) - z(s_2)|^{1/2}$*. Note the difference in scale.*

$\hat{g}(d)$, to data. The fitting is done by weighted least squares, where points in the neighbourhood are weighted according to the distance from d, i.e. points close to d are given a higher weight than those further away. The size of the neighbourhood is chosen by the value of the nearest neighbour bandwidth, $f = q/n$, which is a fraction of the total number of observations n.

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

$$W(u) = \begin{cases} (1 - |u|^3)^3 & |u| < 1 \\ 0 & |u| > 1 \end{cases} \tag{A.7}$$

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

$$w_i(d) = W\left(\frac{\parallel d - d_i \parallel}{\mathrm{dist}(d)}\right) \tag{A.8}$$

where $\parallel d - d_i \parallel$ is the Euclidean distance between d and d_i, and dist(d) is the distance of the q-nearest d_i to d. By combining (A.7) and (A.8) it is seen that the weights $w_i(d)$ decrease when d_i increase in distance from d. In this study polynomials of second order will be used in the estimation, i.e. the estimated value $\hat{\gamma}_i$ can be written as

$$\hat{\gamma}_i = \hat{g}(d_i) = \beta_0 + \beta_1 d_i + \beta_2 d_i^2 \tag{A.9}$$

The values of β are found by minimizing (A.10).

$$\sum_i w_i(d)(\gamma_i - \hat{\gamma_i})^2 \tag{A.10}$$

When applying LOESS the bandwidth f has to be chosen. This can be done using apriori information or optimization methods like Akaike's information criterion (AIC). In the case of a parametric regression AIC is given as

$$AIC = n \log \hat{\sigma_\epsilon}^2 + 2(p+1) \tag{A.11}$$

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

$$\hat{\sigma_\epsilon}^2 = \frac{1}{n}\sum_{i=1}^{n}(\gamma_i - \hat{\gamma_i})^2 \tag{A.12}$$

It is seen that AIC is a function of the goodness of the fit and the complexity of the model, i.e. the criterion has the form

$$\log(\hat{\sigma_\epsilon}^2) + \psi(\boldsymbol{L}) \tag{A.13}$$

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

$$\hat{\boldsymbol{\gamma}} = \boldsymbol{L}\boldsymbol{\gamma} \tag{A.14}$$

The smoothing parameter is selected as the one that minimizes the criterion. 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

$$\begin{aligned} AIC &= n \log \hat{\sigma_\epsilon}^2 + 2(\text{trace}(\boldsymbol{L}) + 1) \qquad \text{where} \\ \hat{\sigma_\epsilon}^2 &= \frac{1}{n}\sum_{i=1}^{n}(\gamma_i - \hat{\gamma_i})^2 = \frac{\boldsymbol{\gamma}^T(\boldsymbol{I}-\boldsymbol{L})^T(\boldsymbol{I}-\boldsymbol{L})\boldsymbol{y}}{n} \end{aligned} \tag{A.15}$$

Figure A.5 shows estimation of the experimental semivariogram using the three methods described in section A.2

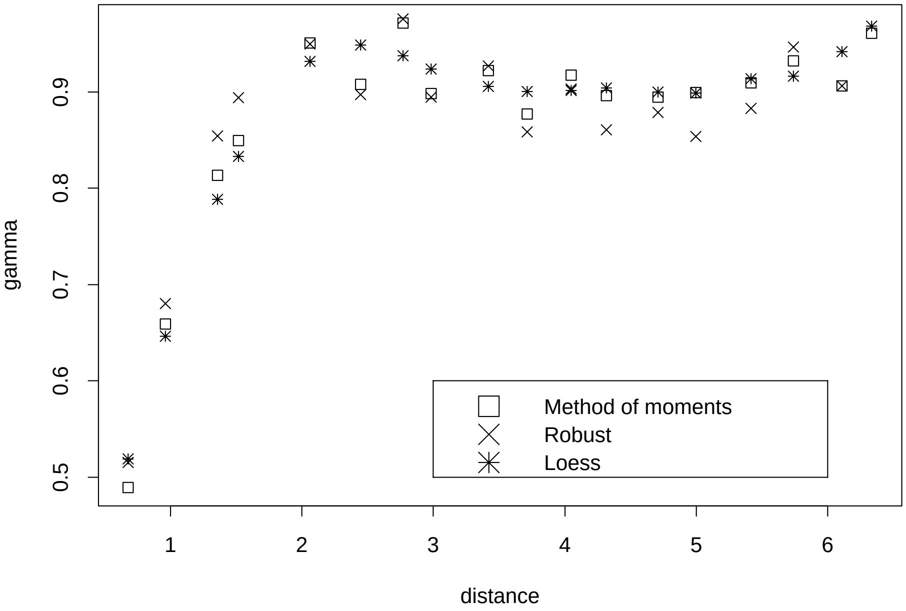

Figure A.5: *Experimental semivariograms computed by the three methods described above. Data are generated as a Gaussian random field with spherical covariance structure. The parameters are: Range=2, sill=1, nugget effect=0.*

A.3 Estimating the parameters of the semivariogram model

Spatial predictions can be calculated using the following model

$$Z(s) = m(s) + \epsilon(s) \tag{A.16}$$

where s is the location given by (x, y), $Z(s)$ is a random function, $m(s)$ is called the trend, and $\epsilon(s)$ is the residual. The expectation of $Z(s)$ is $E\{Z(s)\} = m(s)$, and the trend is modelled as a linear combination of known functions, usually low-order polynomia, multiplied by unknown coefficients, i.e.

$$m(s) = \sum_{j}^{P} f_j(s)\beta_j \tag{A.17}$$

which in matrix notation can be written as

$$\begin{aligned} \boldsymbol{\mu} &= \boldsymbol{X\beta} \\ \boldsymbol{\mu}' &= [m(s_1), m(s_2), \cdots, m(s_n)] \end{aligned} \tag{A.18}$$

where $\boldsymbol{X}$ is a matrix of known regressors and $\boldsymbol{\beta}$ is a vector of unknown coefficients. As an example $\boldsymbol{X}$ and $\boldsymbol{\beta}$ for a quadratic polynomial drift, P=6, and n observations of Z, are given as

$$\begin{aligned} \boldsymbol{X} &= \begin{bmatrix} 1 & x_1 & y_1 & x_1^2 & y_1^2 & x_1y_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & y_n & x_n^2 & y_n^2 & x_ny_n \end{bmatrix} \\ \boldsymbol{\beta}' &= [\beta_1, \beta_2, \cdots, \beta_6] \end{aligned} \tag{A.19}$$

Prior to the computation of spatial predictions, the experimental semivariogram has to be replaced by a parametric semivariogram model, $\gamma^*(d, \boldsymbol{\theta})$. The parameters, $\boldsymbol{\theta}$, of the model can be estimated by least squares fitting of the experimental semivariogram, while maximum likelihood or restricted maximum likelihood estimation is not based on any prior computation of the experimental semivariogram. It is not just any function which can be a valid semivariogram model. $-\gamma^*(d, \boldsymbol{\theta})$ must be conditionally positive definite, i.e. for all $s_1, \cdots, s_n \in R^2$, and for all $\lambda_1, \cdots, \lambda_n \in R$, n coefficients satisfying $\sum \lambda_i = 0$, then

$$-\sum_i \sum_j \lambda_i \lambda_j \gamma^*(s_i - s_j, \boldsymbol{\theta}) \geq 0 \tag{A.20}$$

Moreover $\gamma^*(d, \boldsymbol{\theta})$ must increase less rapidly than $|d|^2$ for $|d| \to \infty$, i.e.

$$\lim_{|d| \to \infty} \frac{\gamma^*(d, \boldsymbol{\theta})}{|d|^2} \to 0 \tag{A.21}$$

The conditions (A.20) and (A.21), must be fulfilled in order to ensure that the kriging equations have one, and only one, stable solution. Two semivariogram models, which are valid according to (A.20) and (A.21) will be considered in this study. These are the spherical model given by

$$\gamma^*(d, \boldsymbol{\theta}) = \begin{cases} 0 & d = 0 \\ C_0 + C_1\Big(\frac{3}{2}\frac{d}{R} - \frac{1}{2}\frac{d^3}{R^3}\Big) & 0 < d < R \\ C_0 + C_1 & d \geq R \end{cases} \tag{A.22}$$

and the exponential model, as shown below

$$\gamma^*(d, \boldsymbol{\theta}) = \begin{cases} 0 & d = 0 \\ C_0 + C_1\Big(1 - \exp(-\frac{d}{R})\Big) & 0 < d \end{cases} \tag{A.23}$$

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

A.3.1 Least squares methods

Given the experimental semivariogram, $\hat{\gamma}(d)$ at k distances, d_k, we want to fit a parametric semivariogram model, $\gamma^*(d, \boldsymbol{\theta})$, e.g. the exponential or spherical, where $\boldsymbol{\theta}$ is the parameter vector containing the sill, range and nugget effect. $\boldsymbol{\theta}$ can be found by non-linear ordinary least squares (OLS) regression, in which $\boldsymbol{\theta}$ is chosen to minimize

$$\{\hat{\boldsymbol{\gamma}}(d) - \boldsymbol{\gamma}^*(d, \boldsymbol{\theta})\}^T \{\hat{\boldsymbol{\gamma}}(d) - \boldsymbol{\gamma}^*(d, \boldsymbol{\theta})\} \tag{A.24}$$

Another and more efficient way of finding $\boldsymbol{\theta}$ is to apply weighted least squares (WLS) regression, i.e. choose $\boldsymbol{\theta}$ to minimize

$$\{\hat{\boldsymbol{\gamma}}(d) - \boldsymbol{\gamma}^*(d, \boldsymbol{\theta})\}^T \boldsymbol{W}(\boldsymbol{\theta})^{-1} \{\hat{\boldsymbol{\gamma}}(d) - \boldsymbol{\gamma}^*(d, \boldsymbol{\theta})\}. \tag{A.25}$$

Here $\boldsymbol{W}(\boldsymbol{\theta})$ is a diagonal matrix containing the variances of $\hat{\boldsymbol{\gamma}}(d)$. Cressie (1985) suggested to choose $\boldsymbol{\theta}$ to minimize

$$\sum_j |N(h_j)| \big(\frac{\hat{\gamma}(d)}{\gamma^*(d, \boldsymbol{\theta})} - 1\big)^2 \tag{A.26}$$

This criterion is an approximation to WLS, and takes into account the number of data pairs corresponding to single lags, and uses this for describing the uncertainty of the estimation of $\gamma^*(d, \boldsymbol{\theta})$. It does not account for correlation between lags.

Figure A.6 shows ordinary and weighted least squares estimation of the experimental semivariogram, computed by the method of moments.

A.4 Maximum likelihood estimation

In this section two methods for estimating the parameters of the semivariogram model are described. These are not, like the least squares methods, based on an experimental semivariogram.

A.4.1 Maximum likelihood

As an alternative to the non-parametric least squares methods described in section A.3.1, parametric methods as maximum likelihood (ML) and restricted maximum likelihood (REML) can be applied for estimating the parameters of

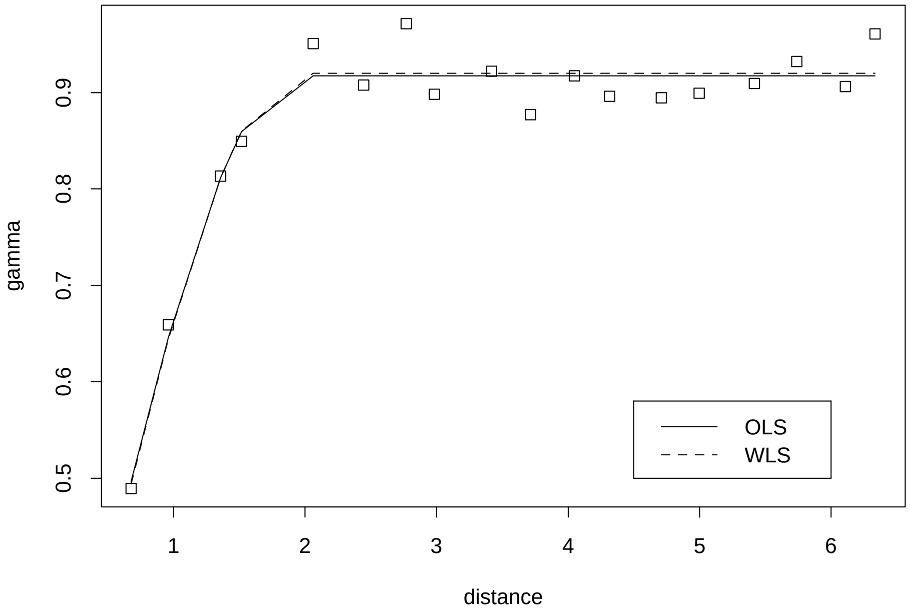

Figure A.6: *Experimental semivariogram, estimated by the method of moments, and spherical semivariogram models fitted using ordinary and weighted least squares. The weighting is done according to Cressie (1985).*

the semivariogram model. These methods rely crucially on the Gaussian assumption, i.e.

$$\boldsymbol{Z} \in N(\boldsymbol{X\beta}, \boldsymbol{\Sigma}) \tag{A.27}$$

where $\boldsymbol{X}$ is a matrix of known regressors, $\boldsymbol{\beta}$ is a vector of unknown coefficients, see (A.18), and $\boldsymbol{\Sigma}$ is the covariance matrix of the observations. The covariance matrix can be factored as

$$\boldsymbol{\Sigma} = \alpha \boldsymbol{V}(\boldsymbol{\theta}) \tag{A.28}$$

where α is a scale parameter and $\boldsymbol{V}(\boldsymbol{\theta})$ is a matrix of standardized covariances. $\boldsymbol{Z}$, defined by (A.27), has the probability density function

$$(2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp(-1/2(\boldsymbol{Z}-\boldsymbol{X\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z}-\boldsymbol{X\beta})) \tag{A.29}$$

and the negative log likelihood of that is given by

$$l(\boldsymbol{\beta},\alpha,\boldsymbol{\theta}) = \frac{n}{2}\log(2\pi)+\frac{n}{2}\log\alpha+\frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\theta})|+\frac{1}{2\alpha}(\boldsymbol{Z}-\boldsymbol{X\beta})^T\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{Z}-\boldsymbol{X\beta}) \tag{A.30}$$

If we define

$$\begin{aligned} \hat{\beta}(\boldsymbol{\theta}) &= (\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{Z} \\ \boldsymbol{G}^2(\boldsymbol{\theta}) &= (\boldsymbol{Z}-\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}))^T\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{Z}-\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})) \end{aligned} \tag{A.31}$$

we get the following negative log likelihood

$$l(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\alpha,\boldsymbol{\theta}) = \frac{n}{2}\log(2\pi)+\frac{n}{2}\log\alpha+\frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\theta})|+\frac{1}{2\alpha}\boldsymbol{G}^2(\boldsymbol{\theta}) \tag{A.32}$$

(A.32) can be minimized numerically with respect to α and $\boldsymbol{\theta}$ or analytically with respect to α by defining

$$\hat{\alpha}(\boldsymbol{\theta}) = \frac{\boldsymbol{G}^2(\boldsymbol{\theta})}{n} \tag{A.33}$$

In that case we have to minimize

$$l(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\hat{\alpha}(\boldsymbol{\theta}),\boldsymbol{\theta}) = \frac{n}{2}\log(2\pi)+\frac{n}{2}\log\frac{\boldsymbol{G}^2(\boldsymbol{\theta})}{n}+\frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\theta})|+\frac{n}{2} \tag{A.34}$$

with respect to $\boldsymbol{\theta}$ (Zimmerman and Zimmerman, 1991; Pardo-Iguzquiza, 1997; Smith, 2001).

A.4.2 Restricted maximum likelihood

The motivation of using restricted maximum likelihood is that the simultaneous estimation of both the trend and semivariogram parameters produces biased parameter estimates. Instead of working with the original data, linear combinations of data are used, which serve to filter out the trend (Pardo-Iguzquiza, 1997). We have

$$\boldsymbol{W} \in N(\boldsymbol{0}, \boldsymbol{A}^T \boldsymbol{\Sigma} \boldsymbol{A}) \tag{A.35}$$

where $\boldsymbol{W} = \boldsymbol{A}^T \boldsymbol{Z}$ is a vector of $n-q$ linear independent contrasts, and $\boldsymbol{A}^T \boldsymbol{X} = \boldsymbol{0}$. The negative log likehood function of $\boldsymbol{W}$ is

$$\begin{aligned} l_W(\alpha, \boldsymbol{\theta}) &= \frac{n-q}{2}\log(2\pi) + \frac{n-q}{2}\log(\alpha) + \frac{1}{2}\log|\boldsymbol{A}^T\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{A}| \\ &\quad + \frac{1}{2\alpha}\boldsymbol{W}^T(\boldsymbol{A}^T\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{A})^{-1}\boldsymbol{W} \end{aligned} \tag{A.36}$$

where α and $\boldsymbol{V}(\boldsymbol{\theta})$ are given by (A.28). When chosing $\boldsymbol{A}$ to satisfy $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}$, (A.36) can be simplified to

$$\begin{aligned} l_W(\alpha, \boldsymbol{\theta}) &= \frac{n-q}{2}\log(2\pi) + \frac{n-q}{2}\log(\alpha) - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{X}| \\ &\quad + \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}| + \frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\theta})| + \frac{1}{2\alpha}\boldsymbol{G^2}(\boldsymbol{\theta}) \end{aligned} \tag{A.37}$$

where $\boldsymbol{G^2}(\boldsymbol{\theta})$ is the same as in (A.34). (A.37) is minimized with respect to α by setting $\tilde{\alpha} = \boldsymbol{G^2}(\boldsymbol{\theta})/(n-q)$, and (A.37) can be reduced to

$$\begin{aligned} l^*_W(\boldsymbol{\theta}) &= l_W(\tilde{\alpha}, \boldsymbol{\theta}) \\ &= \frac{n-q}{2}\log(2\pi) + \frac{n-q}{2}\log(\frac{\boldsymbol{G^2}(\boldsymbol{\theta})}{n-q}) - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{X}| \\ &\quad + \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}| + \frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\theta})| + \frac{n-q}{2} \end{aligned} \tag{A.38}$$

Comparing (A.34) and (A.38), it is seen that the coefficient $\frac{n}{2}$ changes to $\frac{n-q}{2}$, and that there is an additional term of $\frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}|$.

Figure A.7 shows the estimated semivarigram model using maximum likelihood and restricted maximum likelihood. The experimental semivariogram, estimated by the method of moments, have been added to the Figure, even though the likelihood estimation methods are not based on an experimental semivariogram.

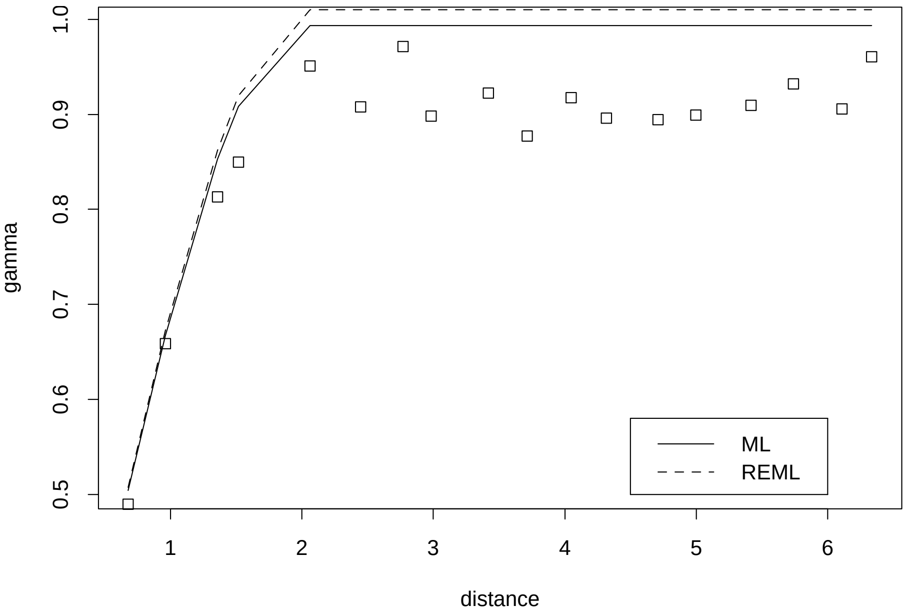

Figure A.7: *Experimental semivariogram, estimated by the method of moments, and spherical semivariogram models fitted using maximum likelihood and restricted maximum likelihood.*

A.5 Results and discussion

The above described methods for estimating the semivariogram have been combined into eight approaches. The combinations are described in the introduction of this paper, and are illustrated in Figure A.1. Table A.1 gives a summary, where the methods are given a number, which is used as a reference when presenting the results in Figure A.8 and in Table A.4 and A.5. The computations are made by simulating spatial data as a Gaussian random field with a known covariance structure, and comparing the estimated parameters with the true values.

The comparison is executed for different values of the sill, range and nugget effect, and for different sample sizes, types of semivariogram models and grid stuctures, i.e. for six different factors. If two levels of each factor are considered, it means that the input can be combined in $2^6 = 64$ different ways. The two levels of the six factors, that will be used, are shown in Table A.2. Changing

Approach number	Methods for estimation of the semivariogram
1	Method of moments + Ordinary least squares
2	Method of moments + Weighted least squares
3	Robust method + Ordinary least squares
4	Robust method + Weighted least squares
5	LOESS + Ordinary least squares
6	LOESS + Weighted least squares
7	Maximum likelihood
8	Restricted maximum likelihood

Table A.1: *Eight approaches for estimating the semivariogram.*

these six factors does not change the distribution of data. We only considered data which are normally distributed, i.e. generated as a Gaussian random field. One should keep this in mind when analyzing environmental data, which are not necessarily normally distributed. The calculations are very time-demanding, and instead of making all 64 experiments, a $2^{(6-3)}$ factor experiment have been designed by confounding the two-factor interactions, AB, AC and BC, with the main factors D, E and F. The resulting design of the eight experimental trials is shown in Table A.3. For each experimental trial, 100 realizations of a Gaussian random field are generated, and the mean value and variance of the parameter estimates are calculated. The difference between the mean and "true" value is a measure of bias, while the variance is a measure of the uncertainty of the parameter estimation. As described in section A.2.3 the bandwidth for locally

Levels	Range (A)	Sill (B)	Nugget (C)	Grid (D)	Sample size (E)	Model (F)
Low	2	1	0	Regular	100	Exponential
High	4	2	1	Irregular	225	Spherical

Table A.2: *Factors and levels of factors for which the comparison is executed.*

Experimental trial	A	B	C	D=AB	E=AC	F=BC	Design
1	-	-	-	+	+	+	DEF
2	+	-	-	-	-	+	AF
3	-	+	-	-	+	-	BE
4	+	+	-	+	-	-	ABD
5	-	-	+	+	-	-	CD
6	+	-	+	-	+	-	ACE
7	-	+	+	-	-	+	BCF
8	+	+	+	+	+	+	ABCDEF

Table A.3: *Design of a $2^{(6-3)}$ factor experiment.*

weighted regression of the semivariogram cloud is found using Akaikes information criterion (AIC). It was found that a bandwidth of approximately 0.5 minimizes AIC, thus this value seems to be optimal when smoothing the semivariogram cloud by locally weighted regression.

The results of the comparison study are shown in appendix, for each combination of design and estimation approach. Furthermore, Figure A.8 shows the results where we have summed over designs. It is seen, in the upper part of the Figure, that the variances of the three semivariogram parameters are smallest when maximum likelihood is used for estimation. Also restricted maximum likelihood results in relatively small variances. The six different least squares approaches seem to give estimation variances which are quite similar, however, regarding the nugget effect LOESS smoothing of the semivariogram cloud results in estimation variances, which are nearly as small as those obtained by the maximum likelihood methods.

The lower part of Figure A.8 shows the sum of the absolute values of biases, summed over designs. Only restricted maximum likelihood seems to give significantly smaller biases than the other approaches, which agrees very well with the studies of Swallow and Monahan (1984); McGilchrist (1989), while the biases of the maximum likelihood estimators are very similar to the values found for the nonparametric least squares estimators.

Table A.4 and A.5 in the appendix indicate that the ratio between the sill and nugget effect, sill/nugget effect, influences the variance of the estimator. This is especially true for the maximum likelihood estimators, and the choice between using maximum likelihood and least squares estimation is therefore believed to depend on this ratio. The relationship between the variance of the estimator and the sill/nugget effect - ratio was studied by simulating 100 realizations of a Gaussian random field on a regular 15 × 15 grid. The covariance structure was chosen to be spherical with a range=6 and sill=4, while the sill/nugget effect - ratio was varied by varying the nugget effect from 1 to 9 in steps of 1. Two methods, for estimation of the three parameters, were considered. These were maximum likelihood estimation and estimation of the experimental semivariogram by the methods of moments, with subsequent parameter estimation by ordinary least squares. The results are shown in Figure A.9, where the estimation variances, computed by maximum likelihood, have been fitted using a function of the form variance $= \exp[(\text{sill/nugget effect}) \cdot a] \cdot b$.

It is clearly seen that the variance of the maximum likelihood estimator increases when decreasing the sill/nugget effect - ratio for all the three parameters. When estimating the nugget effect the variance of the least squares estimator shows a similar dependence, while no dependence is found when estimating the range and the sill. The corresponding results for the biases are not shown in Figure

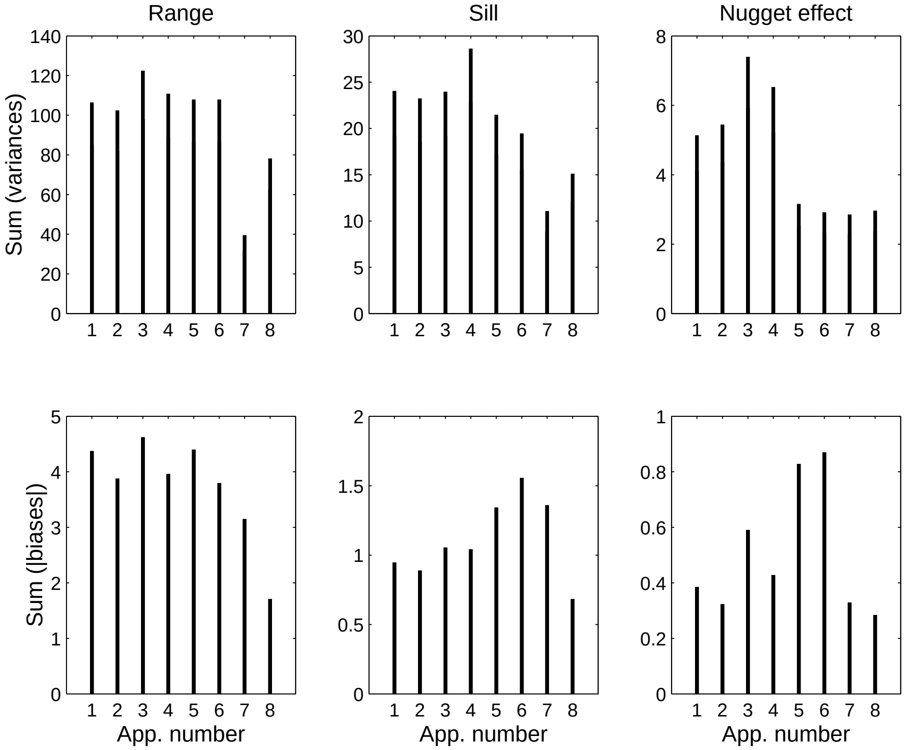

Figure A.8: *Variances and biases, summed over designs, of the three parameters of the semivariogram model, computed by eight different approaches.*

A.9. However, a dependence between the sill/nugget effect - ratio and the biases of the maximum likelihood estimators have not been found. Taking into account the complexity of the maximum likelihood methods, as well as the fact that the computations of the maximum likelihood estimators are much more time-demanding, we believe that these methods should only be applied when the sill/nugget effect - ratio is high, e.g. > 1, otherwise the advantages of the methods are insignificant.

We also applied maximum likelihood and least squares estimation to a real dataset, containing measurements of salinity at 71 sampling stations in the Kattegat basin. The least squares approach we considered was estimation of the experimental semivariogram by the method of moments, with subsequent parameter estimation by ordinary least squares. The experimental semivariogram is shown as circled dots, $\circ$, in Figure A.10, while the two lines show spherical semivariogram models estimated by ordinary least squares and maximum likelihood. In this case the estimated models are very similar, and they go through the experimental semivariogram. In general, this is not necessarily the case for the maximum likelihood estimator, see Figure A.7, because it is not based on the experimental semivariogram. Figure A.11 shows spatial predictions computed by ordinary kriging, see e.g. Cressie (1993); Isaaks and Srivastava (1989), and based on semivariogram models estimated by ordinary least squares and maximum likelihood. It is seen that both predictions and standard deviations are almost identical for the two cases, which is caused by the similarity of the estimated semivariogram models. However, the standard deviations of predictions based on maximum likelihood are slightly lower than in the case of ordinary least squares.

A.6 Conclusion

This paper describes and compares eight different approaches for estimating spatial variability, in form of the semivariogram. The comparison showed that maximum likelihood estimation results in the smallest variances of the semivariogram parameters, while the smallest biases were found by using restricted maximum likelihood. Furthermore, it was found that the variances of the maximum likelihood estimators depend on the sill/nugget effect - ratio, i.e. higher sill/nugget effect - ratios result in lower estimation variances. For the least squares methods a similar dependence is only found for the nugget effect-estimator. This means that the advantage of using maximum likelihood is large when the sill/nugget effect - ratio is high. Taking into account the complexity of the maximum likelihood methods, as well as the fact that the computations of the maximum likelihood estimators are much more time-demanding, we therefore recommend

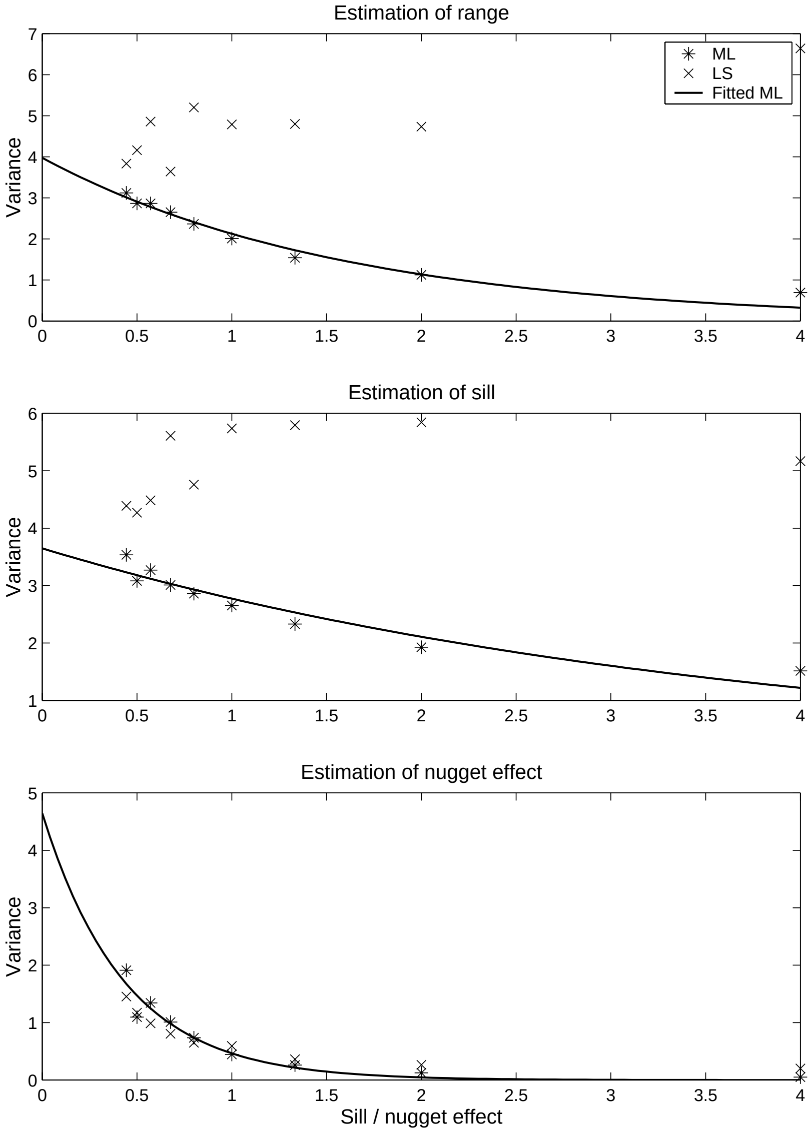

Figure A.9: *The variance of the three semivariogram parameters as a function of the sill / nugget effect - ratio. ML is maximum likelihood estimation, and LS least squares estimation.*

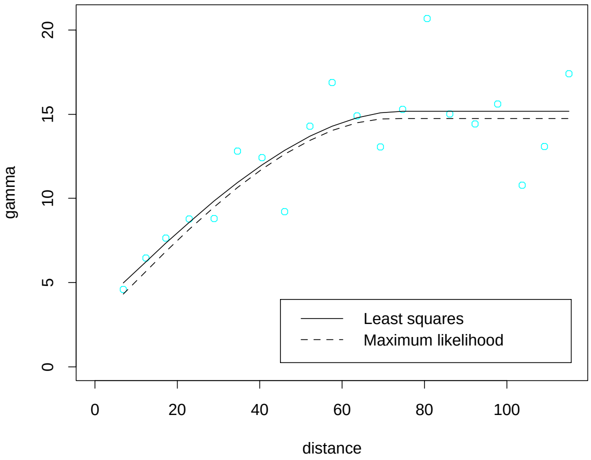

Figure A.10: *Experimental semivariogram estimated by the method of moments, and semivariogram models estimated by maximum likelihood and ordinary least squares.*

only to apply maximum likelihood when the sill/nugget effect - ratio is high. We also applied maximum likelihood and least squares estimation to a real dataset, containing measurements of salinity at 71 sampling stations in the Kattegat basin. This showed that the calculation of spatial predictions is insensitive to the choice of estimation method, but also that the uncertainties of kriging predictions were reduced when applying maximum likelihood.

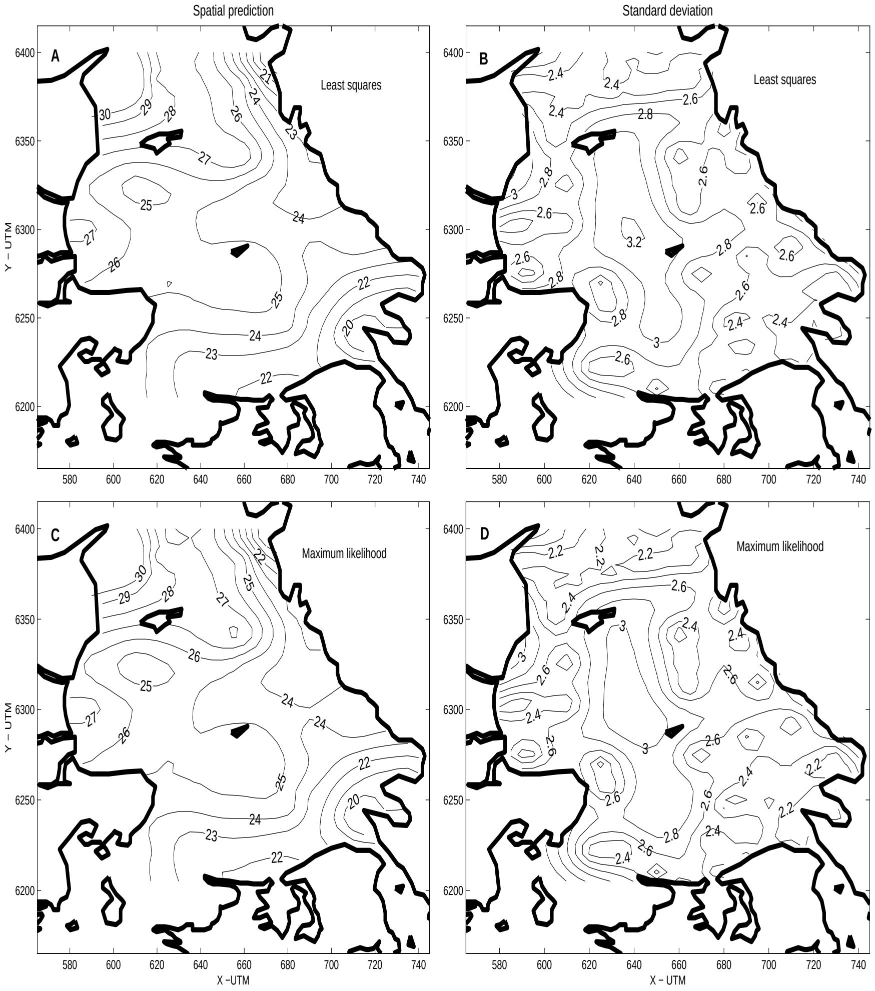

Figure A.11: *Spatial predictions of salinity in Kattegat. A) Spatial predictions using a spherical semivariogram model estimated by ordinary least squares. B) Standard deviations of the spatial predictions based on least squares estimation. C) Spatial predictions using a spherical semivariogram model estimated by maximum likelihood. D) Standard deviations of the spatial predictions based on maximum likelihood estimation.*

Appendix

Design	Estimation	Statistics					
	Method	R		C_1		C_0	
		Mean	Var	Mean	Var	Mean	Var
DEF	1	2.1269	0.6950	0.9676	0.0501	0.0061	0.0037
DEF	2	2.0195	0.3806	0.9742	0.0445	0.0068	0.0018
DEF	3	2.1466	0.6528	1.0046	0.0578	-0.0035	0.0044
DEF	4	2.0737	0.4399	1.0104	0.0521	0.0020	0.0018
DEF	5	2.2388	0.6528	0.9517	0.0699	0.0219	0.0127
DEF	6	2.3644	1.0014	0.8621	0.0441	0.1101	0.0141
DEF	7	1.9902	0.0607	0.9525	0.0200	0.0025	0.000010267
DEF	8	2.0292	0.0681	0.9758	0.0238	0.0024	0.0000099098
AF	1	4.3069	2.1599	1.0546	0.1147	-0.0217	0.0274
AF	2	4.2351	1.5599	1.0568	0.1074	-0.0189	0.0192
AF	3	4.4409	2.6931	1.1203	0.1476	-0.0417	0.0443
AF	4	4.2610	1.7039	1.1372	0.1355	-0.0479	0.0304
AF	5	4.2254	1.7630	1.0554	0.1066	-0.0254	0.0232
AF	6	4.1259	1.2323	1.0631	0.1112	-0.0309	0.0197
AF	7	3.9572	0.1632	0.8902	0.0289	0.0243	0.0014
AF	8	4.0626	0.2941	0.9263	0.0313	0.0221	0.0014
BE	1	2.8757	5.5567	2.4023	1.2232	-0.0763	0.1436
BE	2	2.5432	3.1543	2.2774	0.9921	-0.0470	0.0909
BE	3	2.7684	5.6289	2.4760	0.9075	-0.1657	0.2142
BE	4	2.5763	3.3459	2.3804	0.8961	-0.0930	0.1364
BE	5	2.8220	5.3005	2.3464	1.1047	-0.0649	0.1459
BE	6	2.6454	3.6152	2.2628	1.0432	-0.0147	0.0819
BE	7	2.0767	0.7648	1.8981	0.4101	0.0337	0.0033
BE	8	2.4636	1.2846	2.1599	0.6065	0.0421	0.0042
ABD	1	2.7738	3.6863	1.7165	1.3000	-0.0524	0.0055
ABD	2	3.0885	5.2050	1.7027	1.1885	-0.0061	0.00016507
ABD	3	2.9028	5.1217	1.8345	1.4387	-0.0699	0.0071
ABD	4	3.3448	7.1188	1.9348	2.0866	-0.0137	0.00068714
ABD	5	2.6596	3.4763	1.6652	1.1856	-0.0571	0.0087
ABD	6	3.2272	5.4987	1.4823	0.7619	0.0478	0.0061
ABD	7	2.9197	2.1371	1.4388	0.4425	0.0016	0.0000064270
ABD	8	3.6560	3.7145	1.7800	0.8303	0.0017	0.0000061003

Table A.4: *Comparison of different approaches for estimating the semivariogram.*

Design	Approach	Statistics					
	Number	R		C_1		C_0	
		Mean	Var	Mean	Var	Mean	Var
CD	1	1.6195	1.0732	1.0467	0.2898	0.9115	0.0383
CD	2	1.6211	1.5559	1.0421	0.3545	0.9298	0.0317
CD	3	1.5786	1.8349	1.0447	0.2101	0.8864	0.0556
CD	4	1.6304	1.5809	1.0602	0.2752	0.9246	0.0435
CD	5	1.7560	1.7783	1.0399	0.3010	0.9261	0.0337
CD	6	1.7337	1.8024	1.0446	0.3585	0.9291	0.0351
CD	7	1.5439	0.9189	0.8632	0.1102	0.9650	0.0268
CD	8	2.1606	3.1402	1.0181	0.2046	0.9741	0.0280
ACE	1	3.2535	4.4260	1.0347	0.3121	0.9039	0.0532
ACE	2	3.2131	4.1673	1.0647	0.3709	0.9068	0.0431
ACE	3	3.1144	4.2107	1.1038	0.3126	0.8430	0.0995
ACE	4	3.0798	3.3359	1.0675	0.3032	0.8726	0.0716
ACE	5	3.2990	5.1093	1.0251	0.3452	0.9110	0.0465
ACE	6	3.3340	5.0313	1.0767	0.3719	0.8896	0.0488
ACE	7	2.7544	2.7727	0.8680	0.1821	0.9270	0.0399
ACE	8	3.7645	5.6019	1.0087	0.2201	0.9364	0.0404
BCF	1	2.5470	1.4421	2.0719	0.8575	0.9571	0.7232
BCF	2	2.6377	1.6351	2.1132	0.9125	0.9588	0.8663
BCF	3	2.7340	1.7259	2.0654	1.0293	0.9733	1.0151
BCF	4	2.7098	1.3765	2.1469	1.0649	0.9597	0.9846
BCF	5	2.7312	1.5441	1.5639	0.4862	1.4749	0.3167
BCF	6	2.7521	1.2148	1.5844	0.4847	1.4820	0.3312
BCF	7	2.1768	0.5224	2.1076	0.6775	0.8554	0.4825
BCF	8	2.3154	0.8403	2.1422	0.6852	0.8830	0.5026
ABCDEF	1	4.1631	2.2504	2.0206	0.6641	1.0016	0.0324
ABCDEF	2	4.3650	2.8321	2.0117	0.6790	1.0398	0.0360
ABCDEF	3	4.1290	2.6264	2.0738	0.6922	0.9874	0.0394
ABCDEF	4	4.3964	3.2681	2.1746	0.9132	1.0282	0.0363
ABCDEF	5	4.0957	1.9661	2.0568	0.6966	0.9787	0.0443
ABCDEF	6	4.2044	2.1928	2.0383	0.7170	1.0035	0.0464
ABCDEF	7	3.9399	0.5598	1.8378	0.3426	1.0144	0.0173
ABCDEF	8	4.0960	0.7040	1.9640	0.4212	1.0097	0.0168

Table A.5: *Comparison of different approaches for estimating the semivariogram.*

References

Asli, M. and Marcotte, D. (1995). Comparison of approaches to spatial estimation in a bivariate context. *Mathematical Geology*, **27**(5), 641–658.

Brus, D., DeGruijter, J., Marsman, B., Visschers, R., Bregt, A., Breeuwsma, A., and Bouma, J. (1996). The performance of spatial interpolation methods and chloropleth maps to estimate properties at points: A soil survey case study. *Environmetrics*, **7**(1), 1–16.

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

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

Cressie, N. (1993). *Statistics for spatial data.* Wiley, New York.

Cressie, N. and Hawkins, D. (1980). Robust estimation of the semivariogram. *Mathematical Geology*, **12**, 115–125.

Hosseini, E., Gallichand, J., and Caron, J. (1993). Comparison of several interpolators for smoothing hydraulic conductivity data in south west Iran. *Transactions of the American Society of Agricultural Engineers*, **36**(6), 1687–1693.

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

Laslett, G. (1994). Kriging and splines: An empirical comparison of their predictive performance in some applications. *Journal of the American Statistical Association*, **89**(426), 391–409.

McGilchrist, C. (1989). Bias of ML and REML in regression models with ARMA errors. *Journal of Statistical Computation and Simulation*, **32**, 127–136.

Pardo-Iguzquiza, E. (1997). MLREML: A computer program for the inference of spatial covariance parameters by maximum likelihood and restricted maximum likelihood. *Computers & Geosciences*, **23**(2), 153–162.

Smith, R. (2001). Environmental statistics. Department of Statistics, University of North Carolina. Prepared in connection with NSF-CBMS regional conference in the mathematical sciences: Environmental statistics, University of Washington, June 25-29, 2001.

Swallow, W. and Monahan, J. (1984). Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics*, **26**, 47–57.

Zimmerman, D. and Zimmerman, M. (1991). A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictions. *Technometrics*, **33**(1), 77–91.

Appendix B

Space-time modeling of environmental monitoring data

The paper is published in:

Lophaven, S., Carstensen, J., and Rootzén, H. (2004). Space-time modeling of environmental monitoring data. *Environmental and Ecological Statistics* vol. 11, p. 237-256.

Space-time modeling of environmental monitoring data

Søren Lophaven[1], Jacob Carstensen[2], and Helle Rootzén[1]

[1] Informatics and Mathematical Modelling, Technical University of Denmark
[2] Department of Marine Ecology, National Environmental Research Institute of Denmark

Abstract

This study describes and applies statistical methods for space-time modeling of data from environmental monitoring programs, e.g. within areas such as climate change, air pollution and aquatic environment. Such data are often characterized by sparse sampling in both the temporal and spatial dimensions. In order to improve the amount of information on the physical system in question we suggest using statistical modeling methods for monitoring data. Model predictions combined with observations could be analysed directly to assess the environmental state or as forcing functions for time series models and deterministic, hydrodynamic models. To illustrate the approach we applied the proposed modeling methods to data from the Danish and Swedish marine monitoring programs. Time series with a weekly resolution were predicted from observations of dissolved inorganic nitrogen (DIN) from the Kattegat basin (1993 - 1997). DIN observations were sparse, irregularly distributed and comprised approximately 10 % of the generated time series.

KEY WORDS: *Space-time modeling, dissolved inorganic nitrogen, Kattegat*

B.1 Introduction

Anthropogenic disturbances in the environment have increased substantially over the last century ranging from a global scale (e.g. climate change, ozone depletion) over regional scales (e.g. eutrophication, acidification) to local scales (e.g. air pollution in cities, point source discharge of harmful substances). Monitoring programs have been established in many industrialized countries to assess the magnitude and consequences of human stress on the environment. However, the complexity of the many interacting processes and the costs associated with environmental monitoring poses a paradox that frequently results in only partial fulfillment of monitoring objectives. It is therefore important that these data are exploited to the fullest extent for optimal use of the limited resources available for environmental monitoring.

Environmental processes reflect temporal and spatial variations on a variety of scales that are only partially captured in monitoring data. In particular, the marine environment comprises a complex mosaic of interacting processes, which range from small-scale microbial processes to global-scale oceanic circulation. On the other hand, monitoring at sea by traditional shipboard sampling is associated with large costs for personnel and equipment, and new technologies aiming at reducing costs have not yet proven adequate to substitute for monitoring vessels. Consequently, the spatial and temporal coverage of data is limited and often irregular, and temporal and spatial variations can only be assessed on a coarse resolution scale, unless methods are employed that integrate monitoring data in time and space.

The aim of this study is to describe statistical methods for temporal and spatial modeling of environmental data characterized by sparse sampling in time and space. The method description leads to four different approaches which are illustrated by observations of dissolved inorganic nitrogen (DIN) from the Kattegat basin during 1993-1997. The approaches assume that the log-transformed DIN observations are uncorrelated or spatially correlated, respectively. Furthermore, two different back transforms from the log to the original scale are presented, and hence the two approaches for modeling log-transformed DIN and the two back transforms can be combined into four ways of modeling DIN. The methods are general and can be applied to other sources of monitoring data as well, e.g. air pollution and climate data. The combination of model predictions and observations provide an improvement for assessing effects of proposed nutrient reductions by means of statistical analyses. Moreover, model predictions combined with observations can be applied as forcing functions for time series models and deterministic, hydrodynamic models. This will advance the knowledge of the chemical and biological processes in the marine environment, and reduce uncertainties of regional nutrient and carbon budgets.

B.2 Study area and data material

During the 1980s numerous episodes of oxygen deficiency, covering large areas of the Danish estuaries, were observed (Kronvang et al., 1993). This resulted in the adoption of the Action Plan on the Aquatic Environment in 1987, which required that total discharge of nitrogen from diffuse sources (agriculture) and point sources (municipal wastewater treatment plants and industrial outfalls) were to be reduced by 50 % from a total of 290,000 tonnes per year in 1987 to around 145,000 tonnes per year in 1993. During the same period phosphorus discharges were to be reduced by 80 % from a total of approximately 12,000 tonnes per year to 2,200 tonnes per year. In connection with the adoption of this plan a monitoring program, DNAMAP, was established. The purpose of the program was to characterise the state of the aquatic environment and to document the effects of the measures being taken to reduce nutrient delivery to the marine environment (Kronvang et al., 1993). As a result of DNAMAP, Danish estuaries are among the best-monitored marine systems in the world.

The Kattegat basin is a transition zone between the North Sea and the Baltic Sea (Figure B.1A) with a surface area of 22,290 km^2, a volume of 533 km^3 and a mean depth of approximately 24 meters (Gustafsson, 2000). The area is dominated by advective transport of low-saline water from the Baltic Sea as a surface current and water with a high salinity from the North Sea as a bottom current. This advection creates a strong salinity stratification located at 15-20 meters depth throughout most of the year (Andersson and Rydberg, 1988). Observations for this study were made at 65 stations in Kattegat (Figure B.1B) during a five-year period 1993 - 1997) by Danish and Swedish authorities. A variety of water quality parameters were measured in samples from depths through the water column.

In this study we have chosen to focus on DIN, which is the sum of the following nitrogen constituents: ammonium (NH_4^+-N), nitrite (NO_2^--N) and nitrate (NO_3^--N). DIN is an important parameter, because algae growth in Kattegat is generally nitrogen limited (Granéli, 1987). Observations from the top 10 m of the water column were averaged to produce surface DIN, which is readily available for algae production.

DIN data comprise 1932 observations scattered over 5 years and 65 stations (Figure B.2) with only a few weeks having more than 20 surface values at various stations (Figure B.2) whereas 60 % of the weeks had fewer than 8 observations. Many stations had fewer than 15 observations over the five-year period (Figure B.2B) and only 4 stations were sampled more than 100 times corresponding to biweekly sampling. If we consider the data matrix consisting of 65 stations and 260 weeks, i.e. a data matrix of 65 × 260 = 16900 cells, then monitoring data would fill in approximately 10 % of the cells. The remaining 90 % of missing values are to be predicted by the proposed model. Results will be presented for stations 20004, 1001 and SI2 (Figure B.1B). Stations 20004 and 1001 represent coastal and open-water stations, respectively, whereas SI2 was chosen to show the method performance for a station with few observations.

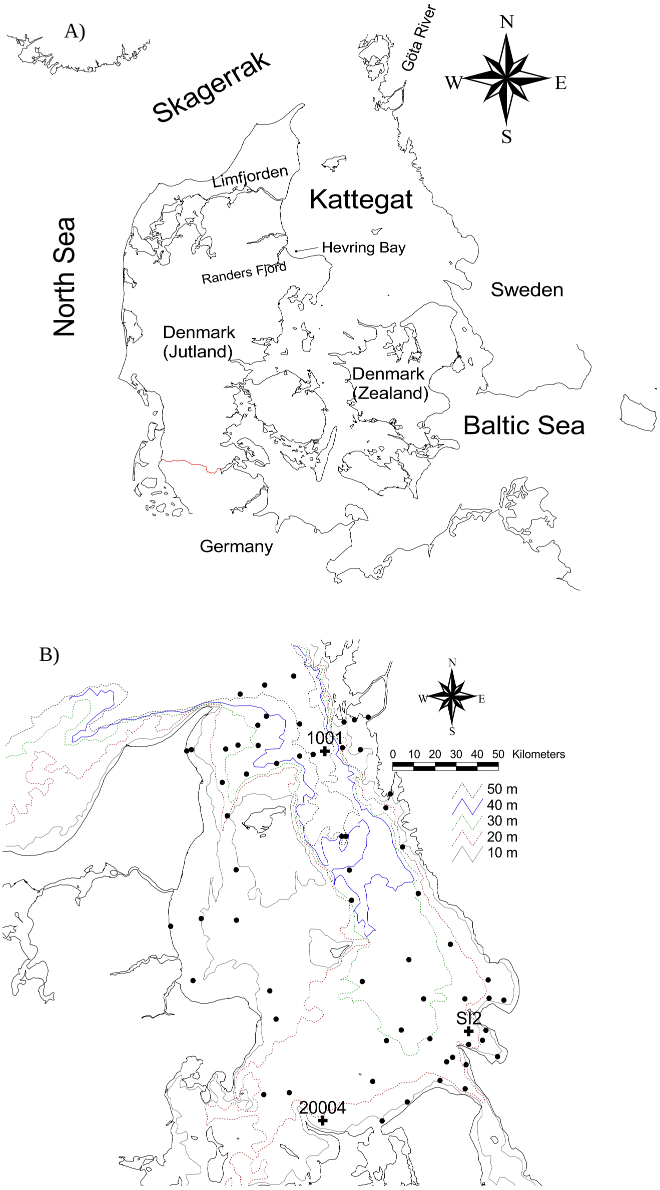

Figure B.1: *The study area. A) Kattegat as a transitional sea between the North Sea and the Baltic Sea. B) Locations of sampling stations (•), and the depth contours (in meters) in Kattegat. Three stations selected for presentation of results are marked with (+).*

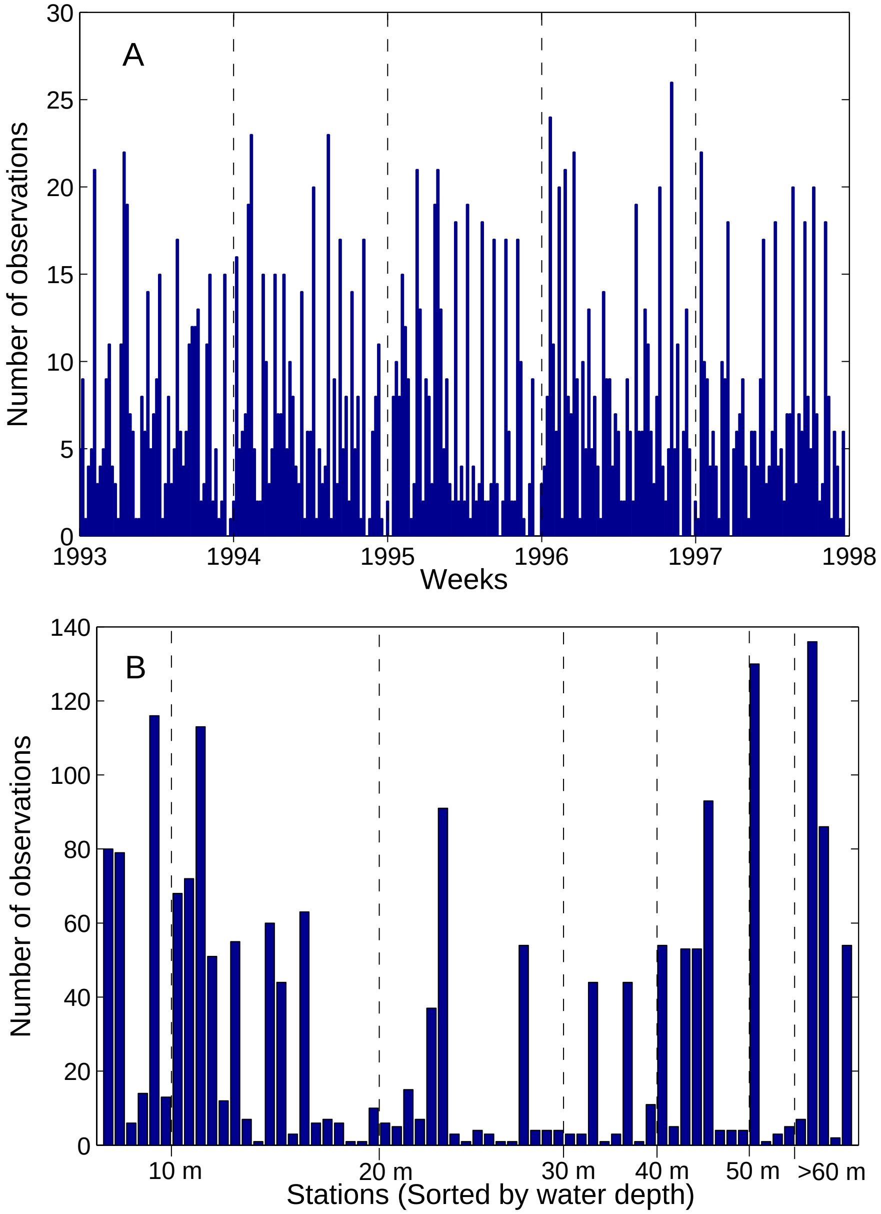

Figure B.2: *A) Number of stations with DIN observations for each week during the study period. B) Number of observations for each station over the entire study period.*

B.3 Space-time modeling methods

Methods for modeling space-time data are usually developed with a specific application in mind. Most studies apply a model of the general decomposed form

$$Z(s,t) = \mu(s,t) + \epsilon(s,t) \tag{B.1}$$

where $\mu(s,t)$ is mean component or trend modeled as a deterministic function depending on space s and time t, and $\epsilon(s,t)$ is the residual component describing fluctuations around the mean in space and time.

The mean component

The mean component $\mu(s,t)$ is generally modeled by means of a deterministic function in both space and time. For example Haas (1998) used a space-time model for predicting wet sulfate deposition with the mean component modeled as

$$\mu_j(s,t) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_3 s_1^2 + \beta_4 s_2^2 + \beta_5 s_1 s_2 + \beta_6 t + \beta_{j+6}, \qquad s = (s_1, s_2) \tag{B.2}$$

where s_1 and s_2 are spatial coordinates, t is time and the four seasonality parameters, β_7 through β_{10}, are constrained by $\sum_{j=1}^{4} \beta_{j+6} = 0$. The model assumed that spatial and temporal variations were not interacting. Carroll et al. (1997) used a deterministic mean component for predicting urban ozone concentrations in Harris County, Texas, which was only a function of time

$$\mu(s,t) = \mu(t) = \alpha_{\text{hour}} + \beta_{\text{month}} + \gamma_1 \text{temp}(t) + \gamma_2 \text{temp}^2(t) \tag{B.3}$$

where temp(t) is the median of the temperatures reported at various places in Harris County at time t, α_{hour} accounts for the overall hourly level of ozone and β_{month} for the overall monthly level of ozone. Spatial variations were included in the residual component, and the model was used to examine the population exposure of ozone, as well as to evaluate the siting of monitors.

The residual component

The residual component $\epsilon(s,t)$ in (B.1) is assumed to be a second order stationary stochastic process with expected value

$$E(\epsilon(s,t)) = 0 \tag{B.4}$$

and covariance function

$$C_{st}(h_s, h_t) = \text{Cov}(\epsilon(s + h_s, t + h_t), \epsilon(s, t)) \tag{B.5}$$

where h_s is the spatial and h_t the temporal separation distance. One way to model the covariance in (B.5) is to separate it into a spatial and a temporal component. Three types of separability are frequently used: the product, the sum and the product-sum model, given by

$$\begin{aligned} C_{st}(h_s, h_t) &= C_s(h_s)C_t(h_t) \qquad \text{(B.6)} \\ C_{st}(h_s, h_t) &= C_s(h_s) + C_t(h_t) \\ C_{st}(h_s, h_t) &= k_1 C_s(h_s)C_t(h_t) + k_2 C_s(h_s) + k_3 C_t(h_t) \end{aligned}$$

A brief discussion of separable space-time covariance models can be found in De Cesare et al. (2001a). The product model was applied in Haas (1995, 1998) for wet sulfate deposition, whereas the product-sum model was used in De Cesare et al. (2001b) for modeling NO_2 concentrations in the Milan district. The product and the product-sum covariance models (B.6) in these studies were implemented by modifying the GSLIB FORTRAN 77 routines, originally developed specifically for modeling spatial data (De Cesare et al., 2002). Another class of space-time covariance models is the non-separable models, which do not assume that the temporal and spatial components can be separated. A description of this class of models is found in Cressie and Huang (1999). De Iaco et al. (2001, 2002) also described non-separable space-time covariance models as generalizations of the separable product and product-sum models in (B.6). Brown et al. (2001) discussed the use of both separable and non-separable models for calibration of radar rainfall data. They argued, with respect to the spatial resolution of the rainfall data, that a separable model adequately approximated data. Meiring et al. (1998) applied a non-separable model for estimation of hourly ozone levels, and compared their predictions with the SARMAP photochemical air-quality model for a region of northern California. They found that their model improved prediction of hourly ozone levels compared to SARMAP.

B.3.1 The DIN model

Our modeling strategy reflects the fact that DIN is sparsely sampled and that DIN exhibits variations in time and space which cannot be adequately described by a continuous and yet simple function of s and t. Furthermore, the highly dynamic properties of DIN concentrations in Kattegat required a temporal resolution in the predictions greater than that given by seasonal components as in Haas (1998). A temporal resolution of one week was the greatest attainable frequency from the monitoring data in Kattegat.

In the following let $Z_0(s,t)$ denote the surface DIN concentration and let $Z(s,t) = \text{Ln}(Z_0(s,t))$ be the log-transform of $Z_0(s,t)$. $Z(s,t)$ is modeled using the general decomposed form (B.1), where the mean component describes variations between stations and weeks by means of indices for each station and each week, i.e.

$$\mu_{kl}(s,t) = station_k + week_l \qquad k = 1, \cdots, n_{station} \quad \text{and} \quad l = 1, \cdots, n_{week} \tag{B.7}$$

This implies that the spatial variation is modeled by $n_{station}$ levels, one for each station, and the temporal variation is modeled by n_{week} levels corresponding to a weekly resolution. No interaction between space and time is assumed. Another consequence of the discrete mean component model (B.7) is that predictions of $Z(s,t)$ cannot be generated at non-monitored sites or for weeks without monitoring data at any station. Using the mean component in (B.7) the model for $Z(s,t)$ can be written as

$$Z_{kl}(s,t) = station_k + week_l + \epsilon(s,t), \qquad \epsilon(s,t) \sim \text{N}(0, \sigma^2\mathbf{\Sigma}) \tag{B.8}$$

where $\sigma^2\mathbf{\Sigma}$ is the covariance matrix, i.e. in general we have $\boldsymbol{C}_{st}(h_s, h_t) = \sigma^2\mathbf{\Sigma}$. In matrix form (B.8) can be written as

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{B.9}$$

where $\boldsymbol{Z}$ is an observation vector of log-transformed surface DIN, and $\boldsymbol{X}$ is the design matrix containing indicator variables, i.e. ones for combinations of station and week for which an observation exists, and zeros at all other places in the matrix. Furthermore, $\boldsymbol{\beta}$ is the parameter vector, and $\boldsymbol{\epsilon}$ a vector of model residuals. $\boldsymbol{\beta}$ is estimated as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\mathbf{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{\Sigma}^{-1}\boldsymbol{Z} \tag{B.10}$$

The simplest way to model the residual component $\epsilon(s,t)$ in (B.8) is to assume that the covariance function is σ^2 when h_t=h_s=0 and zero otherwise. This means that no temporal or spatial correlation in $Z(s,t)$ is included in the model, and this corresponds to a general linear model for normal distributed, uncorrelated residuals. In this case $\mathbf{\Sigma}$ in (B.10) is the identity matrix, and a predicted value of $Z(s,t)$ is obtained by

$$\hat{Z}(s,t) = \boldsymbol{x_p}'\hat{\boldsymbol{\beta}} \tag{B.11}$$

where $\boldsymbol{x_p}$ is a column vector of indicator variables corresponding to the combination of station and week for which we want to predict $Z(s,t)$. The corresponding prediction variance is

$$\mathrm{V}(\hat{Z}(s,t)) = \boldsymbol{x_p}'\mathrm{D}(\hat{\boldsymbol{\beta}})\boldsymbol{x_p} \tag{B.12}$$

where $\mathrm{D}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}$ is the dispersion matrix of $\hat{\boldsymbol{\beta}}$. In this case the model is still capable of modeling DIN at individual stations, because it uses data from surrounding stations through the week-effect in the mean component, however, it does not consider distances between stations.

If the residuals from this model are in fact spatially correlated we propose to model $\epsilon(s,t)$ by means of a spatial covariance structure $C_{st}(h_s,h_t) = C_s(h_s)$ assuming the temporal correlation in the residuals to be neglectable, see Cressie (1993). We used a spherical model for the spatial covariance structure as this model has shown to be adequate in Lophaven (2001). If the residuals are temporally correlated the covariance structure should include this, e.g. by means of an autoregressive process, see Box and Jenkins (1970). We have chosen not to focus on temporal covariance structures, because the majority of stations had a low sampling frequency and observations were irregularly distributed over the 5 year period (Figure B.2). As a consequence spatial and temporal covariance structures could not be simultaneously estimated. The modeling strategy pursued in this study was first to apply the described model assuming uncorrelated residuals. Secondly the residuals from this model were analysed by computing seasonal spatial experimental semivariograms, where the four seasons are defined as the periods December - February, March - May, June - August and September - November. The experimental semivariogram is used to model the semivariogram $\gamma(h)$, and having a semivariogram model with a sill the transition $C(h) = C(0) - \gamma(h)$ can be used to convert the semivariogam model to the covariance function, see Cressie (1993). The use of a semivariogram model estimated based on an experimental semivariogram for the model residuals results in a biased estimator $\hat{\boldsymbol{\beta}}$, see Cressie (1993). Seasonal spatial experimental semivariograms were computed by considering all possible datapairs for each of the weeks in each of the four seasons. For all datapairs the squared differences were computed and separated into 14 spatial lags each with a width of 10 km. The number and width of the spatial lags were chosen in order to have a reasonable number (>30) of data pairs in each lag. For each lag the average divided by 2 was computed to produce seasonal spatial experimental semivariograms. If the residuals were found to be spatially correlated, the spatial covariance structure was estimated for each of the four seasons, and the mean component (B.7) was subsequently estimated by means of these four seasonal structures, i.e. weeks within the same season had the same spatial covariance structure. In this case the elements of the matrix $\boldsymbol{\Sigma}$ in (B.10) only depend on the spatial distance and direction between monitoring stations. For observations from different weeks the corresponding element in the matrix is zero, while observations from the same week are spatially correlated, and the value of the corresponding matrix element is given by one of the four seasonal covariance structures. A prediction of $Z(s,t)$ is computed as

$$\hat{Z}(s,t) = \boldsymbol{x_p}'\hat{\boldsymbol{\beta}} + \boldsymbol{c}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{B.13}$$

where $\boldsymbol{c}$ is a vector of spatial covariances between the monitoring station at which we want to predict and surrounding monitoring stations, where observations were made in the same week. The elements of $\boldsymbol{c}$ corresponding to observations from different weeks are zero. The corresponding prediction variance is

$$\mathrm{V}(\hat{Z}(s,t)) = \sigma^2 - \boldsymbol{c}'\boldsymbol{\Sigma}^{-1}\boldsymbol{c} + (\boldsymbol{x_p} - \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{c})'(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x_p} - \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{c}) \tag{B.14}$$

The model was investigated using log-transformed observations because of the skewed distribution of DIN. Predictions from the model were back-transformed to the original scale using two different approaches a) back-transform into the mean value, i.e.

$$\hat{Z}_0(s,t) = \exp\{\hat{Z}(s,t) + \mathrm{V}(\hat{Z}(s,t))/2\} \tag{B.15}$$

where $\hat{Z}_0(s,t)$ is the prediction of surface DIN, and b) transform into the median value on the original scale, i.e.

$$\hat{Z}_0(s,t) = \exp\{\hat{Z}(s,t)\} \tag{B.16}$$

The reason for investigating these two approaches was that a great number of predictions were associated with high prediction variances due to the highly irregular sampling in time and space, consequently affecting the back-transform into mean values. Journel (1980) described the use of the back-transform (B.15) when $\hat{Z}(s,t)$ is a kriging estimator. In our model $\hat{Z}(s,t)$ is not a kriging estimator and the work by Journel (1980) is therefore used only for motivation.

The two back-transforms (B.15) and (B.16) combined with models assuming uncorrelated and spatially correlated residuals result in four different methods for modeling DIN, as shown in Table B.1

Name of approach	Model residuals	Back-transform
1a	uncorrelated	into mean
1b	uncorrelated	into median
2a	spatially correlated	into mean
2b	spatially correlated	into median

Table B.1: *The four approaches.*

Furthermore, the models assuming uncorrelated and spatially correlated residuals were compared on the log-scale by means of cross validation. This was carried out by removing log-transformed observations from a single year and station from the estimation data set, and subsequently predicting these based on the estimated model. The procedure was repeated for all possible (N_2) combinations of station and week for which observations exist, and the goodness of the model (GOM) value was computed as

$$\mathrm{GOM} = \frac{1}{N_2}\sum_{j=1}^{N_2}\{\frac{1}{N_1(ij)}\sum_{i=1}^{N_1(ij)}(Z_{ij} - \hat{Z}_{ij})^2\} \tag{B.17}$$

where $N_1(ij)$ is the number of observations removed from the estimation data set for the ij'th combination of year and station, and Z_{ij} and $\hat{Z}_{ij}$ are the log-transformed observations removed and predictions of these, respectively. For the DIN observations in the Kattegat N_2 was 211 and $N_1(ij)$ varied between 1 and 30. The GOM statistic in (B.17) is computed on the log-scale, and does therefore not take the two different back-transforms into account. Consequently, the GOM statistic is not influenced by the bias introduced when applying the back-transform in (B.16). Other possible statistics could be used instead of or in addition to the GOM, e.g., the errors in the GOM could be normalized by dividing the error by the prediction standard deviation.

B.4 Results and discussion

B.4.1 Data preparation

At eight of the monitoring stations DIN was observed only once during the entire study period. Consequently, this single observation will have a high influence on the predictions at the remaining weeks for this particular station. Furthermore, the estimate of this particular station level will be associated with a large uncertainty that will affect the back-transform to DIN mean concentration levels. For example, a relatively high DIN concentration at a station with only one observation would lead to overestimation of the station effect for this single station. Therefore data from the eight monitoring stations with only one observation of DIN were removed prior to applying the proposed model. We also investigated discarding stations with two or three DIN observations, as well as for weeks with one observation. This would reduce the model dimension by another eight stations and 24 weeks. Weeks with only one observation did not influence the predictions to the same extent as stations with one observation, because the within-stations variation was larger than the within-weeks variation, and it was therefore not neccesary to remove observations from weeks with only one observation. Furthermore, we did not find that observations from stations with two or three observations caused an over- or underestimation of the station levels for these stations. Consequently, these observations were not removed from the dataset. In summary, only DIN data from stations with one observation were discarded before the model was employed on 57 stations and 260 weeks.

B.4.2 Modeling of space-time data

Surface DIN concentrations in Kattegat are expected to be high during wintertime corresponding to low algae production in this period. In spring enhanced light conditions and increasing temperatures causes growth of algae, and consequently DIN becomes depleted from the surface layer. DIN concentrations remain low during summer when algae production is nitrogen limited, increasing again with the first autumn storms when nutrient-rich bottom water is entrained into the surface layer by increasing winds and buoyancy. The four proposed methods for space-time modeling (Table B.1) of DIN resulted in predicted time series that all described the expected seasonal behavior of DIN at the three stations of interest (Figures B.3-B.4 and B.6-B.7).

The main difference between the time series for the different stations is the mean DIN concentration level, described by $station_k$ in (B.7). Estimates of $station_k$ revealed that mean DIN levels ranged from 0.42 to 3.71 for the 57 stations with highest values at coastal stations, in particular, along the Jutland coast.

DIN peaks in the time series were occasionally predicted by all four approaches, which were due to high concentrations at surrounding stations observed within the same week and partially due to high prediction variance for the mean back-transform (B.15). There are three gaps in the time series, when DIN was not observed at any station in the Kattegat and consequently the DIN level could not be estimated for these particular weeks. However, the model can be used for the time series for the many monitoring stations with few observations (e.g. SI2).

When analysing sparse, irregularly sampled data the prediction variance, given by (B.12) or (B.14), might be substantially influencing the back-transform into the mean value (B.15). The back-transform into median values (B.16) resulted in less spiked dynamics although not all peaks were removed. As an example, the high peaks in winter time 1996 at station 20004 and SI2 by the mean back-transform were reduced substantially by the median back-transform. The Kattegat is a relatively large marginal sea and the physical and biogeochemical processes acting on DIN levels, e.g. nutrient input from land and atmosphere, remineralisation of organic matter or upwelling/entrainment of nutrient-rich

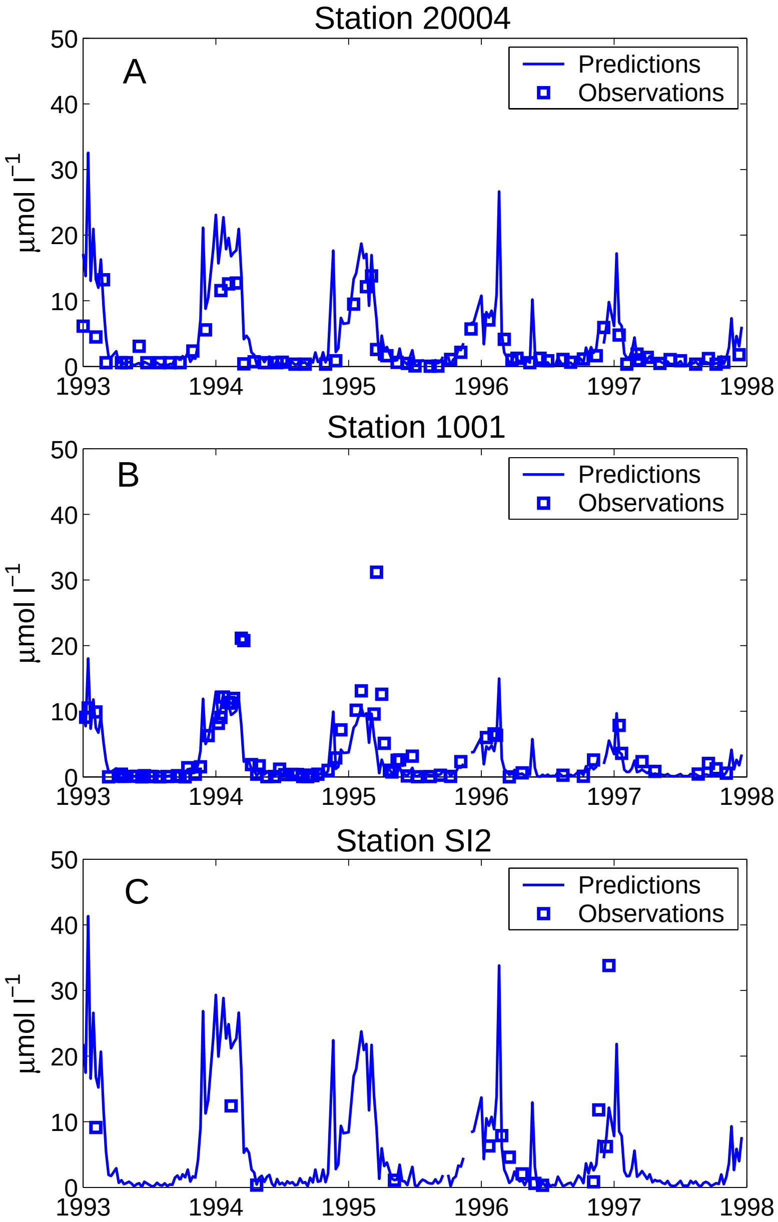

Figure B.3: *Modeling of time series of DIN assuming uncorrelated residuals. Predictions were back-transformed into the mean value (approach 1a). A) Coastal station. B) Open-water station. C) Station with few observations.*

bottom water into the surface layer, cannot account for the spiked behavior modeled using the mean back-transform, because these processes take time, and hence, the median back-transform appears to produce more realistic predictions of surface DIN levels.

The seasonal spatial experimental semivariograms showed that the residuals within weeks were spatially correlated (Figure B.5), i.e. a spatial covariance structure should be included in the model (B.8). The experimental semivariograms were computed using a maximum distance of 120 kilometres, lag distances of 10 kilometres, and distance tolerances of 5 kilometres. Seasonal semivariogram models were estimated by weighted least squares regression of the experimental semivariogram, where the weights are the number of data pairs in each spatial lag, and these all have a range and a sill/nugget effect - ratio which is of the same order of magnitude, supporting the assumption of a time-independent spatial covariance structure, i.e. the spatial dependency was formulated for each season but applied separately for each week in the five-year period. This means that the parameters of the spherical semivariogram model changed from season to season (Figure B.5), and that these models were used for the respective seasons for all years. However, in our model the seasonal semivariogram models were applied to each week, and thereby we assume that the spatial variation on a weekly basis is the same for all weeks within the respective seasons. The reason is that weekly semivariogram models could not be estimated due to lack of available data on a weekly basis (Figure B.2). When modeling DIN from a given week observations from other weeks were not included in the covariance structure, i.e. elements of $\boldsymbol{\Sigma}$ and $\boldsymbol{c}$ in (B.13) corresponding to observations from other weeks were zero. However, observations from other weeks were used to estimate the effect of $station_k$ in (B.8). Directional experimental seasonal semivariograms have been estimated in the same way as described above for the omnidirectional experimental seasonal semivariograms. These showed that the experimental seasonal semivariograms did not depend on direction, and anisotropy was therefore not included in the spatial dependency. The results, when using a spherical spatial covariance structure, as formulated above, in the model (B.8), are shown in Figures B.6 and B.7 for the same three stations as in Figures B.3 and B.4. It is seen that a larger week to week variation in DIN concentrations was predicted when including the spatial covariance structure.

The two approaches assuming uncorrelated and spatially correlated residuals were compared statistically on the log-scale by cross validation. Results using the goodness of model (GOM) measure given by (B.17) are shown in Table B.2, and it is seen that the spherical covariance structure improved the GOM relative to assuming uncorrelated residuals.

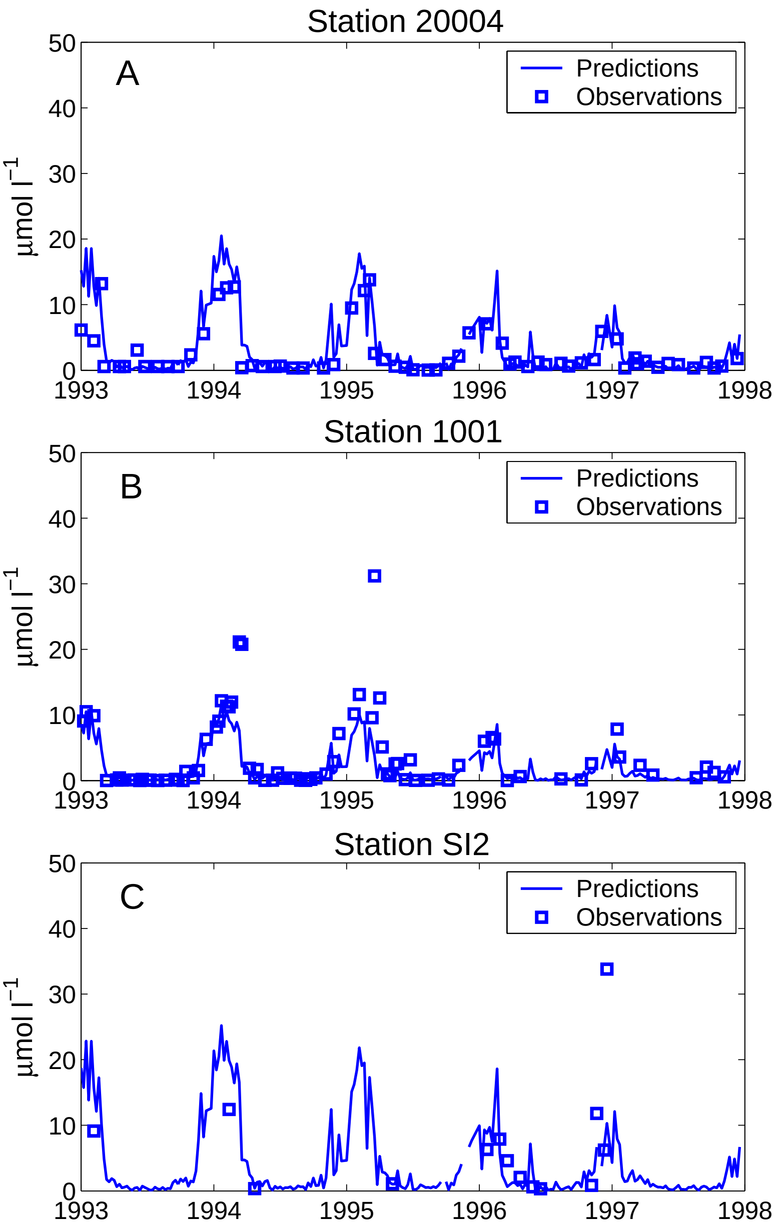

Figure B.4: *Modeling of time series of DIN assuming uncorrelated residuals. Predictions were back-transformed into the median value (approach 1b). A) Coastal station. B) Open-water station. C) Station with few observations.*

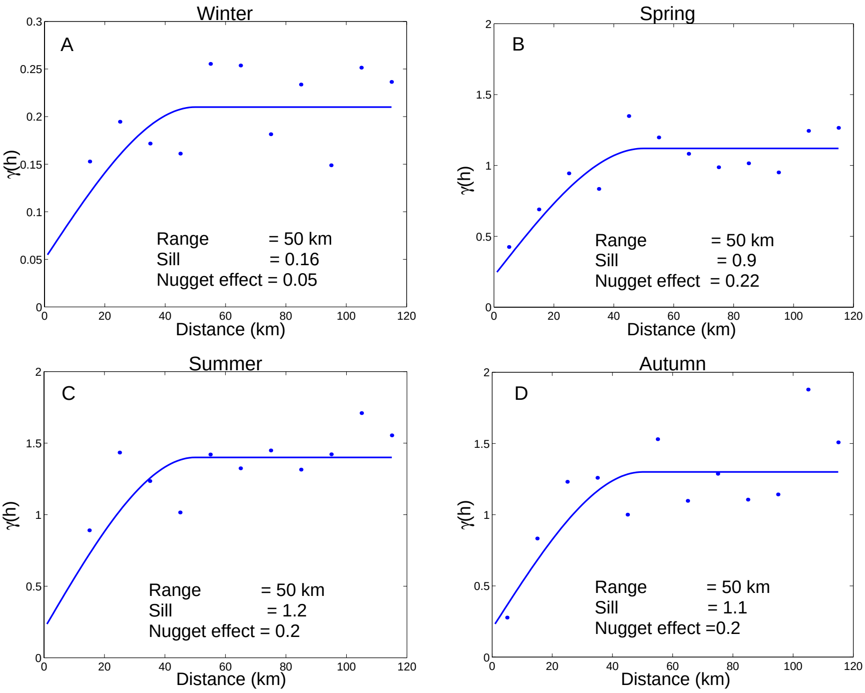

Figure B.5: *Seasonal experimental semivariograms and spherical semivariogram models based on residuals from approach 1a. A) December - February, B) March - May, C) June - August and D)September - November.*

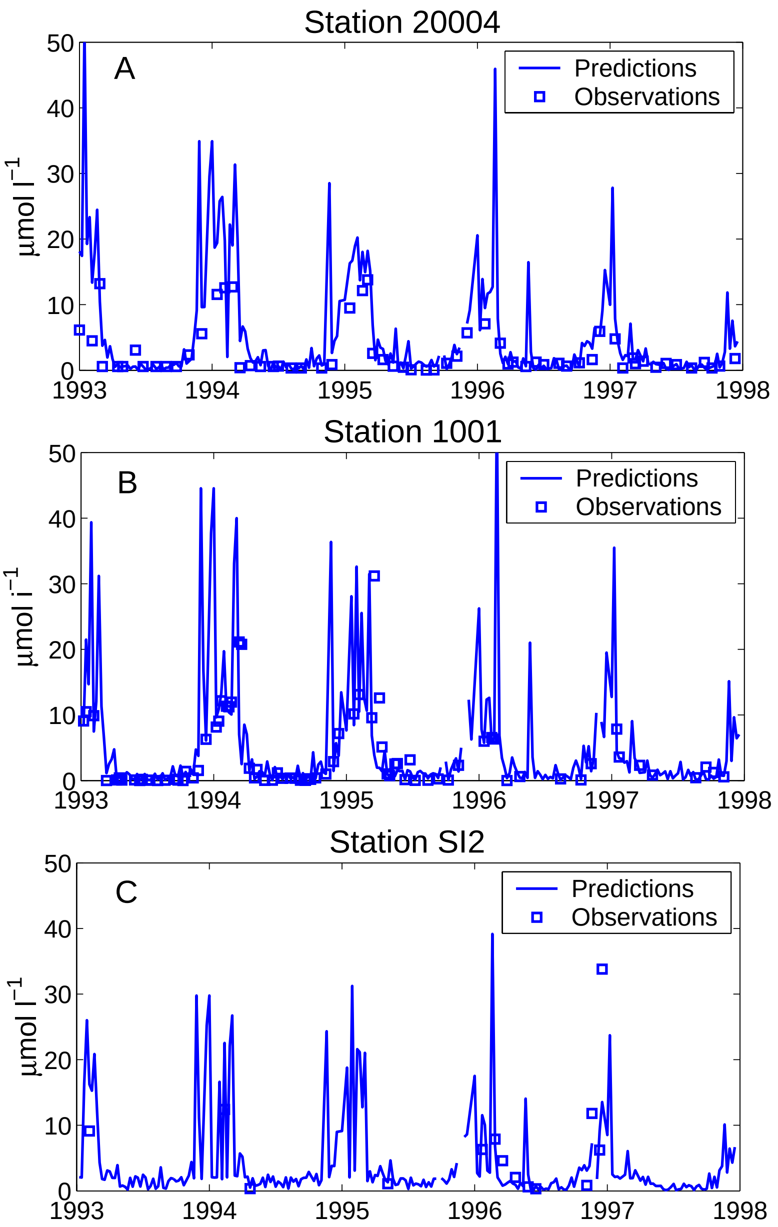

Figure B.6: *Modeling of time series of DIN using a spherical spatial covariance structure. Predictions were back-transformed into the mean value (approach 2a). A) Coastal station. B) Open-water station. C) Station with few observations.*

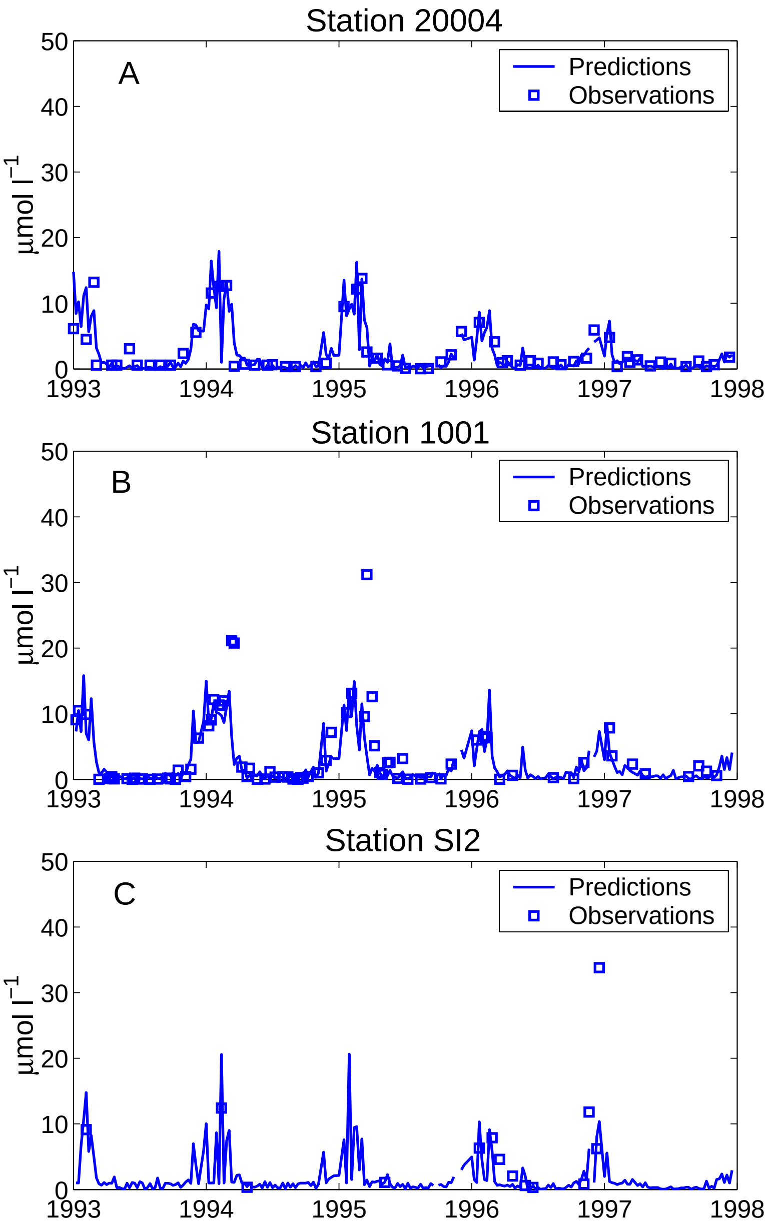

Figure B.7: *Modeling of time series of DIN using a spherical spatial covariance structure. Predictions were back-transformed into the median value (approach 2b). A) Coastal station. B) Open-water station. C) Station with few observations.*

Model	Goodness of model
Assuming uncorrelated residuals	3.59
Spherical covariance structure	1.66

Table B.2: *Performance of two models compared on the log-scale.*

B.5 Application of modeled space-time data

We demonstrate the application of modeled time series by computing weekly maps of the surface DIN concentration in the Kattegat area (Figure B.8). Maps of the corresponding kriging standard deviations are shown in Figure B.9. Three different weeks, with different numbers of observations (Table B.3), were chosen, and spatial predictions were obtained by ordinary kriging using a unique neighborhood, see Cressie (1993), and a spherical semivariogram model.

Week	Number of observations
Second week of July 1993	13
First week of April 1994	5
Third week of May 1994	10

Table B.3: *The number of observations in the three weeks for which the maps in Figure B.8 was computed.*

Spatial predictions were calculated based on both the raw DIN observations and on the combination of DIN observations and model predictions. In the latter case the semivariograms were reestimated using DIN observations combined with model predictions, while it was not possible to reestimate the semivariograms using only the raw DIN observations, due to the small number of weekly observations. Consequently, the same semivariogram models, as those reestimated from DIN observations combined with model predictions, were used when calculating spatial predictions based on DIN observations. Moreover, when calculating spatial predictions and the corresponding kriging standard deviations based on DIN observations combined with model predictions the uncertainty in the model predictions was not incorporated, which affects the kriging standard deviations mapped in Figures B.9B, D and F. Maps based on DIN observations combined with model predictions were based on a larger amount of data (observed and predicted) and therefore include substantially more detail. In fact, only the map produced from 13 observations in a week in July 1993 (Figure B.8A) was capable of describing the spatial variations with appropriate detail, i.e. similar to the maps computed based on the combination of DIN observations and model predictions. The maps from weeks in April and May 1994 were based on 5 and 10 observations only, resulting in large areas of the Kattegat with a predicted DIN level close to the mean value. Relatively few weeks have 13 or more DIN observations (Figure B.2A), and the spatial variation may consequently be investigated for very few periods unless DIN observations are aggregated to lower temporal resolution, e.g. months or seasons.

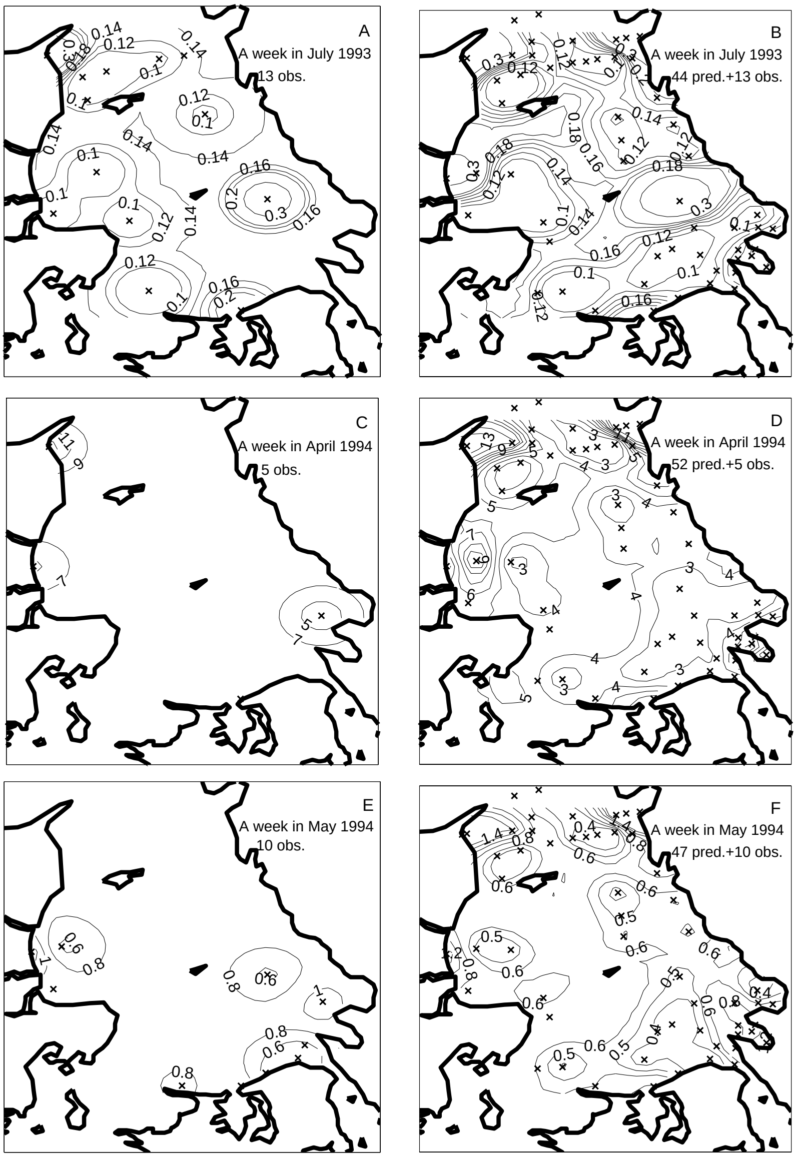

Figure B.8: *Weekly maps of surface DIN in the Kattegat predicted by means of ordinary kriging. The mapping in the left panel was based on raw DIN observations, whereas the mapping in the right panel was based on model predictions combined with DIN observations using approach 2b.* × *indicate locations of stations. Corresponding iso-lines were used for plotting A and B, C and D as well as E and F.*

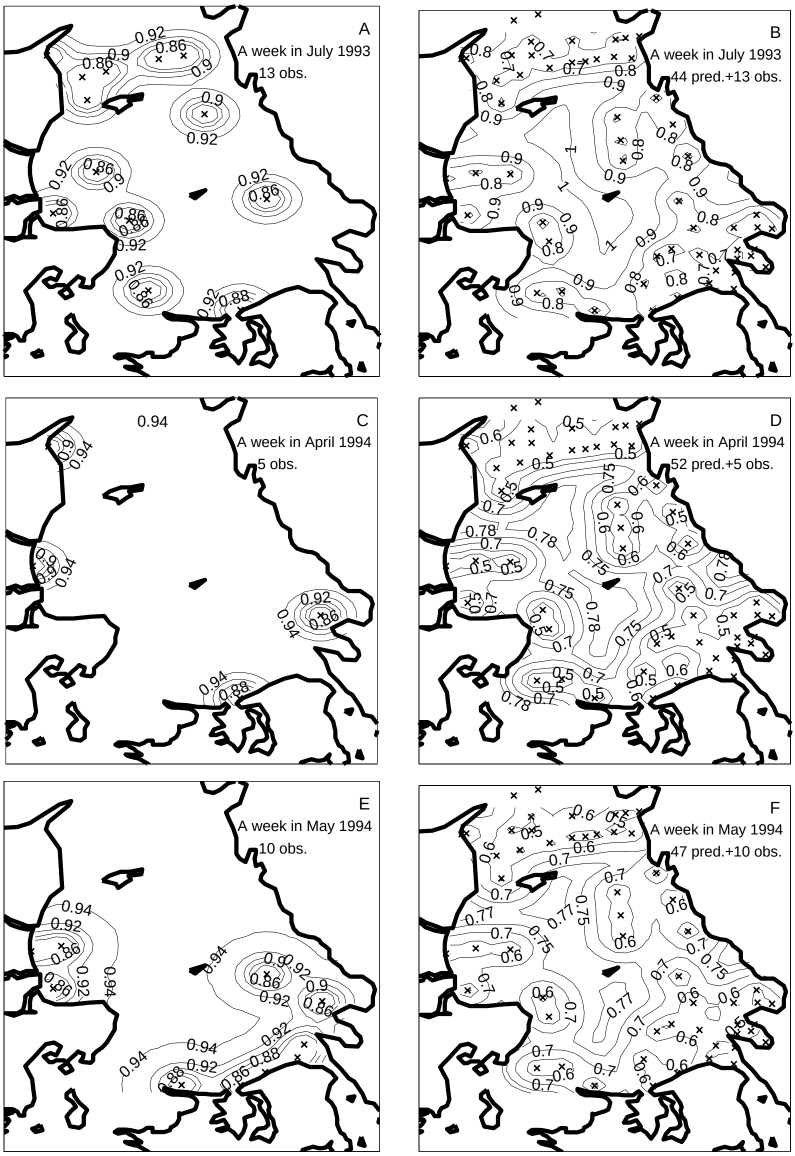

Figure B.9: *Weekly maps of the kriging standard deviation corresponding to the maps of surface DIN shown in Figure B.8. The mapping in the left panel was based on raw DIN observations, whereas the mapping in the right panel was based on model predictions combined with DIN observations using approach 2b.* × *indicate locations of stations.*

The maps computed from DIN observations combined with model predictions tend to have a common pattern (Figure B.8) which is mainly due to the station means of the model (B.8). High concentrations were predicted along the coast of Jutland due to sediment interactions in the shallow water, upwelling and discharges of nutrient-rich water from tributaries. The high DIN concentrations in the north eastern part of the Kattegat were caused by discharge of nutrient-rich water from Göta River and the Skagerrak - Kattegat frontal system.

B.6 Conclusion

This paper describes modeling of sparsely sampled space-time data. Environmental monitoring programs very often lead to such datasets, which is due to a limited amount of economical resources. It is important to consider and use the correlation in data in order to extract as much information as possible from the sparsely sampled monitoring data. The model in the present study is formulated using the general decomposed form, i.e. the sum of a mean and a residual component. The mean and residual components describe the space-time trend of the response variable and the fluctuations around the mean, respectively. For modeling of sparsely sampled data we recommend discretizing the mean component so that it describes variations between monitoring stations and time intervals by means of indices for each monitoring station and each time interval. All other space-time models that have been found in the readily available literature apply a continuous function for modeling the mean component. The proposed model is generally applicable for many types of environmental data reflecting space-time interaction, as exemplified in the present study by DIN concentrations in the Kattegat.

Acknowledgements

This study was in parts carried out under the STAMP project funded by the Nordic Council of Ministers (Nordic working group for monitoring and data analysis). We thank Aleksander Toompuu and Åsa Danielsson for valuable suggestions in the modeling phase. Data were supplied by the National Environmental Research Institute in Denmark and the Swedish Meteorological and Hydrological Institute. We also thank two anonymous referees for their many helpful comments.

References

Andersson, L. and Rydberg, L. (1988). Trends in nutrient and oxygen conditions within the Kattegat: Effects of local nutrient supply. *Estuarine, Coastal and Shelf Science*, **26**, 559–579.

Box, G. and Jenkins, G. (1970). *Time series analysis, forecasting and control.* Holden-Day, San Francisco.

Brown, P., Diggle, P., Lord, M., and Young, P. (2001). Space-time calibration of radar rainfall data. *Journal of the Royal Statistical Society, series C*, **50**, 221–241.

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.

Cressie, N. (1993). *Statistics for spatial data.* Wiley, New York.

Cressie, N. and Huang, H. (1999). Classes of nonseparable, spatiotemporal stationary covariance functions. *Journal of the American Statistical Association*, **94**, 1330–1340.

De Cesare, L., Myers, D., and Posa, D. (2001a). Estimating and modelling space-time correlation structures. *Statistics & Probability Letters*, **51**, 9–14.

De Cesare, L., Myers, D., and Posa, D. (2001b). Product-sum covariance for space-time modelling: an environmental application. *Environmetrics*, **12**, 11–23.

De Cesare, L., Myers, D., and Posa, D. (2002). FORTRAN programs for space-time modeling. *Computers & Geosciences*, **28**, 205–212.

De Iaco, S., Myers, D., and Posa, D. (2001). Space-time analysis using a general product-sum model. *Statistics & Probability Letters*, **52**(1), 21–28.

De Iaco, S., Myers, D., and Posa, D. (2002). Nonseparable space-time covariance models: Some parametric families. *Mathematical Geology*, **34**(1), 23–42.

Granéli, E. (1987). Nutrient limitation of phytoplankton biomass in a brackish water bay highly influenced by river discharge. *Estuarine, Coastal and Shelf Science*, **25**, 563–569.

Gustafsson, B. (2000). Time-dependent modeling of the Baltic entrance area. 1. quantification of circulation and residence times in the Kattegat and the straits of the Baltic sill. *Estuaries*, **23**, 231–252.

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

Journel, A. (1980). The lognormal approach to predicting local distributions of selective mining unit grades. *Mathematical Geology*, **12**, 285–303.

Kronvang, B., Ærtebjerg, G., Grant, R., Kristensen, P., Hovmand, M., and Kirkegaard, J. (1993). Nationwide monitoring of nutrients and their ecological effects: State of the Danish aquatic environment. *Ambio*, **22**, 176–187.

Lophaven, S. (2001). Reconstruction of data from the marine environment. Master's thesis, Technical University of Denmark, Kgs. Lyngby (Denmark).

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

Appendix C

Space-time modeling of dissolved inorganic nitrogen

The paper is published in:

Lophaven, S., Carstensen, J., and Rootzén, H. (2003). Space-time modeling of dissolved inorganic nitrogen. In *Bulletin of the International Statistical Institute 54th Session*, Contributed Papers, vol. LX(1), pp. 749-750, International Statistical Institute, Berlin, Germany, August 13-20, 2003.

Space-time modeling of dissolved inorganic nitrogen

Søren Lophaven[1], Jacob Carstensen[2], and Helle Rootzén[1]

[1] Informatics and Mathematical Modelling, Technical University of Denmark
[2] Department of Marine Ecology, National Environmental Research Institute of Denmark

Abstract

This paper describes an approach for predicting space-time phenomena. The method is tested on observations of dissolved inorganic nitrogen in the Kattegat. The covariance structure is modelled as a function of both space and time, assuming that it can be separated into a spatial and a temporal component. Results are presented for 2 of 65 monitoring stations in the Kattegat, and show that the model is capable of predicting the temporal dynamics of dissolved inorganic nitrogen at these 2 stations.

KEY WORDS: *Space-time semivariogram, separable covariance structure, kriging*

C.1 Introduction

This paper describes an approach for modeling data measured at a given time and location. The proposed modeling approach is applied to observations from the Kattegat of dissolved inorganic nitrogen (DIN), which is the sum of the following nitrogen constituents: ammonium (NH_4^+-N), nitrite (NO_2^--N) and nitrate (NO_3^--N). DIN is an important parameter, because algae growth in the Kattegat is generally nitrogen limited. Observations have been made at 65 monitoring stations during 1993-1997.

C.2 The modeling approach

The applied model is of the decomposed form

$$Z(s,t) = \mu(t) + \varepsilon(s,t) \tag{C.1}$$

where Z is log-transformed DIN, which depends on the location s, given by a x- and y-coordinate, and on time t. μ is the mean component, and ε is the residual component describing fluctuations around the mean in space and time. The mean component describes the temporal dynamics of DIN, and is modeled as

$$\mu_{ij}(t) = \alpha_i + \beta_j + \delta_1 \sin\left(\frac{2\pi t}{52} + \varphi_1\right) + \delta_2 \sin\left(\frac{2\pi t}{26} + \varphi_2\right) \tag{C.2}$$

where α_i are the monitoring station effects, and β_j the year effects where the years were defined as starting and ending the first of July. This was done because the concentrations of DIN are low in the summer period, and the discontinuity, introduced by the year effects β_j, in the limit from year to year is therefore reduced. Furthermore, δ_1, δ_2, φ_1 and φ_2 are the yearly amplitude, the half-year amplitude, the yearly phase shift, and the half-year phase shift, respectively, while t is time given in weeks. The residual component is assumed to be a second order stationary stochastic process with expected value and a covariance function given by

$$\begin{aligned} \mathrm{E}(\varepsilon(s,t)) &= 0 \qquad\qquad (\text{C.3})\\ \mathrm{C}_{st}(h_s,h_t) &= \mathrm{Cov}(\varepsilon(s+h_s,t+h_t),\varepsilon(s,t)) \end{aligned}$$

One way to model the covariance in (3) is to separate it into a spatial and a temporal component, e.g. by using the product model given by

$$\mathrm{C}_{st}(h_s,h_t) = \mathrm{C}_s(h_s)\mathrm{C}_t(h_t) \tag{C.4}$$

A discussion of the product model and of other separable space-time covariance models can be found in De Cesare et al. (2001a). The product model was applied in Haas (1995) for modeling of wet sulfate deposition.

C.3 Results

In Figure C.1 results are presented for the two monitoring stations 1001 and 20004, representing open-sea and coastal stations, respectively. For station 1001 the model predictions seem to fit the observations quite well, and the model is able to predict the high DIN peaks during winter. For station 20004 these peaks are also predicted even though the observed concentrations are much smaller. This is caused by high winter concentrations at the surrounding stations.

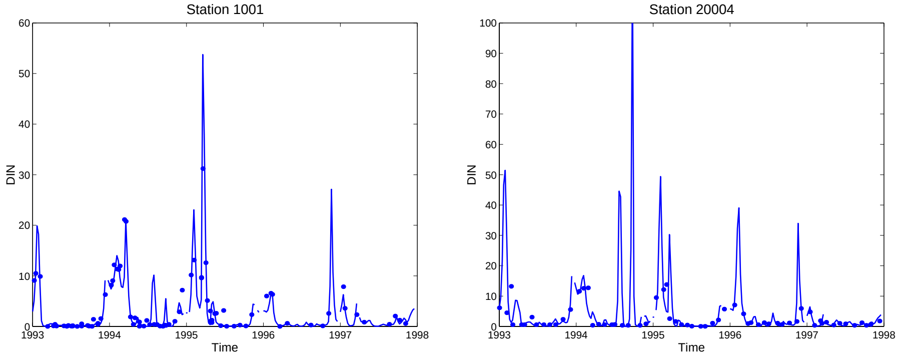

Figure C.1: *Predictions and observations of DIN at the two monitoring stations 1001 and 20004 in the Kattegat.*

References

De Cesare, L., Myers, D., and Posa, D. (2001a). Estimating and modelling space-time correlation structures. *Statistics & Probability Letters*, **51**, 9–14.

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.

Appendix D

Stochastic modelling of dissolved inorganic nitrogen

The paper is submitted:

Lophaven, S., Carstensen, J., and Rootzén, H. (2004). Stochastic modelling of dissolved inorganic nitrogen. *Ecological Modelling.*

Stochastic modelling of dissolved inorganic nitrogen

Søren Lophaven[1], Jacob Carstensen[2,3], and Helle Rootzén[1]

[1] Informatics and Mathematical Modelling, Technical University of Denmark
[2] Department of Marine Ecology, National Environmental Research Institute of Denmark
[3] Inland and Marine Waters Unit, Joint Research Centre, TP 280, I-21020 Ispra, Italy

Abstract

Environmental monitoring datasets often contain a large amount of missing values, and are characterised as being sampled over time on a distinct number of locations in the area of interest. This paper proposes a stochastic approach for modelling such data in space and time, by taking the spatial and temporal correlations in data into account. It has been applied to observations of dissolved inorganic nitrogen in the Kattegat during the period 1993-1997. Modelling results are shown as maps of the spatial distribution of DIN in four weeks, representing the four seasons, and as time series of DIN at three different locations. However, the model approach could be applied to any space-time point given by a location in the Kattegat area and a week in the five-year period 1993-1997. The results can be interpreted from a biological and physical point of view. Thus for the specific application the approach seems to perform very well. The results obtained could be used to improve status reporting of the environment, or as forcing functions for time series models and deterministic, hydrodynamic ecosystem models.

KEY WORDS: *dissolved inorganic nitrogen, geostatistics, space-time modelling, the Kattegat*

D.1 Introduction

Environmental monitoring programs have been established in many industrialized countries to assess the magnitude and consequences of human stresses on the environment. It is important that the monitoring data are exploited to the fullest extent for optimal use of the limited resources available for environmental monitoring.

Environmental processes reflect temporal and spatial variations on a variety of scales that are only partially captured in monitoring data. In particular, the marine environment comprises a complex mosaic of interacting processes, which range from small-scale microbial processes to global-scale oceanic circulation. On the other hand, monitoring at sea by traditional shipboard sampling is associated with large costs for personnel and equipment, and new technologies aiming at reducing costs have not yet proven adequate to substitute for monitoring vessels. Consequently, the spatial and temporal coverage of data is limited and often irregular, and temporal and spatial variations can only be assessed on a coarse resolution scale, unless methods are employed that integrate monitoring data in time and space.

The aim of this study is to describe and apply a statistical approach, which can be used for modelling space-time phenomena by taking account of the temporal as well as spatial correlation in data. The proposed model is able to cope with a high number of missing values, and it is applied to dissolved inorganic nitrogen (DIN) observed in the Kattegat (Figure D.1). In principle the proposed method is general, and could be applied to other types of environmental monitoring data as well, e.g. air pollution and climate data. The model predictions provide an improvement for reporting the state of the environment and for assessing effects of proposed nutrient reductions by means of statistical analyses. Moreover, model predictions combined with observations can be applied as forcing functions for time series models and deterministic, hydrodynamic ecosystem models. This will advance the knowledge of the biogeochemical processes in the marine environment, and reduce uncertainties of regional nutrient and carbon budgets.

D.2 Study area - The Kattegat

During the 1980s numerous episodes of oxygen deficiency, covering large areas of the Danish estuaries, were observed (Kronvang et al., 1993). This resulted in the adoption of the Action Plan on the Aquatic Environment in 1987, which required that total discharge of nitrogen from diffuse sources (agriculture) and point sources (municipal wastewater treatment plants and industrial outfalls) were to be reduced by 50 % from a total of 290,000 tonnes per year in 1987 to around 145,000 tonnes per year in 1993. During the same period phosphorus discharges were to be reduced by 80 % from a total of approximately 12,000 tonnes per year to 2,200 tonnes per year. In connection with the adoption of this plan a monitoring program, the Danish National Aquatic Monitoring and Assessment Program (DNAMAP), was established. The purpose of the program was to characterise the state of the aquatic environment and to document the effects of the measures being taken to reduce nutrient inputs to the marine environment (Kronvang et al., 1993).

The Kattegat basin is a transition zone between the North Sea and the Baltic Sea (Figure D.1A) with a surface area of 22,290 km^2, a volume of 533 km^3 and a mean depth of approximately 24 meters (Gustafsson, 2000). The area is dominated by advective transport of low-saline water from the Baltic Sea as a surface current and water with a high salinity from the North Sea as a bottom current. This advection creates a strong salinity stratification located at 15-20 meters depth throughout most of the year (Andersson and Rydberg, 1988).

D.3 Data material

The observations used for this study were sampled at 65 stations in the Kattegat (Figure D.1B) during a five-year period (1993 - 1997) by Danish and Swedish authorities. Various water quality parameters were measured from the samples taken at various depths within the water column.

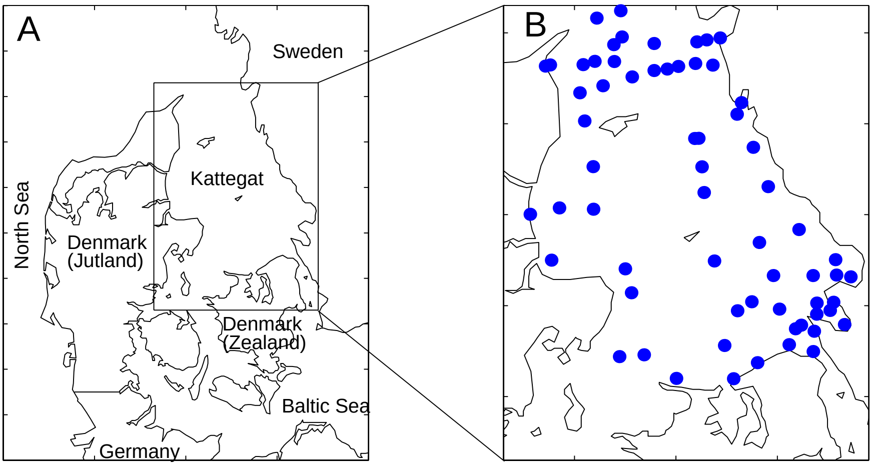

Figure D.1: *A) The Kattegat is the transitional area between the North Sea and the Baltic Sea. B) Locations of the 65 monitoring stations (•) with DIN observations.*

In this study we have chosen to focus on surface concentrations (0-10 m) of DIN. DIN is the inorganic form of nitrogen that is used as nitrogen source by the primary producers. The concentration of DIN in the surface layer is mainly controlled by the biological activity through uptake, regeneration processes transforming organic nitrogen into DIN, mixing between surface and deeper waters across the pycnocline, and direct inputs from land and atmosphere. Normally the concentration in the surface layer is relatively low (0-2 μmol l^{-1}) during the productive season (March-September), whereas DIN accumulates in the surface waters until the onset of the spring bloom. DIN is an important monitoring parameter, because primary production in the Kattegat is considered to be nitrogen limited (Granéli, 1987). In this study the DIN data comprised 1832 observations scattered over 5 years and 65 stations with only a few weeks having more than 20 surface values at various stations, and 60 % of the weeks had fewer than 8 observations. Many stations had fewer than 15 observations over the entire five-year period and only 4 stations were sampled more than 100 times corresponding to biweekly sampling. Considering that the data matrix consisted of 65 stations and 260 weeks, i.e. a data matrix of 65 × 260 = 16900 cells, then the actual DIN data accounted for approximately 10 % of the cells. Thus, given that a weekly resolution was desired, the dataset was characterized by having a high number of missing values.

D.4 Modelling DIN in space and time

This section describes a simple statistical approach which utilize the spatial and temporal correlation in the DIN data, and can handle missing data values. This kind of model was used, rather than a deterministic approach, because of the complexity of DIN dynamics in the Kattegat. The model enables us to predict the DIN concentration in both space and time. Until recently, space-time processes has been a relatively unexplored research area. However, over the last decade a number of statistical methods for analyzing and modelling space-time data have been proposed. A thorough review of geostatistical space-time models is given in Kyriakidis and Journel (1999), while more recent applications were described by Brown et al. (2001); De Cesare et al. (2001a); Figueira et al. (2001). The DIN model that we propose is of the general decomposed form

$$Y(x,t) = m(x,t) + \epsilon(x,t) \tag{D.1}$$

where $Y(x,t)$ is the log-transform of DIN. Thus, the approach was applied to log-transformed DIN, because this change of scale was found to improve modelling. Moreover, $m(x,t)$ is the mean component modelled as a deterministic function depending on space $x = (x_1, x_2)$ and time t, and $\epsilon(x,t)$ is the residual component describing fluctuations around the mean in space and time.

The DIN modelling was separated into three steps. In the first step the mean component was estimated at the 65 monitoring stations by assuming uncorrelated residuals. In the second step the mean component at the monitoring stations was extended to the whole Kattegat area. Finally, in step three the residuals from the first step were analysed and modelled using a separable space-time covariance structure.

D.4.1 Step 1

The mean component was used to describe the temporal dynamics of dissolved inorganic nitrogen at the 65 monitoring stations in the Kattegat. This meant that in the first step the mean component was not extended to non-sampling locations. The mean component was given by

$$m_{ij}(t) = \alpha_i + \beta_j + \delta_1 \sin\left(\frac{2\pi t}{52} + \varphi_1\right) + \delta_2 \sin\left(\frac{2\pi t}{26} + \varphi_2\right) \tag{D.2}$$

where α_i were the station effects, and β_j the year effects where the years were defined as starting and ending the first of July. This was done because the concentrations of DIN were low in the summer period, and the discontinuity, introduced by the year effects β_j, in the limit from year to year was therefore reduced. Furthermore, δ_1, δ_2, φ_1 and φ_2 were the yearly amplitude, the half-year amplitude, the yearly phase shift, and the half-year phase shift, respectively, while t was time given in weeks. Hence, a temporal resolution of one week is used. The mean component was denoted $m_{ij}(t)$ to indicate that it corresponded to the ith monitoring station and the jth year, and was a function of time t. Thus, the mean component is given by a level for each monitoring station and year in addition to some seasonal variation. Equation (D.2) can be rewritten using the addition formulas

$$\begin{aligned} m_{ij}(t) &= \alpha_i + \beta_j + \delta_1\left(\sin\left(\frac{2\pi t}{52}\right)\cos(\varphi_1) + \cos\left(\frac{2\pi t}{52}\right)\sin(\varphi_1)\right) \\ &\quad +\delta_2\left(\sin\left(\frac{2\pi t}{26}\right)\cos(\varphi_2) + \cos\left(\frac{2\pi t}{26}\right)\sin(\varphi_2)\right) \end{aligned} \tag{D.3}$$

and by introducing four new parameters: $\psi_1 = \delta_1\cos(\varphi_1)$, $\psi_2 = \delta_1\sin(\varphi_1)$, $\psi_3 = \delta_2\cos(\varphi_2)$, $\psi_4 = \delta_2\sin(\varphi_2)$, it is seen that the model is linear in the parameters. In the first step, the parameters of the mean component were estimated using the general linear model

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y} \tag{D.4}$$

where $\hat{\boldsymbol{\beta}}$ was a 75×1 vector of estimated parameters $(\alpha_1,\ldots,\alpha_{65},\beta_1,\ldots,\beta_6,\psi_1,\ldots,\psi_4)$, $\boldsymbol{X}$ was a 1832×75 design matrix, and $\boldsymbol{y}$ was a 1832×1 vector containing log-transformed values of DIN. The use of (D.4) implied that the residuals in step 1 were assumed to be independent, and identically distributed, $\epsilon_{ij} \in N(0, \sigma_0^2)$, where σ_0^2 is estimated by

$$\hat{\sigma}_0^2 = \frac{1}{N-p-1}\sum_{k=1}^{N}(y_i - \hat{y}_i)^2 \qquad \text{(D.5)}$$

In (D.5) $N = 1832$ was the number of observations and $p = 75$ was the number of model parameters. The variance of $\hat{\boldsymbol{\beta}}$ was given by

$$\text{Var}[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X^T X})^{-1}\sigma^2 \qquad \text{(D.6)}$$

After estimation of $\boldsymbol{\beta}$ the mean component was computed by

$$\hat{\boldsymbol{m}} = \boldsymbol{X}_f\hat{\boldsymbol{\beta}} \qquad \text{(D.7)}$$

where each row in the matrix $\boldsymbol{X}_f$ defined for which station, year and week the mean component was to be computed. Variances of the computed mean component were

$$\text{Var}[\hat{\boldsymbol{m}}] = \boldsymbol{X}_f \text{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{X}_f^T \qquad \text{(D.8)}$$

The original parameters in (D.2), which can be physically interpreted, were found from

$$\begin{aligned} \varphi_1 &= \arctan\left(\frac{\psi_2}{\psi_1}\right) \Rightarrow \delta_1 = \frac{\psi_2}{\sin(\varphi_1)} \\ \varphi_2 &= \arctan\left(\frac{\psi_4}{\psi_3}\right) \Rightarrow \delta_2 = \frac{\psi_4}{\sin(\varphi_2)} \end{aligned} \qquad \text{(D.9)}$$

D.4.2 Step 2

In the second step the mean component was extended to non-sampling locations, i.e. $m_{ij}(t) \rightarrow m_j(x,t)$, by geostatistical modelling of the station effect (Cressie, 1993; Chilès and Delfiner, 1999; Stein, 1999; Wackernagel, 2003). Given the station effects α_i, $i = 1,\ldots,65$ at spatial locations x_i it was assumed that these could be modelled by

$$\alpha_i = S(x_i) + Z_i, \qquad i = 1, ..., 65 \qquad \text{(D.10)}$$

where $S(x)$ was a stationary Gaussian process with expectation $\text{E}[S(x)] = \mu$, variance $\text{Var}[S(x)] = \sigma^2$ and correlation function $\rho(u) = \text{Corr}[S(x_i), S(x_j)]$, $u = \parallel x_i - x_j \parallel$ being the spatial distance between x_i and x_j, and $Z_i \sim N(0, \tau^2)$.

In the current application we had $\mu = 0$, which was due to the nature of the station effects. In this case the station effect at location x_0 was predicted by

$$\hat{\alpha}(x_0) = \sigma^2 \boldsymbol{r}^T (\tau^2 I + \sigma^2 \boldsymbol{R})^{-1} \boldsymbol{\alpha} \tag{D.11}$$

where $\boldsymbol{R}$ was a symmetric 65×65 matrix with elements $\rho(\| x_i - x_j \|)$, $\boldsymbol{r}$ is a 65×1 vector with elements $\rho(\| x_0 - x_i \|)$ and $\boldsymbol{\alpha}$ is a 65×1 vector of station effects. The variances of the station effects α_i in step 1 were estimated by (D.6). These variances were incorporated in the spatial predictions (D.11) by adding them to the corresponding diagonal elements of the covariance matrix $\sigma^2 \boldsymbol{R}$. Estimation of the parameters of the geostatistical model was based on the semivariogram defined by

$$\gamma(u) = \frac{1}{2}\mathrm{Var}[\alpha(x_i) - \alpha(x_j)] = \frac{1}{2}\mathrm{E}[(\alpha(x_i) - \alpha(x_j))^2], \quad u = \| x_i - x_j \| \tag{D.12}$$

The sample semivariogram was estimated by substituting the expectation in (D.12) with an average, i.e.

$$\hat{\gamma}(u) = \frac{1}{2N_0(u)} \sum (\alpha(x_i) - \alpha(x_j))^2 \tag{D.13}$$

where $N_0(u)$ was the number of pairs of data, i.e. station effects. The relationship between the semivariogram and the correlation function was

$$\gamma(u) = \tau^2 + \sigma^2(1 - \rho(u)) \tag{D.14}$$

where τ^2 was the nugget effect, which described small-scale random variation and measurement errors. In the second step the exponential correlation function was used

$$\rho(u) = \exp(-|u|/\phi) \tag{D.15}$$

where ϕ was the correlation parameter. Parameter estimation was based on the sample semivariogram which was fitted with the semivariogram in (D.14) by means of weighted least squares regression, where the weights were the number of datapairs in each distance bin.

D.4.3 Step 3

In the third step the residuals $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{m}}$ from the first step were analysed and modelled using an approach very similar to step 2, but in step 3 distances are measured in both space and time. It was assumed that the residuals were realisations of a second order stationary space-time random field

$$\epsilon = \{\epsilon(x,t), \quad x \in D, \quad t \in T\} \tag{D.16}$$

with expected value $\mathrm{E}[\epsilon(x,t)] = 0$. In this case the space-time covariance function

$$\Sigma_{xt}(u_x, u_t) = \mathrm{Cov}[\epsilon(x_i,t_i), \epsilon(x_j,t_j)] \tag{D.17}$$

and the semivariogram

$$\begin{aligned}\gamma_{xt}(u_x,u_t) &= \frac{\mathrm{Var}[\epsilon(x_i,t_i) - \epsilon(x_j,t_j)]}{2} \\ &= \frac{\mathrm{E}[(\epsilon(x_i,t_i) - \epsilon(x_j,t_j))^2]}{2}\end{aligned} \tag{D.18}$$

depended solely on the lag vector (u_x, u_t). To estimate the experimental space-time semivariogram from data the expectation in (D.18) is replaced by the mean value, yielding

$$\hat{\gamma}_{xt}(u_x,u_t) = \frac{1}{2N_0(u_x,u_t)} \sum (\epsilon(x_i,t_i) - \epsilon(x_j,t_j))^2 \tag{D.19}$$

where $N_0(u_x,u_t)$ was the number of datapairs with spatial distance u_x and temporal distance u_t. When modelling the space-time covariance structure Σ_{xt} we assumed that this could be separated into a spatial and a temporal component in the following way

$$\Sigma_{xt}(u_x,u_t) = \Sigma_x(u_x)\Sigma_t(h_t) \tag{D.20}$$

where $\Sigma_x(u_x)$ and $\Sigma_t(h_t)$ were the spatial and temporal component, respectively (De Cesare et al., 2001b; De Iaco et al., 2001). In this paper we used a common variance σ^2, i.e. $\sigma_x^2 = \sigma_t^2$, which implies that the space-time semivariogram reaches a common level for long separation distances, and exponential correlation functions for both the temporal and the spatial component. In this case the space-time covariance was given by

$$\Sigma_{xt}(u_x,u_t) = \sigma^2 \exp(-|u_x|/\phi_x)\sigma^2 \exp(-|u_t|/\phi_t) \tag{D.21}$$

and the space-time semivariogram with a common nugget effect as

$$\gamma_{xt}(u_x,u_t) = \tau^2 + \sigma^2(\sigma^2 - \sigma^2 \exp(-|u_x|/\phi_x)\exp(-|u_t|/\phi_t)) \tag{D.22}$$

After having estimated the space-time covariance function for the residuals, predictions were computed by

$$\hat{\epsilon}(x_0,t_0) = \sigma_{xt}^2 \boldsymbol{r}_{xt}^T(\tau_{xt}^2 I + \sigma_{xt}^2 \boldsymbol{R}_{xt})^{-1}\boldsymbol{\epsilon} \tag{D.23}$$

As for the station effects the variances of the residuals in step 1 (i.e. $\sigma_0^2\mathbf{1} + \mathrm{Var}[\hat{\boldsymbol{m}}]$) were added to the diagonal of the space-time covariance matrix.

D.5 Results

In this section results for DIN in the Kattegat are presented for four different weeks, representing the four seasons. These weeks were chosen as from January 1997, April 1997, July 1997 and September 1997. For each of these four weeks spatial predictions were computed in a regular grid covering the Kattegat having a grid spacing of 5 kilometers in both directions. Furthermore, modelled time series for three arbitrarily selected locations in the Kattegat are presented. However, the model approach could be applied to any space-time point given by a location in the Kattegat area and a week in the five-year period 1993-1997. When estimating $\hat{\boldsymbol{\beta}}$ in (D.4) we found ψ_1=1.62, ψ_2=0.78, $\psi_3 = 0.20$ and $\psi_4 = -0.21$ and the original parameters of (D.2) were consequently calculated as

$$\begin{aligned}\varphi_1 &= \arctan\left(\frac{0.78}{1.62}\right) = 0.45 \Rightarrow \delta_1 = \frac{0.78}{\sin(0.45)} = 1.80\\ \varphi_2 &= \arctan\left(\frac{-0.21}{0.20}\right) = -0.81 \Rightarrow \delta_2 = \frac{-0.21}{\sin(-0.81)} = 0.29\end{aligned}$$

The estimates of the six year effects $\beta_1, \ldots, \beta_6$ ranged from -0.11 to 0.51, and 65 station effects $\alpha_1, \ldots, \alpha_{65}$ ranged from -1.20 to 2.96. The estimated α_i were in general highest for monitoring stations in the coastal shallow areas and in the northern part (Figure D.2A), whereas the variances of the estimated α_i computed by (D.6) were highest in the northern and eastern parts of the Kattegat (Figure D.2B). These variances are highly dependently on the number of observations at individual stations during the five-year period. In the second step α is extended to non-sampling locations in the Kattegat area by means of a geostatistical model (D.10). Parameter estimation was based on the sample semivariogram (Figure D.2C), as described in section D.4.2. This yielded the estimates $(\tau^2, \sigma^2, \phi) = (0.25, 0.4, 15)$.

Step three consist of modelling the residuals from step 1 by means of space-time geostatistics. To estimate the model parameters in (D.23) we first computed the spatial sample semivariogram $\hat{\gamma}_{xt}(u_x, 0)$. This was done by considering all possible datapairs for each of the weeks. For all datapairs the squared differences of the residuals were computed and separated into spatial bins. For each bin the average divided by 2 was computed to produce the spatial sample semivariogram (Figure D.3A).

From the spatial sample semivariogram it is seen that the residuals were spatially correlated until a separation distance of approximately 60 kilometers. The temporal sample semivariogram was computed in a similar way, but this time we considered all possible datapairs for each of the monitoring stations (Figure D.3B). The residuals were correlated until a temporal separation distance

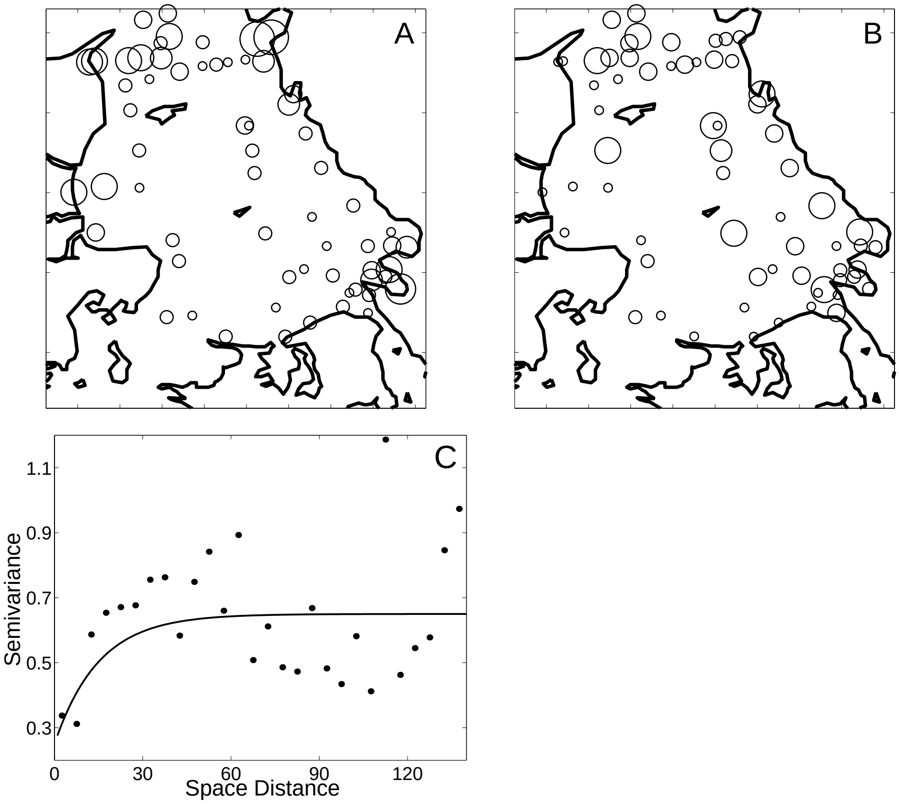

Figure D.2: *A) Estimated station effects ranging from -1.20 to 2.96 indicated by the size of the circle. The locations of monitoring stations are given by the circle centers. B) Standard deviations of the station effects ranging from 0.34 to 1.20 indicated by the size of the circle. C) Sample semivariogram of the station effects.*

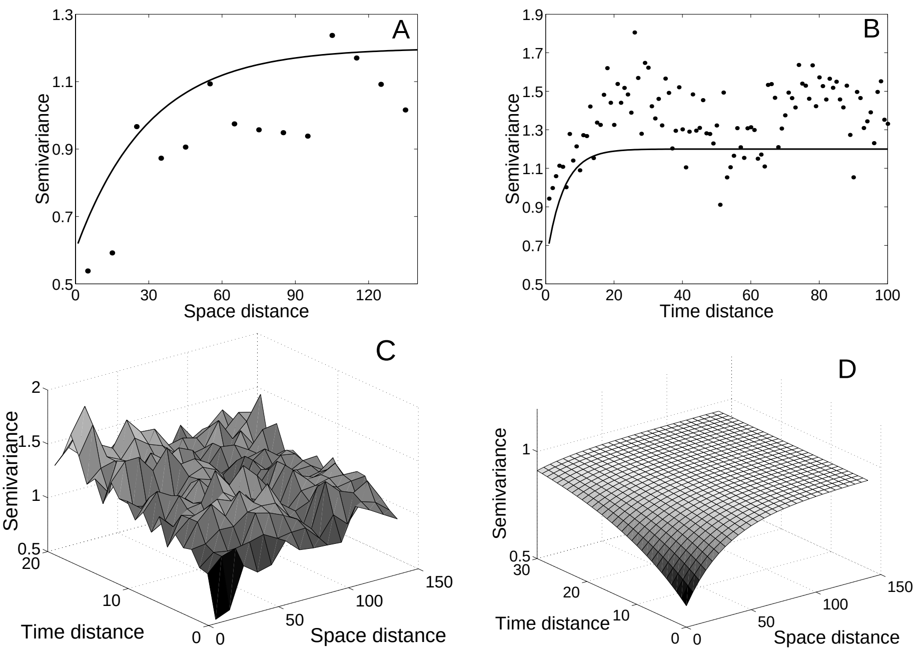

Figure D.3: *A) Spatial sample semivariogram of the residuals B) Temporal sample semivariogram of the residuals C) Space-time sample semivariogram of the residuals D) Space-time semivariogram model corresponding to separability of the space-time dimension.*

of approximately 15 weeks. Furthermore, the plot indicated that some seasonal variation still remained in the residuals. This phenomenon was caused by some high winter concentrations which the estimated mean component did not describe sufficiently. Fitting the two sample semivariograms with common parameters τ^2 and σ^2 gave $(\tau^2, \sigma^2, \phi_x, \phi_t) = (0.6, 0.6, 30, 5)$. When separability according to (D.20) was assumed, these two semivariograms corresponded to the semivariogram surface in Figure D.3D. Comparing the estimated to the sample space-time semivariogram (Figure D.3C and D.3D) the assumption of separability could be examined. Given that the estimated space-time variogram reflected the overall features of the sample space-time variogram we concluded that separability was a valid assumption. Thus, we did not find any reasons for using non-separable space-time covariance structures (Cressie and Huang, 1999; Brown et al., 2000; De Iaco et al., 2002), and consequently the following space-time semivariogram was employed

$$\gamma_{xt}(u_x, u_t) = 0.6 + 0.6(0.6 - 0.6 \exp(-|u_x|/30) \exp(-|u_t|/5)) \tag{D.24}$$

The residuals were predicted using (D.23) and added to the estimated mean component to give predictions on the log-scale. These were back-transformed by means of the exponential function, to yield the results in Figures D.4 and D.5.

D.6 Discussion

Surface DIN concentrations in Kattegat are expected to be high during wintertime corresponding to low primary production in this period. In spring enhanced light conditions and increasing temperatures causes growth of algae, and consequently DIN becomes depleted from the surface layer. DIN concentrations remain low during summer when algae production is nitrogen limited, increasing again with the first autumn storms when nutrient-rich bottom water is entrained into the surface layer by increasing winds and buoyancy. This temporal dynamics is predicted by the model and accounted for by the mean component (Figure D.4).

Furthermore, high concentrations are expected in the coastal areas, in particular along the eastern coast of Jutland, in the north eastern part of the Kattegat near Gothenburg, as well as near the Swedish coast in the south eastern part of the Kattegat. This expected pattern is due to sediment interactions in the shallow water, upwelling and discharges of nutrient-rich water from tributaries. The high concentrations in the north eastern part of the Kattegat are associated with discharges of nutrient-rich water from the Göta River and the Skagerrak - Kattegat frontal system. The described spatial pattern fits very well with the

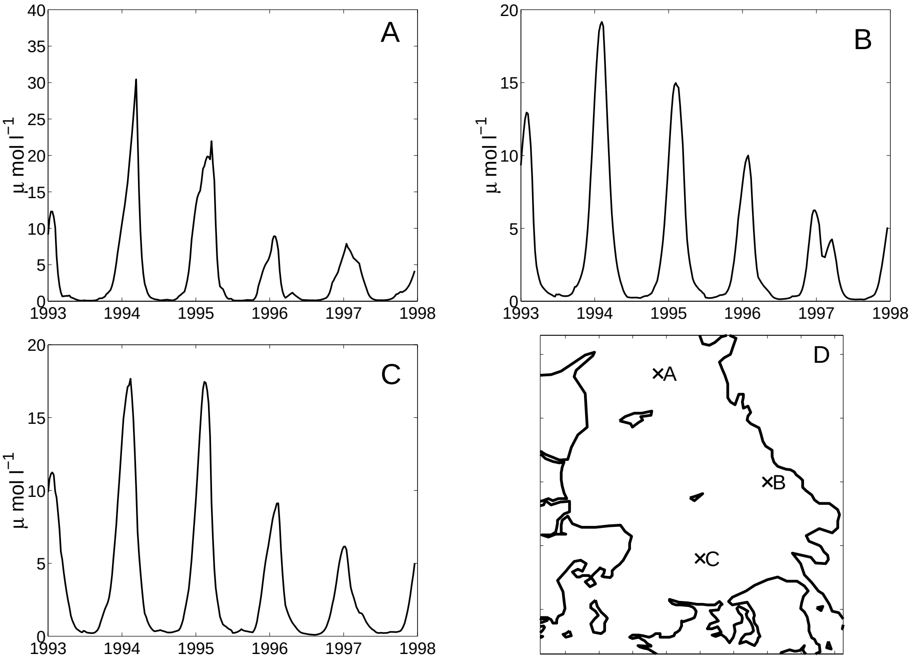

Figure D.4: *A,B,C) Modelled time series of DIN at three arbitrarily selected locations in the Kattegat. D) Map of the Kattegat with locations of the modelled time series.*

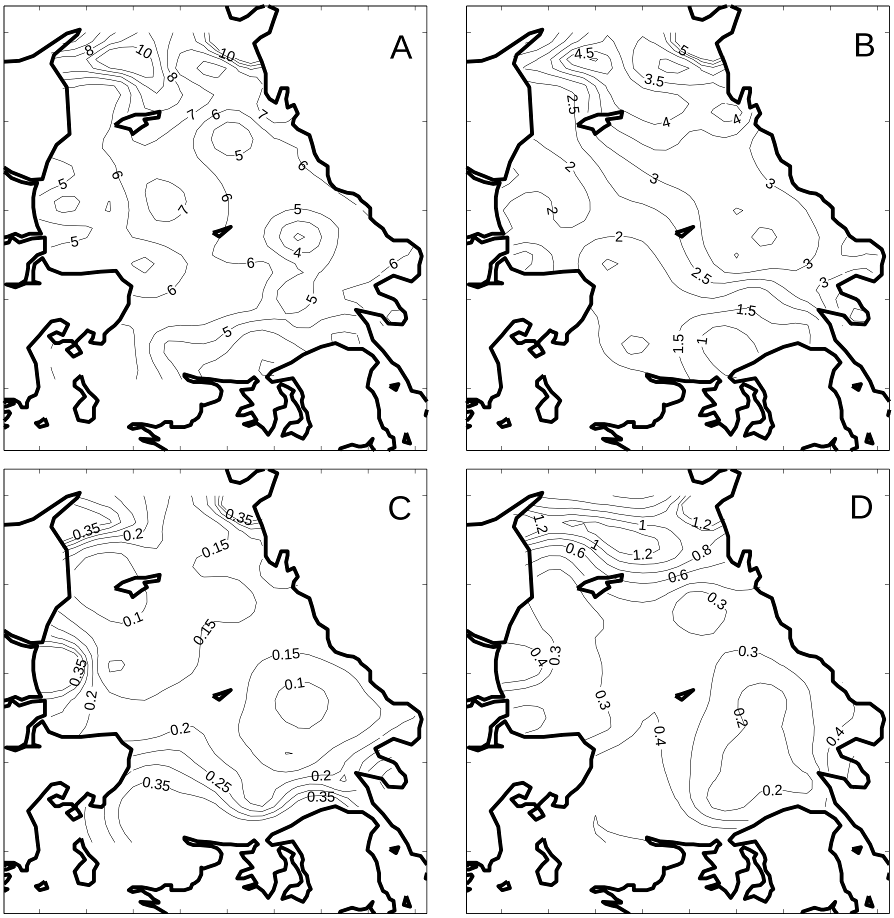

Figure D.5: *A) Spatial predictions of DIN for one week in A) January 1997, B) April 1997, C) July 1997, and D) September 1997.*

size of the station effects (Figure D.2A), which are measures of the overall DIN level at individual stations. It also fits well to the predicted DIN concentrations in the summer week (Figure D.5C), whereas the model does not predict high DIN concentrations in the south eastern part relative to other parts of the Kattegat in the three other weeks (Figure D.5A, D.5B and D.5D). In these three weeks relatively high concentrations were only predicted in the northern part of the Kattegat. This is caused by the fact that high DIN concentrations were not measured in the south eastern and south western part in these weeks as well as in weeks just before and after.

In relation to the application that this study deals with, i.e. DIN concentrations in the Kattegat, we have thought of the proposed model approach as a method for obtaining more information from the monitoring data. Subsequently, as mentioned in the introduction, model predictions could for example be used to reduce uncertainties of nutrient and carbon budgets in the Kattegat. We have aimed at proposing a simple approach which utilize the spatial and temporal correlation in data, can cope with a huge number of missing data values, and enable us to predict the DIN concentration in both space and time. We believe the proposed approach is general applicable for modelling phenomena observed over time at a number of spatially located monitoring stations, as the mean component of the model can be modified to fit the studied phenomenon. We also believe that this simple model could serve as a surrogate model for a more advanced and eventually computationally intensive model, and in that sense it could be used to correct the advanced model or to speed up computations in this.

D.7 Conclusion

This paper describes and applies an approach for modelling the DIN concentration in space and time. The main feature of the model approach is that it takes spatial and temporal correlations in data into account. The model is used to predict maps of the spatial distribution of DIN in four weeks, representing the four seasons, and to model time series of DIN at three different locations. The results can easily be interpreted from a biological and physical point of view, i.e. for this specific application the model performs very well. The proposed model approach is general, and can be applied to monitoring datasets observed over time at a number of locations.

Acknowledgements

Data for this study were supplied by the National Environmental Research Institute in Denmark and the Swedish Meteorological and Hydrological Institute. We acknowledge the effort by regional authorities in both countries for collecting the monitoring data.

References

Andersson, L. and Rydberg, L. (1988). Trends in nutrient and oxygen conditions within the Kattegat: Effects of local nutrient supply. *Estuarine, Coastal and Shelf Science*, **26**, 559–579.

Brown, P., Diggle, P., Lord, M., and Young, P. (2001). Space-time calibration of radar rainfall data. *Journal of the Royal Statistical Society, series C*, **50**, 221–241.

Brown, P., Roberts, G., Kåresen, K., and Tonellato, S. (2000). Blur-generated non-separable space-time models. *Journal of the Royal Statistical Society, series B*, **62**, 847–860.

Chilès, J. and Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty.* Wiley, New York.

Cressie, N. (1993). *Statistics for spatial data.* Wiley, New York.

Cressie, N. and Huang, H. (1999). Classes of nonseparable, spatiotemporal stationary covariance functions. *Journal of the American Statistical Association*, **94**, 1330–1340.

De Cesare, L., Myers, D., and Posa, D. (2001a). Estimating and modelling space-time correlation structures. *Statistics & Probability Letters*, **51**, 9–14.

De Cesare, L., Myers, D., and Posa, D. (2001b). Product-sum covariance for space-time modelling: an environmental application. *Environmetrics*, **12**, 11–23.

De Iaco, S., Myers, D., and Posa, D. (2001). Space-time analysis using a general product-sum model. *Statistics & Probability Letters*, **52**(1), 21–28.

De Iaco, S., Myers, D., and Posa, D. (2002). Nonseparable space-time covariance models: Some parametric families. *Mathematical Geology*, **34**(1), 23–42.

Figueira, R., Sousa, A., Pacheco, A., and Catarino, F. (2001). Use of secondary information in space-time statistics for biomonitoring studies of saline deposition. *Environmetrics*, **12**, 203–217.

Granéli, E. (1987). Nutrient limitation of phytoplankton biomass in a brackish water bay highly influenced by river discharge. *Estuarine, Coastal and Shelf Science*, **25**, 563–569.

Gustafsson, B. (2000). Time-dependent modeling of the Baltic entrance area. 1. quantification of circulation and residence times in the Kattegat and the straits of the Baltic sill. *Estuaries*, **23**, 231–252.

Kronvang, B., Ærtebjerg, G., Grant, R., Kristensen, P., Hovmand, M., and Kirkegaard, J. (1993). Nationwide monitoring of nutrients and their ecological effects: State of the Danish aquatic environment. *Ambio*, **22**, 176–187.

Kyriakidis, P. and Journel, A. (1999). Geostatistical space-time models: A review. *Mathematical Geology*, **31**, 651–684.

Stein, M. (1999). *Interpolation of spatial data: Some theory for kriging*. Springer, New York.

Wackernagel, H. (2003). *Multivariate geostatistics: An introduction with applications*. Springer, Berlin.

Appendix E

Computing spatial designs in R

The paper is submitted:

Lophaven, S., Carstensen, J., and Rootzén, H. (2004). Computing spatial designs in R. *Computers & Geosciences.*

Computing spatial designs in R

Søren Lophaven[1], Jacob Carstensen[2,3], and Helle Rootzén[1]

[1] Informatics and Mathematical Modelling, Technical University of Denmark
[2] Department of Marine Ecology, National Environmental Research Institute of Denmark
[3] Inland and Marine Waters Unit, Joint Research Centre, TP 280, I-21020 Ispra, Italy

Abstract

This paper describes parametric and non-parametric statistical methods for determining the optimal set of locations where samples are to be taken. This is a frequently occurring problem within environmental monitoring. The methods presented here are primarily based on geostatistics and focus on either estimation of model parameters or on spatial prediction. It is shown how these methods can be combined, and how these computations can be made in the statistical analysis and programming environment R.

KEY WORDS: *Bayesian geostatistics, classical geostatistics, semivariogram, space-filling design.*

E.1 Introduction

Today, monitoring networks have been established for reporting the state of the environment, and it is anticipated that the number of such networks will continue to increase in the future. These networks often aim at determining the spatial distribution of one or more pollutants. Ultimately, it is desirable to sample at all possible locations within a specific area of interest, but in practice the design of a monitoring network is limited by economic and operational constraints. In such cases the limited number of locations where samples are to be taken has to be determined. In principle two different design situations exists, i.e. a prospective design situation where the locations for a new set of sampling points have to be determined, and a retrospective design situation where an existing design is modified by adding sampling points to, or deleting sampling points from the design.

In this paper we describe statistical methods for computing the optimal set of sampling points, here referred to as the spatial design, and give some guidelines about when to apply the different methods. Furthermore, we show how the computations can be made in the statistical analysis and programming environment R (Ihaka and Gentleman, 1996; Grunsky, 2002). Some functions for computing spatial designs already exist, i.e. the ossfim-function in the gstat-package (Pebesma, 2004) and the cover.design-function (Royle and Nychka, 1998) in the fields-package. It is shown how these work, as are the four new R-functions (krige.conv.design, warrick.myers.design, zimmerman.homer.design and krige.bayes.design) that we have implemented. Hence, the spatial design problem is treated from a statistical point of view, without considering any practical aspects.

The design methods can be grouped according to whether they focus on estimation of the parameters in a geostatistical model (section E.5 and E.6) or on computing efficient spatial predictions using the estimated model (Section E.3 and E.4). Furthermore, we show that these two groups of methods result in different and conflicting designs, and describe how the methods could be combined (section E.7). Section E.8 discusses various aspects of the described methods. Most of the design methods described in this paper are based on the geostatistical theory, which is therefore briefly presented in section E.2.

E.2 Geostatistics

Geostatistics is the part of spatial statistics which is concerned with continuous spatial variation (Cressie, 1993). Given data y_i, $i = 1, ..., n$ at sampling locations x_i it is assumed that data can be modelled by

$$Y_i = S(x_i) + Z_i, \qquad i = 1, ..., n \tag{E.1}$$

where $S(x)$ is a stationary Gaussian process with expectation $\mathrm{E}[S(x)] = \mu$, variance $\mathrm{Var}[S(x)] = \sigma^2$ and correlation function $\rho(u) = \mathrm{Corr}[S(x_i), S(x_j)]$, $u = \| x_i - x_j \|$ being the spatial distance between x_i and x_j, and $Z_i \sim N(0, \tau^2)$, where τ^2 is the nugget effect. A possible model for describing spatial correlation is the powered exponential correlation function $\rho(u) = \exp[(-u/\phi)^\kappa]$, where $\phi > 0$ and $0 < \kappa \leq 2$ are parameters (Stein, 1999; Diggle et al., 2003). For $\kappa = 1$ and $\kappa = 2$ the powered exponential correlation function is usually called the exponential and the Gaussian correlation function, respectively. The model parameters can be estimated by maximum likelihood (Pardo-Iguzquiza, 1997), or based on the semivariogram, which is related to the correlation function in the following way

$$\gamma(u) = \tau^2 + \sigma^2(1 - \rho(u)) \tag{E.2}$$

The predictor that minimizes $\mathrm{E}[(\hat{S}(x) - S(x))^2]$ is called the kriging predictor. It can be shown that the kriging predictor for $T = S(x_0)$ is

$$\hat{T} = \mu + \sigma^2 r^T (\tau^2 I + \sigma^2 R)^{-1}(y - \mu I) \tag{E.3}$$

with prediction variance

$$\mathrm{Var}[\hat{T}] = \sigma^2 - \sigma^2 r^T (\tau^2 I + \sigma^2 R)^{-1} \sigma^2 r \tag{E.4}$$

where R is a symmetric $n \times n$ matrix with elements $\rho(\parallel x_i - x_j \parallel)$ and r is a $n \times 1$ vector with elements $\rho(\parallel x_0 - x_i \parallel)$, see e.g. Cressie (1993); Chilès and Delfiner (1999); Diggle et al. (2003). An important characteristic of the prediction variance (E.4) is that it does not depend on the data values y_i, which makes it attractive to use as a design criterion.

E.3 Spatial designs based on the prediction variance

McBratney et al. (1981) applied the classical kriging approach to the prospective design situation. They showed that if the model parameters in (E.1) are regarded as well-known then in the isotropic case the regular grid design is optimal for computing efficient spatial predictions. They also implemented their design approach in a Fortran-program (McBratney and Webster, 1981). This determines the necessary grid spacing, given the model parameters. In the anisotropic case the approach finds the necessary grid spacing in the direction of minimum spatial correlation, and the grid mesh is then elongated in the perpendicular direction in proportion to the anisotropy ratio. The Fortran program originally implemented by McBratney and Webster (1981) is now available in gstat (Pebesma, 2004). Assuming that gstat has been installed and loaded, then the R-commands

```
> x <- ossfim(seq(0.1,2,by=0.1),0.1, model = vgm(1,"Exp",0.3))
> plot(x$spacing,x$kriging.se,type="b",lwd=2,xlab="Grid spacing",
ylab="Block kriging standard error")
```

produce a plot of the block kriging standard error against the grid spacing of a regular grid when the exponential correlation function with $\phi = 0.3$, a constant mean μ, a variance $\sigma^2 = 1$, a nugget effect $\tau^2 = 0$, and an isotropic field are used (Figure E.1). We will use this model in the examples throughout the paper.

Various studies have computed retrospective spatial designs based on the prediction variance (E.4). As an example, Spruill and Candela (1990) showed how

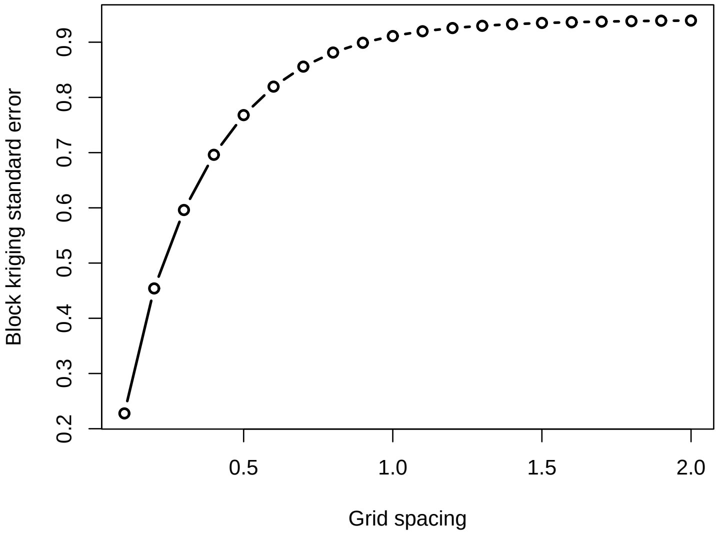

Figure E.1: *The block kriging standard error as a function of grid spacing in a regular design. The block size is 0.1.*

the number of sampling locations in a ground-water monitoring network measuring chloride-concentrations could be reduced from 120 to 99, with only a marginal increase in the prediction variance. Such a design can be computed with the R-function krige.conv.design which is based the function krige.conv in the geoR-package (Ribeiro et al., 2003). Assuming that geoR has been installed and loaded, then the R-commands

```
> simgrf <- grf(30,cov.model="exponential",cov.pars=c(1,0.3))
> x <- krige.conv.design(candidate.start=as.matrix(cbind(simgrf$coords,simgrf$data)),
grid=expand.grid(seq(0,1,l=17),seq(0,1,l=17)),fixed=NULL,cov.parameter=c(1,0.3),
kappa=1,nugget=0,n.add=15,mean.max=1,nruns=5,nn=9)
> plot(x$design[,1],x$design[,2],lwd=2,pch=19,cex=2,xlab="x1",ylab="x2",xlim=c(0,1),
ylim=c(0,1))
> points(simgrf$coords[,1],simgrf$coords[,2],lwd=2,pch=1,cex=2)
> text(0,1,"A",cex=2.5)
```

produce a plot (Figure E.2A) showing the simulated starting design (all points) of 30 points and the 15 points that remain in the design (•) when minimizing the average prediction variance is used as design criterion. It is seen that the 15 remaining sampling points fill up the area of interest, i.e. distances between neighboring points are large.

E.4 Space-filling design

As noted by Royle and Nychka (1998) a basic problem with the above described approach is that one must know the true geostatistical model, i.e. misspecification of the model produces designs that are not optimal. Instead, Royle and Nychka (1998) proposed a space-filling design criterion, which is based on geometry, i.e. it is only a function of the distance between sampling locations and a defined set of non-sampling locations (the candidate set), rather than on a stochastic model like (E.1). Based on results in Johnson et al. (1990) the authors showed that the resulting designs are nearly optimal from a spatial prediction point of view. The idea of a space-filling design is to find a combination of sampling points which fills up the available space in some suitably defined sense. A measure of how well a set of sampling points, $\mathfrak{D}$, covers a location x, can be written as

$$d_p(x, \mathfrak{D}) = \left(\sum_{u \in \mathfrak{D}} \parallel x - u \parallel^p \right)^{1/p} \tag{E.5}$$

where p can be any negative number. An overall geometric coverage criterion, i.e. the space-filling design criterion, can then be written as

$$C_{p,q}(\mathfrak{D}) = \left(\sum_{x \in C} d_p(x, \mathfrak{D})^q \right)^{1/q} \tag{E.6}$$

where $q > 0$, and the sum is taken over all x in the space of possible locations C. The goal is now to choose $\mathfrak{D}$ to minimize $C_{p,q}(\mathfrak{D})$.

Royle and Nychka (1998) also proposed a point-swapping algorithm, used to find the optimal design. This is the algorithm which is implemented in our four R-functions (krige.conv.design, warrick.myers.design, zimmerman.homer.design and krige.bayes.design). The basic idea of the algorithm is simple. For a given point in the current design, replace this point with members of the candidate set. If a particular swap reduces the design criterion, then this new point is included in the design and the old point is moved to the candidate set. This process is repeated for each member of the spatial design until there are no longer any productive swaps. Note, that when the design is modified, the design criterion will always be reduced, i.e. the algorithm will always converge to some solution depending on the randomly chosen starting design set. A space-filling design can be computed with the R-function cover.design in the fields-package. Assuming that this has been installed and loaded, then the R-commands

```
> x <- cover.design(R=sim$coords,nd=15,nruns=5,nn=9)
> plot(x$design[,1],x$design[,2],lwd=2,pch=19,cex=2,xlab="x1",ylab="x2",xlim=c(0,1),
ylim=c(0,1))
```

```
> points(simgrf$coords[,1],simgrf$coords[,2],lwd=2,pch=1,cex=2)
> text(0,1,"B",cex=2.5)
```

produce a plot (Figure E.2B) very similar to Figure E.2A, i.e. the remaining points fill up the area of interest. This property is of course what is indicated by the name ”space-filling design” and agrees with the fact that such designs are nearly optimal from a spatial prediction point of view.

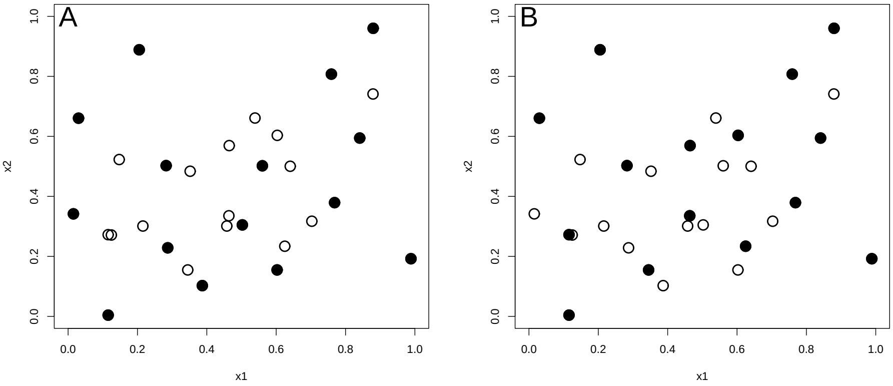

Figure E.2: *Reduction of an existing spatial design (all points) by deleting 15 points (open symbols). A) Design according to the classical prediction variance and B) Design according to the space-filling criterion. Filled circles (●) mark the remaining 15 sampling points.*

E.5 Spatial designs based on the sample semivariogram

One problem about spatial designs based on the prediction variance is that they assume that the model parameters are known, i.e. the uncertainty of the parameter estimates is not included in the design criterion. The space-filling design criterion leads to designs with points filling up the area of interest as well, and in that sense it also generates designs which are efficient from a prediction point of view, without taking proper care of the parameter uncertainties. The model parameters are rarely well-known, and therefore some studies have focused on how to design in order to estimate the semivariogram efficiently.

Some early studies of this was done by Russo (1984); Warrick and Myers (1987) who proposed criteria for finding the optimal set of sampling points in order to estimate the sample semivariogram, used in classical geostatistics for analysing the second moment structure of a spatial stochastic process. Warrick and Myers (1987) proposed minimising the design criterion

$$a \sum_{i=1}^{N} w_i (f_i - f_i^*)^2 + b \sum_{i=1}^{N} m_{1i} + c \sum_{i=1}^{N} m_{2i} \tag{E.7}$$

where f_i is the number of datapairs in the ith bin, f_i^* is the target value of f_i, N is the number of bins, m_{1i} and m_{2i} are respectively the variances of the distances and angles for the ith bin, and finally w_i, a, b and c are constants. These constants can be chosen arbitrarily, however, Warrick and Myers (1987) used the values $a = 4[N(N-1)]^{-2}$, $w_i = 1$, $b = c = 0$ and all f_i^* the same. The criterion in (E.7) results in designs for which the number of datapairs for each distance-angle bin is large, while the average of the distances and angles in each bin are close to the plotted distance and angle, respectively, and the variance of the distances and angles in each bin is small. The approach by Warrick and Myers (1987) is more general than the one by Russo (1984), because the latter only included the average and variance of the distances for each bin. The two criteria are the same if $a = 0$, $b = 1$, and $c = 0$. A spatial design produced according to (E.7) can be computed with the R-function warrick.myers.design, which does not depend on any existing package. The R-commands

```
> x <- warrick.myers.design(candidate.start=simgrf$coords,fixed=NULL,N=8,n.add=15,
a=1,b=1,c=0,n.directions=1,nruns=5,nn=9)
> plot(x$design[,1],x$design[,2],lwd=2,pch=19,cex=2,xlab="x1",ylab="x2",xlim=c(0,1),
ylim=c(0,1))
> points(simgrf$coords[,1],simgrf$coords[,2],lwd=2,pch=1,cex=2)
> text(0,1,"A",cex=2.5)
```

produce a plot (Figure E.3A) with a high proportion of points located close together, the points tend to be spatially clustered. Such a design is in contrast to space-filling designs or designs based on the prediction variance (Figure E.2).

E.6 Spatial designs based on the semivariogram model

Zimmerman and Homer (1991) extended the ideas by Warrick and Myers (1987) to consider not only estimation of the sample semivariogram but also parametric estimation of the semivariogram model. They proposed maximising the design criterion

$$\det(V^T W V) \tag{E.8}$$

where V is a matrix of partial derivatives of the chosen semivariogram model with respect to its parameters and W is a diagonal matrix, with non-zero elements equal to the reciprocal squared values of the assumed semivariogram. The rationale for this is given by Cressie (1985). Zimmerman and Homer (1991) also included the matrix D in their design criterion. This contains partial derivatives of specified parametric functions g_j with respect to the parameters, and the g_j identify the semivariogram attributes which are of particular interest. Thus, the criterion in (E.8) occurs when all parameters of the semivariogram are of interest. Müller and Zimmerman (1999) used a modification of (E.8) in which W is non-diagonal. A very similar approach was suggested by Bogaert and Russo (1999). For computational reasons we have implemented the criterion in (E.8) rather than those suggested by Müller and Zimmerman (1999); Bogaert and Russo (1999). Computations are performed with the R-function zimmerman.homer.design, which does not depend on any existing package. The R-commands

```
> x <- zimmerman.homer.design(candidate.start=simgrf$coords,fixed=NULL,
cov.parameter=c(1,0.3),kappa=1,nugget=0,n.add=15,nruns=5,nn=9)
> plot(x$design[,1],x$design[,2],lwd=2,pch=19,cex=2,xlab="x1",ylab="x2",xlim=c(0,1),
ylim=c(0,1))
> points(simgrf$coords[,1],simgrf$coords[,2],lwd=2,pch=1,cex=2)
> text(0,1,"B",cex=2.5)
```

produce a plot (Figure E.3B) which also has a high proportion of points located close together, even though the clustered pattern is less pronounced than in Figure E.3A.

E.7 Combining spatial designs

The above description shows two groups of spatial designs focusing on either spatial prediction, assuming that the model parameters are known, or on parameter estimation based on the semivariogram. When computation of spatial predictions is the primary goal of the spatial design, some points should also be allocated for estimating the model parameters, because good parameter estimates are required for computing efficient predictions (Stein, 1999). Thus, the design should consist of some points allocated for predicting the spatial distribution and some for estimating the model parameters (Müller, 2001; Martin, 2001).

Müller (2001) showed that for $n_0 \simeq 0.3n$ the number of small bins of the sample semivariogram equals the number of large bins. Here n is the total number of

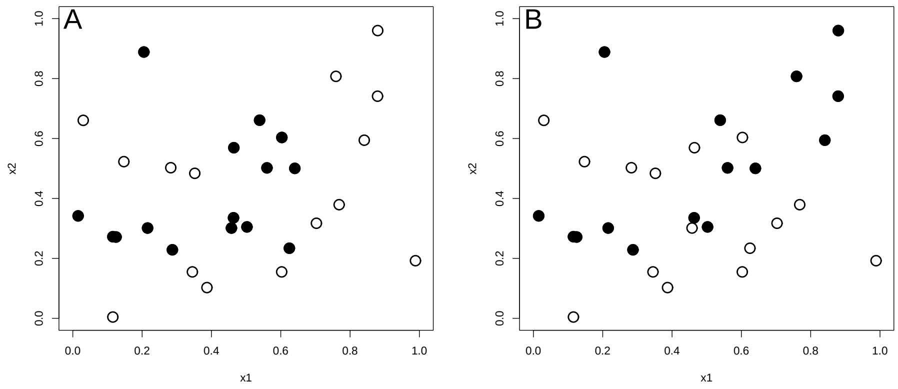

Figure E.3: *Reduction of an existing spatial design (all points) by deleting 15 points (open symbols). A) Design according to the criterion proposed by Warrick and Myers (1987) and B) Design according to the criterion proposed by Zimmerman and Homer (1991). Filled circles (●) mark the remaining 15 sampling points.*

sampling points in the design, and $n_0 < n$ is the number of sampling points in the design used to compute spatial predictions. There are more small bins than large bins when $n_0 < 0.3n$ and vice versa when $n_0 > 0.3n$. To allow for a larger proportion of small bins, Müller (2001) recommended to have $n_0 < 0.3n$, e.g. $n_0 = 0.2n$. Such a combined spatial design is computed by combining one of the R-functions krige.conv.design or cover.design with warrick.myers.design or zimmerman.homer.design. In the example below we combine the functions cover.design and warrick.myers.design to the prospective situation where a spatial design of 30 points is to be constructed. The 30 points are chosen from a candidate set consisting of 200 randomly chosen points. The R-commands

```
> candidate <- cbind(runif(200),runif(200))
> x1 <- cover.design(R=candidate,nd=6,nruns=5,nn=30)
> x2 <- warrick.myers.design(candidate.start=candidate,fixed=x1$best.id,N=10,
n.add=24,a=1,b=1,c=0,n.directions=1,nruns=5,nn=30)
> plot(candidate[,1],candidate[,2],lwd=2,pch="+",xlab="x1",ylab="x2",xlim=c(0,1),
ylim=c(0,1))
> points(x1$design[,1],x1$design[,2],lwd=2,pch=15,cex=2)
> points(x2$design[1:24,1],x2$design[1:24,2],lwd=2,pch=19,cex=2)
```

first allocates six points for spatial prediction according to the space-filling design criterion, and afterwards 24 points, with the first six points fixed, according to the criterion proposed by Warrick and Myers (1987). This means that the first six points are not swapped, but included in the computation of the design criterion. The resulting design (Figure E.4) serves as a compromise between the conflicting issues of focusing on spatial prediction or parameter estimation. However, Figure E.4 also shows that a large proportion of the points are close together in clusters, indicating that we might choose $n_0 = \alpha n$ with $\alpha > 0.2$.

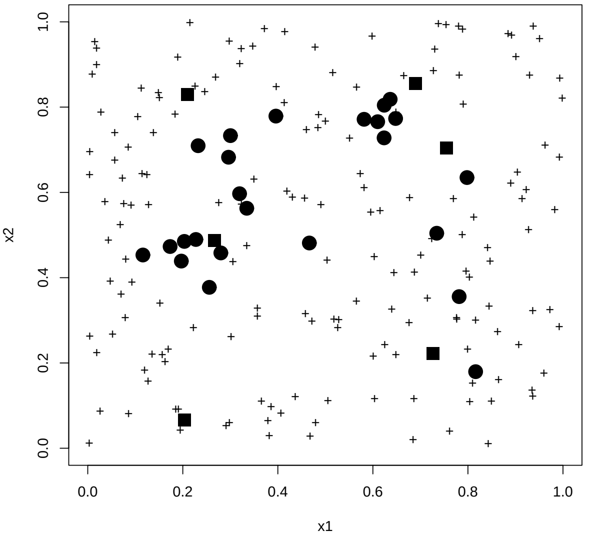

Figure E.4: *A prospective design situation in which a new spatial design is to be constructed. Based on a combination of the space-filling criterion and the design criterion proposed by Warrick and Myers (1987), 30 sampling points (shown by ■ and •) were selected from a candidate set (all points) of 200 points. Filled squares (■) mark the six points selected by means of the space-filling criterion. Filled circles (•) mark the 24 points selected by means of the Warrick and Myers (1987) - criterion.*

Another way to combine spatial designs, which overcomes the problem of choosing α, is to compute spatial predictions and associated variances using a Bayesian approach (Diggle et al., 2003; Handcock and Stein, 1993; Le and Zidek, 1992). In the classical geostatistical approach parameters are estimated, and afterwards these are plugged into (E.3) and (E.4) to compute predictions and associated prediction variances as if the parameter estimates were the truth. Predictions computed using the Bayesian approach can be interpreted as a weighted average of a number of classical predictions, with weights given by the posterior distributions of the parameters. This means, that in the Bayesian approach the uncertainty of the model parameters, described by the posterior distribution, is automatically incorporated in the predictive distribution. Hence, good parameter estimates, i.e. a narrow posterior distribution, leads to more efficient spatial predictions. The Bayesian approach is computationally intensive, but can be used when sampling points are to be deleted from an existing design. When adding points to a design the computational work usually increases dramatically, and this situation is not considered here. The R-function krige.bayes.design is based on the krige.bayes function in geoR, and uses the average or the maximum of the variances of the predictive distributions as a design criterion. Computational details can be found in Ribeiro (1999). Again we investigate the retrospective situation where we want to delete 15 points from the starting design of 30 points (Figures E.2 and E.3). The R-commands

```
> x <- krige.bayes.design(candidate.start=as.matrix(cbind(simgrf$coords,simgrf$data)),
grid=expand.grid(seq(0,1,by=0.1),seq(0,1,by=0.1)),fixed=NULL,n.add=15,mean.max=1,
nruns=1,nn=9)
> plot(x$design[,1],x$design[,2],lwd=2,pch=19,cex=2,xlab="x1",ylab="x2",xlim=c(0,1),
ylim=c(0,1))
> points(simgrf$coords[,1],simgrf$coords[,2],lwd=2,pch=1,cex=2)
```

produce a plot (Figure E.5) of a design with most sampling points allocated to fill up the area of interest, but also of a few points located close together, e.g. the two points approximately at $(x_1, x_2) = (0.15, 0.25)$, allocated to account for the uncertainty in the model parameters.

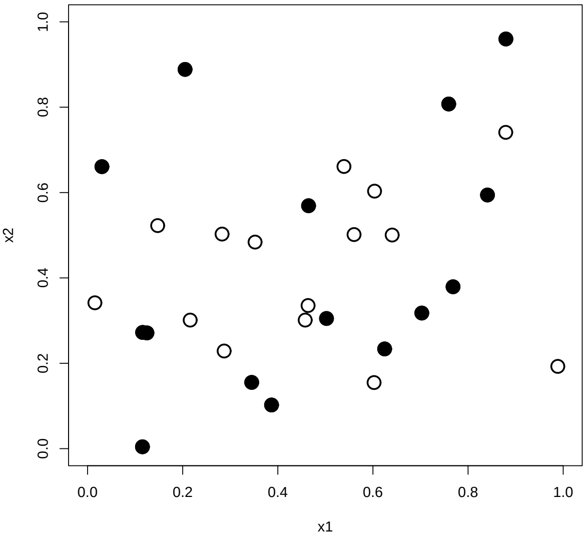

Figure E.5: *Reduction of an existing spatial design (all points) by deleting 15 points (open symbols). Filled circles (•) mark the remaining 15 sampling points. The design is computed by means of the Bayesian design criterion.*

E.8 Discussion and conclusions

The design methods described in sections E.3 - E.6 can be grouped according to whether they focus on estimation of the parameters in a geostatistical model or on computing efficient spatial predictions using the estimated model. They could also be grouped according to whether they are non-parametric (Sections E.4 and E.5) or model-based (Sections E.3 and E.6). Obviously, to apply the latter group of methods prior knowledge about the model parameters are required, while the non-parametric methods only depend on the distances between sampling points in the design.

In a prospective design situation it would be reasonable to apply a method which focus on parameter estimation, and then after some time change the design according to a criterion which focus on spatial prediction. Alternatively, the spatial design could be constructed by means of a combination of two criteria, as described in section E.7. The retrospective design situation where points are to be deleted from an existing design can be handled by the Bayesian design criterion. Although this is computationally intensive it gives designs which ensures that spatial predictions can be computed efficiently, while taking parameter uncertainty into account. In the situation where points are to be added to an existing design, we believe that in most situations it is reasonable to add points by means of a criterion focusing on spatial prediction. The R-functions described in this paper can be used in these situations, and the resulting designs should be compared with expert judgement to make decisions about where to sample, or about which points to delete.

Some assumptions were made in the implementation of the methods, e.g. in the functions zimmerman.homer.design and krige.conv.design isotropy, a constant mean, and a powered exponential correlation function are assumed, while krige.bayes.design also assumes isotropy and a constant mean, as well as a flat prior for ϕ, an improper prior for σ^2 and a nugget effect τ^2 which is fixed in value. Most of these settings can easily be changed in the functions. The effect of our choice of priors is an area which needs further investigation. We expect that for small designs the priors will affect the design substantially, and that using a flat prior for τ^2 rather than keeping it fixed would result in designs with more points close together in clusters.

The point-swapping algorithm was analysed by Royle and Nychka (1998) when applied to the space-filling criterion. They found that the resulting designs are sufficiently close to being optimal, and that the nearest-neighbor search strategy is reasonable, especially when applied to large designs and candidate sets.

Thus, we believe that the described methods cover the different spatial design situations that can arise, and by implementation in the statistical analysis and programming environment R they are easy to use in combination with other statistical analysis.

Appendix

The function krige.conv.design

x $<-$ krige.conv.design(candidate.start,grid,fixed=NULL,cov.parameter,kappa,nugget,n.add, mean.max=1,nruns,nn)

candidate.start: Matrix of candidate points (first two columns) and data values (third column).

grid: Two-column matrix of points in which the predictive distribution is computed.

fixed: Vector specifying points to be forced into the spatial design.

cov.parameter: Two-element vector with values of the covariance parameters (σ^2 and ϕ)

kappa: parameter in the powered exponential correlation function

nugget: The nugget effect τ^2

n.add: Number of points to add to the design.

mean.max: If 1 minimising the average prediction variance is used as design criterion, otherwise the maximum prediction variance is used.

nruns: The number of random starting designs to be optimized.

nn: Number of nearest-neighbors to search over.

The function can be used retrospectively as well as prospectively. Isotropy, a constant mean, and a powered exponential correlation function are assumed. These settings can easily be changed by changing the call to krige.conv.

The function warrick.myers.design

x $<-$ warrick.myers.design(candidate.start,fixed=NULL,N,n.add,a,b,c,n.directions,nruns,nn)

candidate.start: Two-column matrix of candidate points (first two columns).

fixed: Vector specifying points to be forced into the spatial design.

N: Number of bins for the semivariogram

n.add: Number of points to add to the design.

a,b,c: Constants corresponding to Warrick and Myers (1987).

n.directions: Number of directions to consider.

nruns: The number of random starting designs to be optimized.

nn: Number of nearest-neighbors to search over.

The function can be used retrospectively as well as prospectively. A constant mean is assumed and all weights w_i in Warrick and Myers (1987) are =1. Both isotropy and anisotropy can be handled.

The function zimmerman.homer.design

x $< -$ zimmerman.homer.design(candidate.start,fixed=NULL,cov.parameter,kappa,nugget, n.add,nruns,nn)

candidate.start: Matrix of candidate points (first two columns) and data values (third column).

fixed: Vector specifying points to be forced into the spatial design.

cov.parameter: Two-element vector with values of the covariance parameters (σ^2 and ϕ).

kappa: parameter in the powered exponential correlation function

nugget: The nugget effect τ^2.

n.add: Number of points to add to the design.

nruns: The number of random starting designs to be optimized.

nn: Number of nearest-neighbors to search over.

The function can be used retrospectively as well as prospectively. Isotropy, a constant mean, and a powered exponential correlation function are assumed. The correlation can of course be changed, but requires that the partial derivatives of another model are specified.

The function krige.bayes.design

$x <- $ krige.bayes.design(candidate.start,grid,fixed=NULL,n.add,mean.max=1,nruns,nn)

candidate.start: Matrix of candidate points (first two columns) and data values (third column).

grid: Two-column matrix of points in which the predictive distribution is computed.

fixed: Vector specifying points to be forced into the spatial design.

n.add: Number of points to add to the design.

mean.max: If 1 minimising the average prediction variance is used as design criterion, otherwise the maximum prediction variance is used.

nruns: The number of random starting designs to be optimized.

nn: Number of nearest-neighbors to search over.

The function delete points from an existing design by means of the variance of the predictive distribution. Isotropy and a constant mean are assumed. Prior are chosen as the default for krige.bayes (see documentation). These settings can easily be changed by changing the call to krige.bayes.

References

Bogaert, P. and Russo, D. (1999). Optimal spatial sampling design for the estimation of the variogram based on a least squares approach. *Water Resources Research*, **35**(4), 1275–1289.

Chilès, J. and Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty.* Wiley, New York.

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

Cressie, N. (1993). *Statistics for spatial data.* Wiley, New York.

Diggle, P., Ribeiro, P., and Christensen, O. (2003). *Spatial statistics and computational methods*, chapter An introduction to model-based geostatistics, pp. 43–86. Lecture notes in statistics. Springer, New York.

Grunsky, E. (2002). R: A data analysis and statistical environment - an emerging tool for the geosciences. *Computers & Geosciences*, **28**(10), 1219–1222.

Handcock, M. and Stein, M. (1993). A bayesian analysis of kriging. *Technometrics*, **35**, 403–410.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**(3), 299–314.

Johnson, M., Moore, L., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, **26**, 131–148.

Le, N. and Zidek, J. (1992). Interpolation with uncertain covariances: A bayesian alternative to kriging. *Journal of Multivariate Analysis*, **43**, 351–374.

Martin, R. (2001). Comparing and contrasting some environmental and experimental design problems. *Environmetrics*, **12**, 303–317.

McBratney, A. and Webster, R. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables-II. Program and examples. *Computers & Geosciences*, **7**(4), 335–365.

McBratney, A., Webster, R., and Burgess, T. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables-I. Theory and methods. *Computers & Geosciences*, **7**(4), 331–334.

Müller, W. and Zimmerman, D. (1999). Optimal designs for variogram estimation. *Environmetrics*, **10**, 23–37.

Müller, W. G. (2001). *Collecting spatial data: Optimum design of experiments for random fields.* Contributions to statistics. Physica-Verlag, Heidelberg (Germany).

Pardo-Iguzquiza, E. (1997). MLREML: A computer program for the inference of spatial covariance parameters by maximum likelihood and restricted maximum likelihood. *Computers & Geosciences*, **23**(2), 153–162.

Pebesma, E. (2004). Multivariable geostatistics in S: The gstat package. *Computers & Geosciences*, **30**, 683–691.

Ribeiro, P. (1999). Bayesian inference in Gaussian model-based geostatistics. Technical Report St-99-08, Lancaster University.

Ribeiro, P., Christensen, O., and Diggle, P. (2003). Geor and georglm: Software for model-based geostatistics. In Hornik, K., Leisch, F., and Zeileis, A., editors, *Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003).*

Royle, J. and Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in Splus. *Computers & Geosciences*, **24**(5), 479–488.

Russo, D. (1984). Design of an optimal sampling network for estimating the variogram. *Soil Science Society of American Journal*, **48**, 708–716.

Spruill, T. and Candela, L. (1990). Two approaches to design of monitoring networks. *Ground Water*, **28**(3), 430–442.

Stein, M. (1999). *Interpolation of spatial data: Some theory for kriging.* Springer, New York.

Warrick, A. and Myers, D. (1987). Optimization of sampling locations for variogram calculations. *Water Resources Research*, **23**, 496–500.

Zimmerman, D. and Homer, K. (1991). A network design criterion for estimating selected attributes of the semivariogram. *Environmetrics*, **2**(4), 425–441.

Appendix F

A Bayesian geostatistical approach to optimal design of monitoring networks

The paper is submitted:

Lophaven, S., Carstensen, J., and Rootzén, H. (2004). A Bayesian geostatistical approach to optimal design of monitoring networks. *Environmental and Ecological Statistics*.

A Bayesian geostatistical approach to optimal design of monitoring networks

Søren Lophaven[1], Jacob Carstensen[2], and Helle Rootzén[1]

[1] Informatics and Mathematical Modelling, Technical University of Denmark
[2] Department of Marine Ecology, National Environmental Research Institute of Denmark

Abstract

This paper applies a Bayesian geostatistical approach for reducing the number of sampling stations within a marine monitoring program. The approach focuses on designing a monitoring program which is efficient from a prediction point of view, while taking parameter uncertainty for the spatial correlation into account. The novelty is that it combines different classical geostatistical design methods. The number of sampling stations can be reduced from the present 31 to 14 with only a marginal increase in the design criterion. The reduced network consist of sampling stations covering the Kattegat area as well as some stations close to each other. The approach can be applied for revising the monitoring effort in different types of networks provided that spatial correlation prevails.

KEY WORDS: *Bayesian geostatistics, the Kattegat, spatial design*

F.1 Introduction

Environmental pollution has become a major problem on all possible scales from local, regional to global. Today, monitoring networks have been established to obtain information on the spatial distribution of different pollutants as well as their biological effects more recently. Environmental monitoring of air, soil and water pollution is a challenge to the modern society having the aim of determining the spatial distribution on a wide range of substances. Ultimately, it is desirable to sample at all possible locations within a specific area of interest, but in practice the design of a monitoring network is limited by economic and operational constraints. In such cases the limited number of locations where samples are to be taken has to be determined.

Existing monitoring networks have mainly been established on heuristic rules such as appointing a sampling location to be representative for a larger area, rather than defining optimality criteria and determine the location of sampling stations on this basis. One major problem associated with optimizing the monitoring network design is that it requires a priori information on the spatial scales of variation, which may not be readily available in the case of a new monitoring network design. Another problem is that a single criterion encompassing all the different pollutants and biological effects in the monitoring network cannot be established, due to contrasting definitions of optimality. Moreover, the general objectives of a monitoring network is often formulated in rather broad terms making the translation into a more stringent mathematical optimality formulation quite difficult. Statistical methods for design of monitoring networks should not be seen as an alternative to the traditional heuristic rules based on expert judgement, but more as a complementary tool that give modifications or support to the design decisions made.

Most monitoring networks undergo revisions at different times during their life span. Here we shall consider such a situation when the monitoring network has already been established and the necessary a priori information can be obtained from historical data. In section F.2 geostatistics and geostatistical design methods are briefly reviewed. Based on this review an appropriate approach for reducing the number of sampling stations within the Danish National Aquatic Monitoring and Assessment Programme (DNAMAP) in the Kattegat, a coastal marginal sea. Nutrients, oxygen, chlorophyll a, temperature and salinity in addition to a wider range of biological measurements were sampled at the monitoring stations. The application of the approach is described in section F.3.

F.2 Methodology

F.2.1 Geostatistics

Geostatistics is the part of spatial statistics which is concerned with continuous spatial variation (Cressie, 1993). Given data y_i, $i = 1, ..., n$ at spatial locations x_i it is assumed that data can be modelled by

$$Y_i = S(x_i) + Z_i, \qquad i = 1, ..., n \tag{F.1}$$

where $S(x)$ is a stationary Gaussian process with $\mathrm{E}[S(x)] = \mu = F\beta$, where F is a $n \times p$ matrix of covariates, and β is the parameter vector, $\mathrm{Var}[S(x)] = \sigma^2$ and correlation function $\rho(u) = \mathrm{Corr}[S(x_i), S(x_j)]$, $u =\parallel x_i - x_j \parallel$ being the spatial distance between x_i and x_j, and $Z_i \sim N(0, \tau^2)$. A possible model for describing spatial correlation is the exponential correlation function $\rho(u) = \exp(-u/\phi)$, where ϕ is the range parameter. Other widely used models are the spherical, the Gaussian and the Matérn model, the latter of which the exponential and the Gaussian correlation functions are special cases (Diggle et al., 2003). The predictor that minimizes $\mathrm{E}[(\hat{S}(x) - S(x))^2]$ is called the kriging predictor. It can be shown that the kriging predictor for $T = S(x_0)$ is

$$\hat{T} = \mu + \sigma^2 r^T (\tau^2 I + \sigma^2 R)^{-1} (y - \mu I) \tag{F.2}$$

with prediction variance

$$\mathrm{Var}[T|y] = \sigma^2 - \sigma^2 r^T (\tau^2 I + \sigma^2 R)^{-1} \sigma^2 r \tag{F.3}$$

where R is a symmetric $n \times n$ matrix with elements $\rho(\parallel x_i - x_j \parallel)$ and r is a $n \times 1$ vector with elements $\rho(\parallel x_0 - x_i \parallel)$, see e.g. Cressie (1993); Chilès and Delfiner (1999); Diggle et al. (2003). An important characteristic of the prediction variance (F.3) is that at any given location the prediction depends on the distances to the sampling points in the design and the model parameters, which for the exponential correlation function the parameter vector is given by $\theta = (\beta, \sigma^2, \phi, \tau^2)$. The prediction variance does not depend on the data values, which makes it attractive to use as a criterion for constructing or modifying sampling designs.

F.2.2 Geostatistical design methods - Review

A reasonable design criterion could be to minimize the average or the maximum prediction variance over all prediction points. Given the values of the model parameters and a maximum tolerable value of one of these design criteria the necessary grid spacing of regular, rectangular or triangular designs can be calculated (McBratney et al., 1981). Geostatistical design based on the prediction variance (F.3) are widely used in the literature, see e.g. Winkels and Stein (1997); Spruill and Candela (1990); Ben-Jemaa et al. (1995) who studied optimal design of environmental monitoring networks. Royle and Nychka (1998) suggested a design criterion based on geometry, i.e. it is only a function of the distance between sampling and non-sampling (candidate) points, rather than on a stochastic model like (F.1). Based on results in Johnson et al. (1990) the authors showed that the resulting designs are nearly optimal from a spatial prediction point of view. Sampling designs based on the prediction variance have sampling points with almost equal distances between the neighboring locations, and are optimal from a prediction point of view if the model parameters are known. This is usually not the case, particularly not if a monitoring network is to be designed. Thus, these designs are not well-suited for parameter estimation, especially not for estimating the nugget effect τ^2, which is critical to efficient spatial predictions (Stein, 1999).

Rather than focusing on designs for minimizing the average or the maximum prediction variance over all prediction points, various studies concentrate on how to construct sampling designs from which the model parameters can be optimally estimated. Early studies proposed algorithms for finding the optimal combination of sampling points in order to estimate the sample semivariogram, used in geostatistics for analyzing the second moment structure of a spatial stochastic process (Russo, 1984; Russo and Jury, 1988; Warrick and Myers, 1987). Zimmerman and Homer (1991); Müller and Zimmerman (1999); Bogaert and Russo (1999) extended these ideas to consider not only estimation of the sample semivariogram but also parametric estimation of the semivariogram model.

The above description shows two groups of design criteria focusing either on spatial prediction or parameter estimation. Lark (2002) illustrated that errors in the model parameters result in errors in the estimated prediction variances. This means that if the primary goal of the monitoring network is to compute spatial predictions a combination of the two groups of design criteria should be used. This ensures that efficient spatial predictions can be computed while taking proper care of the uncertainties of the model parameters. Hence in this case an efficient design should consist of some sampling points allocated for estimating model parameters and some for computing spatial predictions (Müller, 2001; Martin, 2001).

A way to combine designs for computing spatial predictions with designs for estimating model parameters is to compute spatial predictions and associated prediction variances using a Bayesian approach, see e.g. Diggle et al. (2003); Handcock and Stein (1993); Le and Zidek (1992). In the Bayesian approach the uncertainty of the model parameters, described by the posterior distribution, is automatically incorporated in the predictions. This means that good parameter estimates, i.e. a narrow posterior distribution, leads to more efficient spatial predictions. In the next section a Bayesian geostatistical approach is applied to analyze how the DNAMAP in the Kattegat should be reduced, i.e. which monitoring stations should be removed from the existing monitoring network. The Bayesian approach is chosen, because the monitoring program has already existed for several decades, and information on the geostatistical model can be obtained from the existing monitoring data. This information should be utilized in a model based design approach, rather than applying a non-parametric approach, e.g. as suggested by Warrick and Myers (1987); Royle and Nychka

(1998). Furthermore, we want to focus on accurate prediction, but at the same time take parameter uncertainty into account.

F.2.3 Geostatistical design methods - A Bayesian approach

Consider the stochastic variable Y in (F.1) with probability distribution $pr(y|\theta)$, $\theta = (\beta, \sigma^2, \phi, \tau^2)'$ being the unknown parameter vector. The distribution of Y is given by

$$(Y|\beta, \sigma^2, \phi, \tau^2) \sim N(F\beta, \tau^2 I + \sigma^2 R)) \qquad \text{(F.4)}$$

The likelihood is a function of the parameter vector θ and is given by

$$L(\theta|y) \propto |\tau^2 I + \sigma^2 R|^{\frac{1}{2}} \exp\left(\frac{1}{2}(y - F\beta)'(\tau^2 I + \sigma^2 R)^{-1}(y - F\beta)\right) \qquad \text{(F.5)}$$

In the Bayesian approach Y and θ are considered random quantities. Given data y and prior distributions $pr(\theta) = pr(\beta, \sigma^2, \phi, \tau^2)$ of the parameters the posterior distributions are found using the relation

$$\begin{aligned} pr(\beta, \sigma^2, \phi, \tau^2|y) \quad \propto \quad & pr(\beta, \sigma^2, \phi, \tau^2)|\tau^2 I + \sigma^2 R|^{\frac{1}{2}} \\ & \exp\left(\frac{1}{2}(y - F\beta)'(\tau^2 I + \sigma^2 R)^{-1}(y - F\beta)\right) \end{aligned} \qquad \text{(F.6)}$$

and the predictive distribution $pr(y_0|y)$ by

$$pr(y_0|y) = \int pr(y_0|y, \theta)pr(\theta|y)\mathrm{d}\theta \qquad \text{(F.7)}$$

Classical geostatistical methods estimate the parameters, and then use these to perform predictions as if the estimates were the truth. Predictions computed using the Bayesian approach can be interpreted as a weighted average of a number of classical predictions, with weights given by the posterior distributions of the parameters.

In order to determine the specific monitoring stations to be removed from the network, stations were removed one at a time by the following algorithm. The jth station was removed and values of a Gaussian random field were simulated in the remaining sampling points using the exponential model with parameters estimated from data. This was repeated 1000 times, and for each simulated set, predictions $E(y_0|y_{(j)})$ and prediction variances $V(y_0|y_{(j)})$ were computed in a regular grid of 103 points covering the entire Kattegat area (Figure F.1). For each of the 103 prediction locations the averaged prediction variance ν_{ij} for the ith prediction location, and corresponding to removing the jth station was computed by

$$\nu_{ij} = \frac{1}{1000}\sum_{k=1}^{1000} V(y_0|y_{(j)}), \qquad i = 1, \cdots, 103 \tag{F.8}$$

The maximum of the 103 averaged prediction variances was used as a design criterion C_i, i.e.

$$C_j = \max_i(\nu_{ij}) \tag{F.9}$$

Finally, the monitoring station for which its removal resulted in the smallest design criterion, $\min_j(C_j)$, is permanently removed from the network. The whole process can then be repeated for a number of times, depending on how many stations we want to remove. It is seen that the design criterion is based on minimizing the maximum prediction variance. This approach was chosen in order to ensure efficient predictions everywhere in the Kattegat, rather than on efficient predictions on average.

F.3 Application: Modifying the Danish National Aquatic Monitoring and Assessment Programme in Kattegat

F.3.1 Study area

The Kattegat (Figure F.1A) is a relatively shallow coastal ecosystem significantly affected by man-made eutrophication over the last 4-5 decades (Richardson, 1996; Carstensen, 2003). It comprises the transition zone between the Baltic Sea and the North Sea with an area of 22,290 km^2, a volume of 533 km^3 and a mean depth of 24 meters (Gustafsson, 2000). Monitoring is carried out as a joint effort between Danish and Swedish authorities and includes routine measurements of nutrients, oxygen, chlorophyll a, temperature and salinity in addition to a wider range of biological measurements. The first monitoring program was established in the beginning of the 1970s and in the following monitoring programs the number of stations gradually increased. In recent years the number of sampling stations has been reduced due to reallocation of resources within the monitoring program.

F.3.2 Data description and preparation

In this study we will consider 31 stations (Figure F.1B) and focus on the concentration of dissolved inorganic nitrogen (DIN) in the surface layer monitored with various frequencies during the 5-year period from 1993 to 1997. Nitrogen is assumed to be the nutrient limiting phytoplankton growth in the Kattegat during the productive season (March - October). The concentration of DIN in the surface layer has a seasonal variation with low concentrations during summertime and high concentrations during winter, when primary production and DIN uptake by phytoplankton are low.

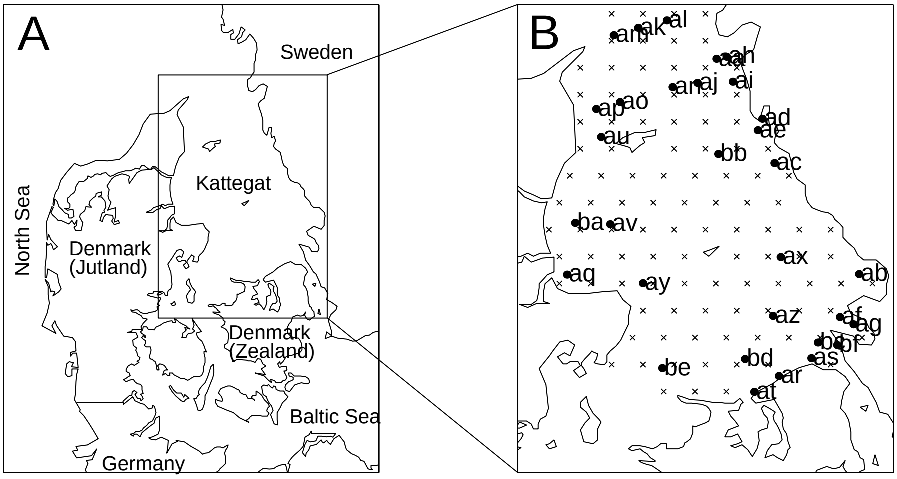

Figure F.1: *The study area. A) Kattegat is a transitional sea between the North Sea and the Baltic Sea. B) Locations of existing monitoring stations (●), denoted by a name of two letters, and prediction locations (×) in the Kattegat.*

Due to the skewed distribution of the DIN values (Figure F.2), data were log-transformed prior to making geostatistical analysis. Afterwards the seasonal variation was removed using a harmonic function for the mean value

$$\mu_{ijt} = \alpha_i + \zeta_j + \delta_1 \sin\left(\frac{2\pi t}{52} + \varphi_1\right) + \delta_2 \sin\left(\frac{2\pi t}{26} + \varphi_2\right) \tag{F.10}$$

where α_i $(i = 1, \cdots, 31)$ are the monitoring station effects, and ζ_j $(j = 1, \cdots, 5)$ the year effects where the years were defined as starting and ending the first of July. This definition of the year effect was chosen the reduce the discontinuity between separate years, since DIN concentrations are generally low during summer. Furthermore, δ_1, δ_2, φ_1 and φ_2 are the yearly amplitude, the half-year amplitude, the yearly phase shift, and the half-year phase shift, respectively, while $t = 1, \cdots, 52$ is the week number. After removing the seasonal variation we analyzed the residuals assuming that (F.10) described the true mean field μ in (F.1).

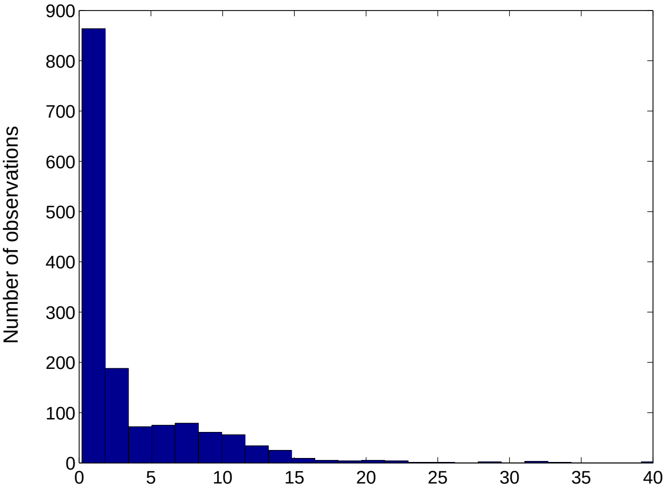

Figure F.2: *Histogram for DIN.*

The monitoring data were unevenly distributed in time and space (Figure F.3), and for many weeks the total number of observations was less than 10, and around 20 for a few weeks only. consequently, this would complicate the geostatistical analyses if these were to be carried out on a weekly basis.

Rather than focusing on individual weeks the sample semivariogram was computed by considering all possible datapairs for each of the weeks in the 5-year period. For all datapairs the squared differences were computed and separated into 14 bins each with a width of 10 km. For each bin the average divided by 2 is computed to produce the spatial lag 0 sample semivariogram, i.e. the spatial semivariogram for no time lag. This can be interpreted as the average spatial sample semivariogram, averaged over all weeks in the period 1993-1997 (Figure F.4). Analysis of directional lag 0 semivariograms showed no anisotropy. Estimates of the parameters of the exponential semivariogram model, found using ordinary least squares regression, were (σ^2, ϕ, τ^2)=(1.08,26.70,0.41). This model was the basis for our Bayesian design approach. Based on the sample semivariogram a uniform prior for ϕ on $(2, 52)$ was chosen, for $(\beta, \sigma^2|\phi)$ a prior proportional to $1/\sigma^2$ was used, whereas the nugget effect $\tau^2 = 0.41$ was assumed known in order to reduce the time of computation.

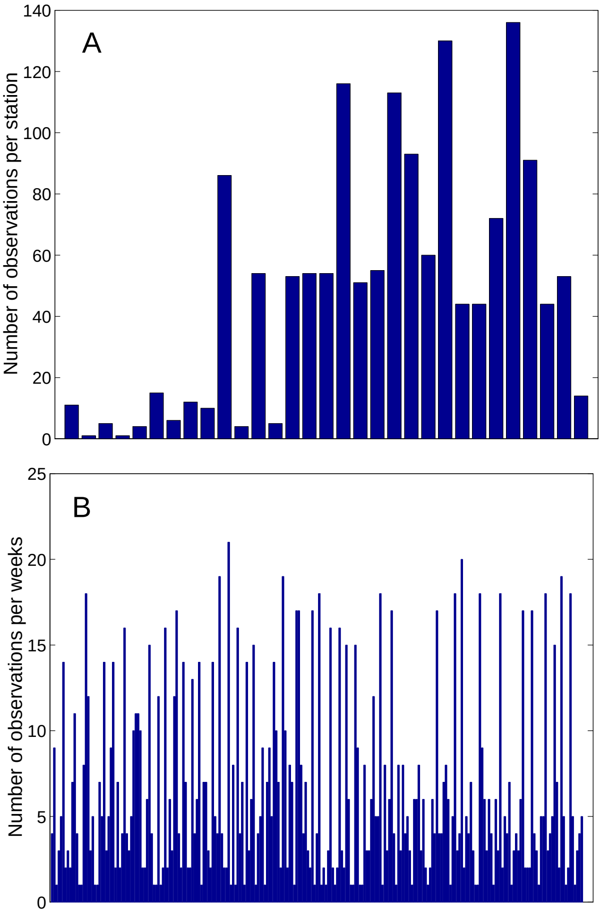

Figure F.3: *A) The number of observations of DIN per monitoring station. B) The number of observations of DIN per week.*

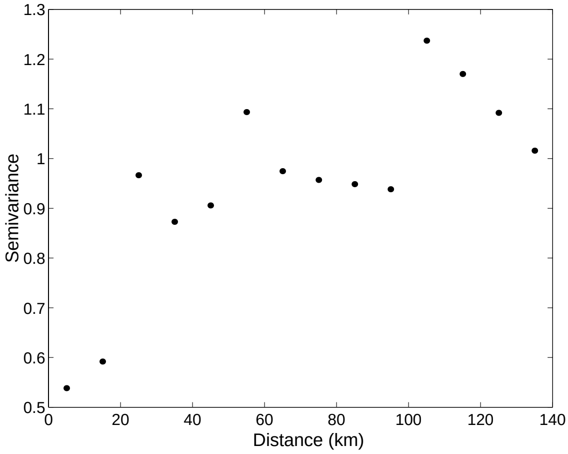

Figure F.4: *Spatial lag 0 sample semivariogram.*

F.3.3 Results and discussion

The design criterion, and hence the maximum prediction variance, for the 103 considered locations ranged between 0.93 and 1.14 when removing one of the 31 stations at a time (Figure F.5A). For the first iteration the removal of station *aj* results in the smallest design criterion, and consequently this station should be removed from the network, although we could probably equally well have removed other stations like *al*, *au* or *az* which also had small values of the design criterion. The reason for this is that the station-specific design criteria (Figure F.5) were associated with uncertainty introduced by the variation between simulations when estimating the prediction variances in each of the 103 prediction locations. Data were simulated due to the lack of data on a weekly basis. If the design problem was purely spatial with a data value for each of a number of sampling locations, no simulated data would be necessary and the uncertainty of the station-specific design criteria would be removed. This means that the proposed design approach is generally applicable when reducing the number of stations from an existing purely spatial monitoring network. In our case the design situation is complicated by the fact that data were measured in both space and time as well as by the lack of data on a weekly basis, which made simulation of data necessary.

Despite this uncertainty we adopted the strategy of removing the station with the smallest design criterion one at a time, corresponding to a backwise elimination approach. When removing more than one station a stepwise approach would probably perform better. However, for computational reasons the strategy of removing the station with the smallest design criterion one at a time was preferred.

Some agreement between the iterations is seen (Figure F.5), e.g. removing monitoring stations *be*, *ao* or *aq* is seen to result is high maximum prediction variances in all three cases. However, the design criterion for some stations changed a lot, e.g. station *al* had a low design criterion in the first iteration, and was very close to being removed, whereas the design criterion for station *al* in the second iteration was significantly increased.

The design criterion, and hence the maximum prediction variance, increased when stations were removed although not monotonically (Figure F.6). The minor decreases in the criterion were caused by the Monte Carlo error introduced when estimating the prediction variance. The design criterion increased rapidly after having removed approximately 17 monitoring, indicating that the monitoring network should comprise at least 14 stations.

In the beginning of the selection process, stations close to other stations were removed. However, after removing the first four stations a station was selected which was relatively distant to other stations, whereas stations situated close to each other, e.g. the three stations numbered 16, 18 and 19, remained in the design. This variation between removing nearby or distant stations became clearer when the distance from the selected station to the nearest station was calculated (Figure F.7). This analysis also showed that the first four monitoring stations removed from the network were located relatively close to other stations while the next three were relatively distant to other stations. This fluctuating pattern continued as monitoring stations were removed from the network and illustrated the compromise between designing for prediction and for estimation of the model parameters. The general trend in Figure F.7 was simply caused by the fact that removing sampling stations from the network increased the overall average distance between remaining stations. Furthermore, if 17 sampling stations were removed from the existing network (Figure F.6), the resulting monitoring network consisted of stations with inter-distances at all ranges.

The prediction performance of the reduced monitoring network consisting of 14 stations was also compared to the existing network of 31 stations by comparing maps of the predicted surfaces based on data from both networks (Figure F.8). Data values were simulated in the 31 sampling locations as a Gaussian random field with parameters estimated from data. Afterwards spatial predictions were computed based on all 31 data values and on data values from the 14 sampling locations in the reduced network. The reduced network was capable of identifying the area with high DIN values along the northern coast of Zealand and in the north-western part of the Kattegat, the low values in the south-western

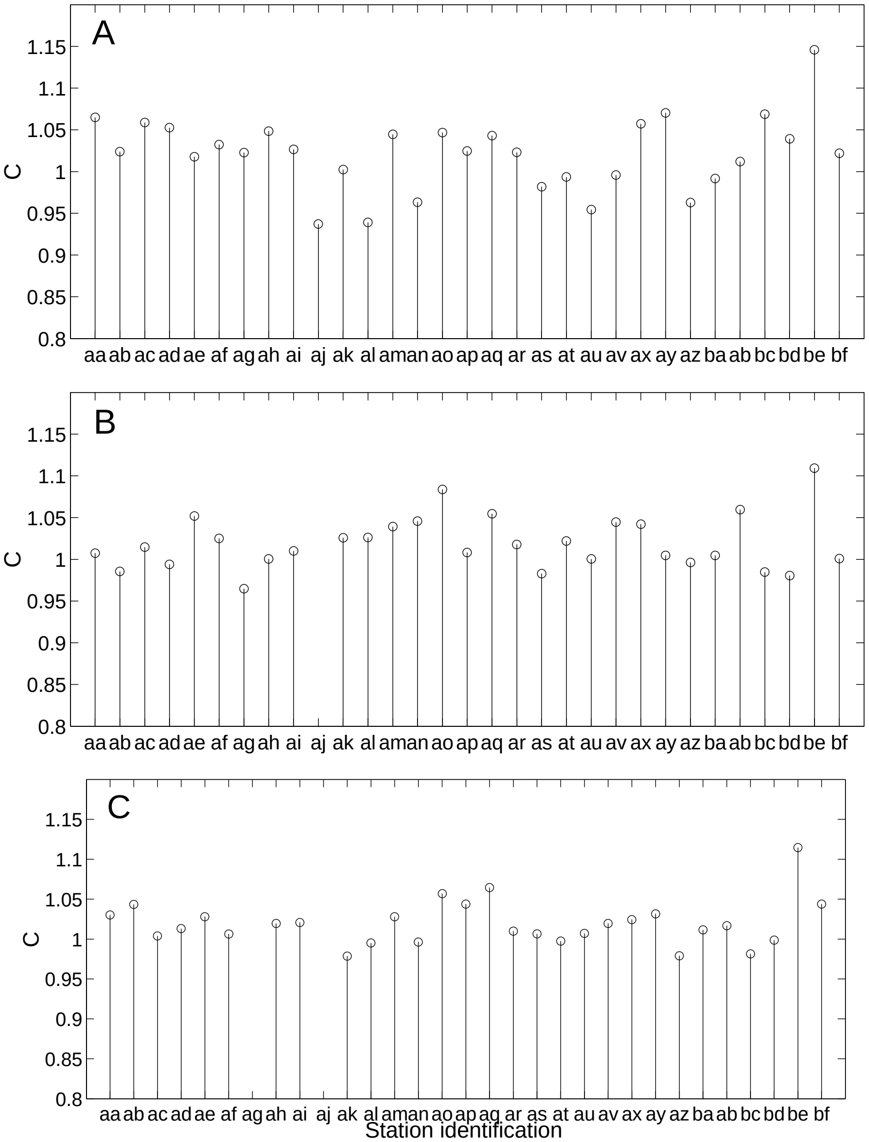

Figure F.5: *Station-specific values of the design criterion for removing the first (A), second (B) and third station (C). The locations of stations are found in Figure F.1.*

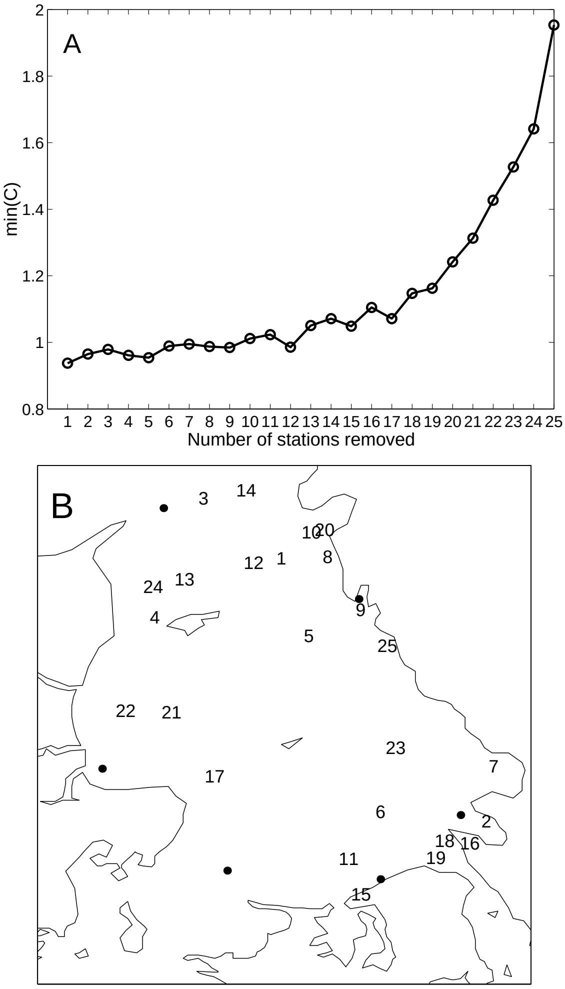

Figure F.6: *A) The minimized design criterion as a function of the number of monitoring stations removed. B) Map of Kattegat with monitoring stations numbered according to their rank in the removing strategy. The six remaining stations after removing 25 are marked by* •.

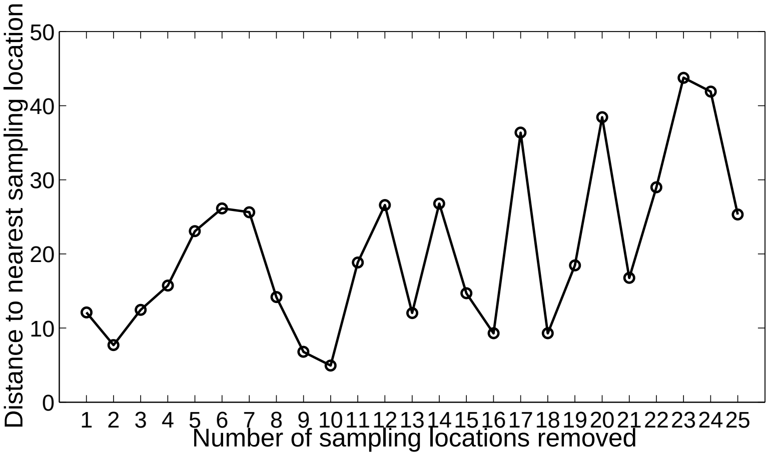

Figure F.7: *The distance from the removed monitoring station to its nearest sampling location as a function of the number of sampling locations removed.*

and south-eastern part of the Kattegat, but some details were missing in the north-eastern part of the Kattegat, due to lack of monitoring stations in this area. At this stage it is important to note that we could probably equally well have chosen 14 other monitoring stations to be our reduced network, i.e. the approach that has been outlined only give suggestions to which stations to remove. What is important about these suggestions is not so much the specific stations which it chose to remove, but more the fact that the reduced network consists of some stations close to each other as well as some more distant.

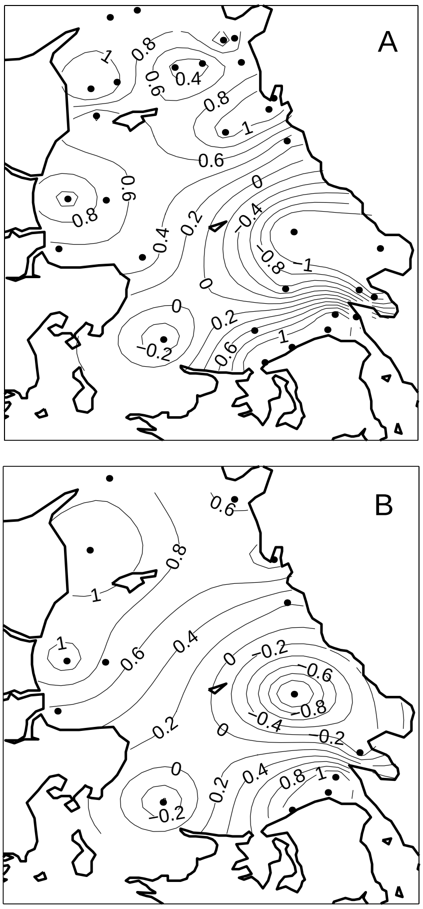

Figure F.8: *Spatial predictions based on a simulated dataset. Sampling locations are indicated by •. A) Predictions computed from the existing monitoring network of 31 stations. B) Predictions computed from the reduced monitoring network of 14 stations.*

F.4 Conclusion

This paper describes classical geostatistical methods for designing monitoring networks, i.e. for finding the locations of the optimal set of sampling stations. These methods focus either on spatial prediction assuming known model parameter or on estimation of model parameter. Rather than using these classical methods, a design approach focusing on spatial prediction, but at the same time taking uncertainty in the model parameters into account, was applied. The approach was used to reduce the number of sampling stations for measuring dissolved inorganic nitrogen in the Kattegat. The approach serves as a compromise between the conflicting issues of designing for prediction and for estimation of the model parameters. It can be used to identify the appropriate number of stations and the specific stations to be removed from existing monitoring networks.

Acknowledgements

We thank the Danish counties for collecting data within DNAMAP and the Swedish Meteorological and Hydrological Institute (SMHI) for providing Swedish monitoring data for this study.

References

Ben-Jemaa, F., Marino, M., and Loaiciga, H. (1995). Sampling design for contaminant distribution in lake sediments. *Journal of Water Resources Planning and Management*, **121**(1), 71–79.

Bogaert, P. and Russo, D. (1999). Optimal spatial sampling design for the estimation of the variogram based on a least squares approach. *Water Resources Research*, **35**(4), 1275–1289.

Carstensen, J. (2003). Spatial and temporal resolution of carbon fluxes in a shallow coastal ecosystem. *Marine Ecology Progress Series*, **252**, 35–50.

Chilès, J. and Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty.* Wiley, New York.

Cressie, N. (1993). *Statistics for spatial data.* Wiley, New York.

Diggle, P., Ribeiro, P., and Christensen, O. (2003). *Spatial statistics and computational methods*, chapter An introduction to model-based geostatistics, pp. 43–86. Lecture notes in statistics. Springer, New York.

Gustafsson, B. (2000). Time-dependent modeling of the Baltic entrance area. 1. quantification of circulation and residence times in the Kattegat and the straits of the Baltic sill. *Estuaries*, **23**, 231–252.

Handcock, M. and Stein, M. (1993). A bayesian analysis of kriging. *Technometrics*, **35**, 403–410.

Johnson, M., Moore, L., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, **26**, 131–148.

Lark, R. (2002). Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, **105**, 49–80.

Le, N. and Zidek, J. (1992). Interpolation with uncertain covariances: A bayesian alternative to kriging. *Journal of Multivariate Analysis*, **43**, 351–374.

Martin, R. (2001). Comparing and contrasting some environmental and experimental design problems. *Environmetrics*, **12**, 303–317.

McBratney, A., Webster, R., and Burgess, T. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables-I.

Theory and methods. *Computers & Geosciences*, **7**(4), 331–334.

Müller, W. and Zimmerman, D. (1999). Optimal designs for variogram estimation. *Environmetrics*, **10**, 23–37.

Müller, W. G. (2001). *Collecting spatial data: Optimum design of experiments for random fields.* Contributions to statistics. Physica-Verlag, Heidelberg (Germany).

Richardson, K. (1996). *Eutrophication in coastal marine ecosystems*, chapter Conclusion, research and eutrophication control, pp. 243–267. American Geophysical Union, Washington D.C.

Royle, J. and Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in Splus. *Computers & Geosciences*, **24**(5), 479–488.

Russo, D. (1984). Design of an optimal sampling network for estimating the variogram. *Soil Science Society of American Journal*, **48**, 708–716.

Russo, D. and Jury, W. (1988). Effect of the sampling network on estimates of the covariance function of stationary fields. *Soil Science Society of American Journal*, **52**, 1228–1234.

Spruill, T. and Candela, L. (1990). Two approaches to design of monitoring networks. *Ground Water*, **28**(3), 430–442.

Stein, M. (1999). *Interpolation of spatial data: Some theory for kriging.* Springer, New York.

Warrick, A. and Myers, D. (1987). Optimization of sampling locations for variogram calculations. *Water Resources Research*, **23**, 496–500.

Winkels, H. and Stein, A. (1997). Optimal cost-effective sampling for monitoring and dredging of contaminated sediments. *Journal of Environmental Quality*, **26**, 933–946.

Zimmerman, D. and Homer, K. (1991). A network design criterion for estimating selected attributes of the semivariogram. *Environmetrics*, **2**(4), 425–441.

Bibliography

Ababou, R., Bagtzoglou, A., and Wood, E. (1994). On the condition number of covariance matrices in kriging, estimation and simulation of random fields. *Mathematical Geology*, **26**, 99–133.

Ærtebjerg, G. (1986). Årsager till og effekter av eutrofieringen i Kattegat och Belthavet. In *22 Nordiska Symposiet om Vattenforskning*, Helsinki (Finland). Nordforsk.

Andersson, L. (1996). Trends in nutrient and oxygen concentrations on the Skagerrak-Kattegat. *Journal of Sea Research*, **35**, 63–71.

Andersson, L. and Rydberg, L. (1988). Trends in nutrient and oxygen conditions within the Kattegat: Effects of local nutrient supply. *Estuarine, Coastal and Shelf Science*, **26**, 559–579.

Anton, K., Liebzeit, G., Rudolph, C., and Wirth, H. (1993). Origin, distribution and accumulation of organic carbon in the Skagerrak. *Marine Geology*, **111**, 287–297.

Asli, M. and Marcotte, D. (1995). Comparison of approaches to spatial estimation in a bivariate context. *Mathematical Geology*, **27**(5), 641–658.

Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical modeling and analysis for spatial data*. Monographs on statistics and applied probability. Chapman and Hall, New York.

Ben-Jemaa, F., Marino, M., and Loaiciga, H. (1995). Sampling design for contaminant distribution in lake sediments. *Journal of Water Resources Planning and Management*, **121**(1), 71–79.

Bogaert, P. and Russo, D. (1999). Optimal spatial sampling design for the estimation of the variogram based on a least squares approach. *Water Resources Research*, **35**(4), 1275–1289.

Box, G. and Jenkins, G. (1970). *Time series analysis, forecasting and control*. Holden-Day, San Francisco.

Brown, P., Diggle, P., Lord, M., and Young, P. (2001). Space-time calibration of radar rainfall data. *Journal of the Royal Statistical Society, series C*, **50**, 221–241.

Brown, P., Roberts, G., Kåresen, K., and Tonellato, S. (2000). Blur-generated non-separable space-time models. *Journal of the Royal Statistical Society, series B*, **62**, 847–860.

Brus, D., DeGruijter, J., Marsman, B., Visschers, R., Bregt, A., Breeuwsma, A., and Bouma, J. (1996). The performance of spatial interpolation methods and chloropleth maps to estimate properties at points: A soil survey case study. *Environmetrics*, **7**(1), 1–16.

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, J. (2003). Spatial and temporal resolution of carbon fluxes in a shallow coastal ecosystem. *Marine Ecology Progress Series*, **252**, 35–50.

Chilès, J. and Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty*. Wiley, New York.

Christensen, O. (2002). *Methodology and applications in non-linear model-based geostatistics*. Ph.D. thesis, Department of Matrhematical Sciences. AAlborg University.

Christensen, P. (1998). The Danish marine environment: has action improved its state ? Marine Research Programme HAV90 62, Danish Environmental Protection Agency.

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

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

Cressie, N. (1993). *Statistics for spatial data.* Wiley, New York.

Cressie, N. and Hawkins, D. (1980). Robust estimation of the semivariogram. *Mathematical Geology*, **12**, 115–125.

Cressie, N. and Huang, H. (1999). Classes of nonseparable, spatiotemporal stationary covariance functions. *Journal of the American Statistical Association*, **94**, 1330–1340.

De Cesare, L., Myers, D., and Posa, D. (2001a). Estimating and modelling space-time correlation structures. *Statistics & Probability Letters*, **51**, 9–14.

De Cesare, L., Myers, D., and Posa, D. (2001b). Product-sum covariance for space-time modelling: an environmental application. *Environmetrics*, **12**, 11–23.

De Cesare, L., Myers, D., and Posa, D. (2002). FORTRAN programs for space-time modeling. *Computers & Geosciences*, **28**, 205–212.

De Iaco, S., Myers, D., and Posa, D. (2001). Space-time analysis using a general product-sum model. *Statistics & Probability Letters*, **52**(1), 21–28.

De Iaco, S., Myers, D., and Posa, D. (2002). Nonseparable space-time covariance models: Some parametric families. *Mathematical Geology*, **34**(1), 23–42.

Diggle, P. and Lophaven, S. (2004). Bayesian geostatistical design. *Scandinavian Journal of Statistics. (submitted).*

Diggle, P., Moyeed, R., and Tawn, J. (1998). Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society, series C*, **47**, 299–350.

Diggle, P., Ribeiro, P., and Christensen, O. (2003). *Spatial statistics and computational methods*, chapter An introduction to model-based geostatistics, pp. 43–86. Lecture notes in statistics. Springer, New York.

Edler, L. (1984). *Gödning av havsområden kring Sverige*, chapter Västerhavet. SNV PM 1808. SEPA.

Figueira, R., Sousa, A., Pacheco, A., and Catarino, F. (2001). Use of secondary information in space-time statistics for biomonitoring studies of saline deposition. *Environmetrics*, **12**, 203–217.

Gneiting, T. (2001). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, **97**, 590–600.

Granéli, E. (1987). Nutrient limitation of phytoplankton biomass in a brackish water bay highly influenced by river discharge. *Estuarine, Coastal and Shelf Science*, **25**, 563–569.

Granéli, E., Wallström, K., Larsson, U., Granéli, W., and Elmgren, R. (1990). Nutient limitation of primary production in the Baltic Sea area. *Ambio*, **19**, 142–151.

Grunsky, E. (2002). R: A data analysis and statistical environment - an emerging tool for the geosciences. *Computers & Geosciences*, **28**(10), 1219–1222.

Gustafsson, B. (2000). Time-dependent modeling of the Baltic entrance area. 1. quantification of circulation and residence times in the Kattegat and the straits of the Baltic sill. *Estuaries*, **23**, 231–252.

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

Handcock, M. and Stein, M. (1993). A bayesian analysis of kriging. *Technometrics*, **35**, 403–410.

Hosseini, E., Gallichand, J., and Caron, J. (1993). Comparison of several interpolators for smoothing hydraulic conductivity data in south west Iran. *Transactions of the American Society of Agricultural Engineers*, **36**(6), 1687–1693.

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

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**(3), 299–314.

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

Jakobsen, F. (1997). Hydrographic investigation of the northern Kattegat front. *Continental Shelf Research*, **17**, 533–554.

Johnson, M., Moore, L., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, **26**, 131–148.

Journel, A. (1980). The lognormal approach to predicting local distributions of selective mining unit grades. *Mathematical Geology*, **12**, 285–303.

Kiørboe, T. (1996). *Eutrophication in coastal marine ecosystems*, chapter Material flux in the water column, pp. 1–20. American Geophysical Union, Washington D.C.

Kitanidis, P. (1986). Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Research*, **22**, 499–507.

Kitanidis, P. (1997). *Introduction to geostatistics: Applications in hydrogeology.* Cambridge University Press, New York.

Kronvang, B., Ærtebjerg, G., Grant, R., Kristensen, P., Hovmand, M., and Kirkegaard, J. (1993). Nationwide monitoring of nutrients and their ecological effects: State of the Danish aquatic environment. *Ambio*, **22**, 176–187.

Kyriakidis, P. and Journel, A. (1999). Geostatistical space-time models: A review. *Mathematical Geology*, **31**, 651–684.

Lark, R. (2002). Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, **105**, 49–80.

Laslett, G. (1994). Kriging and splines: An empirical comparison of their predictive performance in some applications. *Journal of the American Statistical Association*, **89**(426), 391–409.

Le, N. and Zidek, J. (1992). Interpolation with uncertain covariances: A bayesian alternative to kriging. *Journal of Multivariate Analysis*, **43**, 351–374.

Lophaven, S. (2001). Reconstruction of data from the marine environment. Master's thesis, Technical University of Denmark, Kgs. Lyngby (Denmark).

Mardia, K., Goodall, C., Redfern, E., and Alonso, F. (1998). The kriged Kalman filter (with discussion). *Test*, **7**, 217–285.

Martin, R. (2001). Comparing and contrasting some environmental and experimental design problems. *Environmetrics*, **12**, 303–317.

McBratney, A. and Webster, R. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables-II. Program and examples. *Computers & Geosciences*, **7**(4), 335–365.

McBratney, A., Webster, R., and Burgess, T. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables-I. Theory and methods. *Computers & Geosciences*, **7**(4), 331–334.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models.* Chapman and Hall, London.

McGilchrist, C. (1989). Bias of ML and REML in regression models with ARMA errors. *Journal of Statistical Computation and Simulation*, **32**, 127–136.

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

Müller, W. and Zimmerman, D. (1999). Optimal designs for variogram estimation. *Environmetrics*, **10**, 23–37.

Müller, W. G. (2001). *Collecting spatial data: Optimum design of experiments for random fields.* Contributions to statistics. Physica-Verlag, Heidelberg (Germany).

Pardo-Iguzquiza, E. (1997). MLREML: A computer program for the inference of spatial covariance parameters by maximum likelihood and restricted maximum likelihood. *Computers & Geosciences*, **23**(2), 153–162.

Pebesma, E. (2004). Multivariable geostatistics in S: The gstat package. *Computers & Geosciences*, **30**, 683–691.

Rahm, L., Sandén, P., Wulff, F., Stålnacke, P., and Conley, D. (1996). Time series analysis of nutrient inputs to the Baltic Sea and changing DSi/DIN ratios. *Marine Ecology Progress Series*, **130**, 221–228.

Ribeiro, P. (1999). Bayesian inference in Gaussian model-based geostatistics. Technical Report St-99-08, Lancaster University.

Ribeiro, P., Christensen, O., and Diggle, P. (2003). Geor and georglm: Software for model-based geostatistics. In Hornik, K., Leisch, F., and Zeileis, A., editors, *Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003).*

Richardson, K. (1985). Plankton distribution and activity in the North Sea/Skagerrak-Kattegat frontal area in April 1984. *Marine Ecology Progress Series*, **26**, 233–244.

Richardson, K. (1996). *Eutrophication in coastal marine ecosystems*, chapter Conclusion, research and eutrophication control, pp. 243–267. American Geophysical Union, Washington D.C.

Royle, J. and Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in Splus. *Computers & Geosciences*, **24**(5), 479–488.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society, series B*, **63**, 325–338.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, **29**, 31–49.

Russo, D. (1984). Design of an optimal sampling network for estimating the variogram. *Soil Science Society of American Journal*, **48**, 708–716.

Russo, D. and Jury, W. (1988). Effect of the sampling network on estimates of the covariance function of stationary fields. *Soil Science Society of American Journal*, **52**, 1228–1234.

Schlather, M. (1999). Introduction to positive definite functions and to unconditional simulation of random fields. Technical Report St-99-10, Lancaster University.

Smith, R. (2001). Environmental statistics. Department of Statistics, University of North Carolina. Prepared in connection with NSF-CBMS regional conference in the mathematical sciences: Environmental statistics, University of Washington, June 25-29, 2001.

Spruill, T. and Candela, L. (1990). Two approaches to design of monitoring networks. *Ground Water*, **28**(3), 430–442.

Stein, M. (1999). *Interpolation of spatial data: Some theory for kriging*. Springer, New York.

Swallow, W. and Monahan, J. (1984). Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics*, **26**, 47–57.

Wackernagel, H. (2003). *Multivariate geostatistics: An introduction with applications*. Springer, Berlin.

Ward, R., Loftis, J., and McBride, G. (1986). The ”Data-rich but information-poor” syndrome in water quality monitoring. *Environmental Management*, **10**(3), 291–297.

Warrick, A. and Myers, D. (1987). Optimization of sampling locations for variogram calculations. *Water Resources Research*, **23**, 496–500.

Winkels, H. and Stein, A. (1997). Optimal cost-effective sampling for monitoring and dredging of contaminated sediments. *Journal of Environmental Quality*, **26**, 933–946.

Zhang, X., Van Eijkeren, J., and Heemink, A. (1995). On the weighted least-squares method for fitting a semivariogram model. *Computers & Geosciences*, **21**, 605–608.

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

Zimmerman, D. and Homer, K. (1991). A network design criterion for estimating selected attributes of the semivariogram. *Environmetrics*, **2**(4), 425–441.

Zimmerman, D. and Zimmerman, M. (1991). A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictions. *Technometrics*, **33**(1), 77–91.