Interpolation and optimal monitoring in space and time

Promotoren: prof. dr. dr. h.c. D.A.M.K. Rasch
emeritus hoogleraar in de Wiskundige Statistiek

prof. dr. ir. A. Stein
hoogleraar in de Wiskundige en Statistische Methoden,
met speciale aandacht voor de ruimtelijke statistiek

Copromotor: dr. B. van Putten
universitair hoofddocent Wiskundige en Statistische Methoden

Promotiecommissie: prof. dr. N.A.C. Cressie (Ohio State University)

prof. dr. P. van Beek (Wageningen Universiteit)

prof. dr. L. Hordijk (Wageningen Universiteit)

dr. A.L.M. Dekkers (RIVM, Bilthoven)

Eric Peter Jan Boer

Interpolation and optimal monitoring in space and time

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit,
prof. dr. ir. L. Speelman,
in het openbaar te verdedigen
op maandag 25 februari 2002
des namiddags te vier uur in de Aula.

Boer, E.P.J., 2002

Interpolation and optimal monitoring in space and time.
PhD Thesis Wageningen University, Wageningen, the Netherlands.
- With ref.- With summary in Dutch.

ISBN 90-5808-573-2

Voorwoord

Derde kerstdag 2001, een goede tijd om terug te blikken op het werk dat ik de afgelopen jaren heb mogen doen bij de leerstoelgroep Wiskundige en Statistische Methoden. Het resultaat ligt nu voor u, maar dit had niet tot stand gekomen zonder de ondersteuning van een groot aantal mensen.

Ik wil als eerste mijn promotoren noemen, Dieter Rasch en Alfred Stein. Dieter, ik weet nog dat ik mijn twijfels had over mijn toekomstig AIO bestaan. Je maakte de keus heel gemakkelijk, door me voor te houden dat als ik veel geld wilde verdienen maar glazenwasser moest worden. Ik heb geen spijt gehad, en ik ben je dan ook dankbaar voor al je adviezen in de afgelopen jaren. Alfred, jij hebt vooral bij het afronden van dit proefschrift geholpen. Dankzij jou ligt er nu dit resultaat zonder al te lange uitloop.

In het projectvoorstel staat "een open-deur-beleid ten behoeve van het vervullen van een klankbordfunctie zijn vanzelfsprekend". Die klankbordfunctie heeft Aad van Eijnsbergen willen vervullen. Bij hem kon ik altijd terecht met allerhande problemen. Inhoudelijk, maar ook gewoon om even mijn hart te luchten als het een keer tegen zat of als ik niet goed wist wat te doen. Jouw heldere analyses had ik niet graag willen missen. Bij mijn andere begeleider en copromotor, Bram van Putten, heb ik twee keer Inleiding Statistiek gevolgd. De eerste keer als student, de tweede keer als docent om te leren hoe het moet. Bram, het was me een waar genoegen! Ik kan me herinneren dat studenten T-shirts hadden laten bedrukken met jouw foto erop. Als er nog exemplaren over zijn wil ik er graag één van hebben. Eligius Hendrix mag zeker niet in dit rijtje ontbreken. Verschillende keren nam je me op sleeptouw. Op reis naar het buitenland, maar ook bij het schrijven van artikelen. Ons congresbezoek in Florence was een hoogtepunt van de afgelopen jaren. In het vervolg moet je wel een betere smoes verzinnen als je weer van kamer wil ruilen!

Mijn mede-auteurs van hoofdstuk 2, Kirsten de Beurs en Dewi Hartkamp, wil ik danken voor de prettige samenwerking. Kirsten fungeerde tevens als proefkonijn voor mijn beginnende kookkunsten. Ik vind het jammer dat je Nederland voor een tijdje gaat verlaten en naar de VS gaat verhuizen om daar promotieonderzoek te gaan doen, maar ik gun je daar van harte een fijne en leerzame tijd. Arnold Dekkers en het RIVM hebben me geholpen bij het schrijven van hoofdstuk 5, waarvoor mijn hartelijke dank.

Het leven bestaat uit meer dan werk. In Annemarie Pielaat, Mark Huiskes, Maria João Paulo en Anna Rzepiela had ik collega's die dat ook vonden. Een persoonlijk woord is wel op zijn plaats. Annemarie, jij gaat ons voor op allerlei terreinen van het leven. Ik wou dat ik de helft van je vastberadenheid en doorzettingsvermogen had. Als je weer terug bent uit Canada, gaan we dan nog een keer wandelen (inclusief appeltaart met slagroom)? Mark and Anna, thanks for all the low-calorie candle-light diners and your hospitality. I enjoyed our trips on Saturdays very much. I will never forget the moving of "the couch" and the helpful advices of you, Anna. Transport of floor-covering was also not without problems, but we get some experience. João, ik vind het fantastisch dat je het Nederlands al zo goed onder de knie hebt. Afgezien van het weer, vind je het hier goed toeven en dat straal je ook uit. Ik ben blij dat je mijn paranimf wilt zijn.

Graag wil ik ook de andere collega's bij wiskunde bedanken voor hun belangstelling in mijn onderzoek en de fijne tijd die ik bij jullie gehad heb. Het secretariaat stond altijd voor me klaar om allerlei dingen te regelen. Ria, nog bedankt voor je deskundig advies aan een lid van de benoemings-commissie!

In juni 2000 mocht ik paranimf zijn bij de promotie van mijn broer Martin. Het was zeker geen verplichting voor me om jou ook als mijn paranimf te vragen. Martin heeft me in de loop van de jaren veel geholpen, vanaf de middelbare school tot en met mijn promotie-onderzoek. Meestal had ik alleen interesse voor het trucje, waar jij ook wilde weten wat het principe erachter was. Schaken heb ik ook van jou geleerd. Nu je in Wageningen woont, verwacht ik een bezoek aan SVW. Ook jij zal dan kunnen profiteren van de gastvrijheid van Max de Ruiter.

Oma den Hoed, ooms en tantes en verdere familie zijn altijd belangstellend geweest wat die twee boertjes nou toch allemaal uitvoerden. Op 25 februari hoop ik u het één en ander uit te leggen. Mevrouw Blaak wil ik bedanken voor haar goede verzorging gedurende mijn studie en het begin van mijn promotieonderzoek.

Bij de Hervormde Gemeente Wageningen, wijkgemeente 1, voel ik me zeer thuis. Ik wil in het bijzonder de jonge lidmatenkring en de kerkenraad noemen. Het is mij een vreugde om met jullie samen op weg te zijn.

Tot slot, mijn ouders. Dit proefschrift wil ik aan jullie opdragen. Ik ben dankbaar voor alles wat jullie voor mij hebben willen betekenen.

Eric Boer

Stellingen

behorende bij het proefschrift van Eric Boer, *Interpolation and optimal monitoring in space and time,* Wageningen Universiteit, 25 februari 2002.

1.
Helaas geldt voor meetnet-evaluatie en –optimalisatie dat er geen "eenvoudige truc" is.

TNO-rapport – Evaluatie van provinciale grondwatermeetnetten. Deel 2B
Dit proefschrift

2.
In de geostatistische literatuur wordt veel aandacht besteed aan cokriging als methode om additionele informatie te gebruiken voor interpolatie doeleinden. Cokriging is echter tijdrovend om te implementeren en heeft zeer beperkte mogelijkheden.

Dit proefschrift

3.
Statistiek speelt een belangrijke rol in veel toegepast wetenschappelijk onderzoek. Het is dan ook van essentieel belang dat goed statistisch onderwijs een prominente plaats krijgt in de opleidingprogramma's.

4.
In theologische discussies worden de vruchten van de Geest (Galaten 5) vaak node gemist.

5.
Tijdnood komt de kwaliteit van stellingen niet ten goede.

6.
Het openstellen van winkels op zondag is een uiting van de gezindheid "nooit genoeg".

7.
Wie altijd bezig is gelukkig te worden, zal het nooit zijn.

Blaise Pascal

O Vader, dat Uw liefd' ons blijk';
O Zoon, maak ons Uw beeld gelijk;
O Geest, zend Uwen troost ons neer;
Drieënig God, U zij al d' eer.

Avondzang : 7

Aan mijn ouders

Abstract

This thesis shows how statistics can be used for both analysing data and for determining the (optimal) design for collecting data in environmental research. An important question is often where to place monitoring stations to meet the objective of measuring as good as possible. In this thesis it is shown how existing monitoring networks can be adjusted on the basis of quantitative criteria. These criteria are based on aspects of spatial(-temporal) interpolation.

A case study of climate variables in Jalisco State of Mexico is used to investigate the use of interpolation techniques. The climate variables monthly maximum temperature and monthly mean precipitation are predicted on a regular grid of points on the basis of measurements at climate stations. Four forms of kriging and three forms of thin plate splines are discussed. From these techniques, trivariate regression-kriging and trivariate thin plate splines performed best.

The optimal adjustment of existing monitoring networks is investigated for three case studies with different criteria. In the first place, a monitoring network adjustment is investigated for estimation of the semivariance function, whereby the criterium is based on the theory of optimal design of experiments. Secondly, we develop and apply a methodology to reduce an existing monitoring network to find an optimal configuration of a smaller network. In this case a criterion based on locally weighted regression with two different weight functions is used. The methodology is applied to the Dutch national SO_2 network and offers the possibility to include different politically relevant options in the model by weight criteria. As a third case study, a monitoring network for groundwater level is considered. It focusses on a possible reduction of the number of measurements at this monitoring network without losing much information about the groundwater level at the different piezometers. The investigations of a reduction of the number of measurements is based on a geostatistical spatial-temporal model. The results show that the monitoring effort of the network can be reduced.

Finding optimal designs involves several optimization problems. In this thesis several methods are developed and applied to solve these problems. For small problems full enumeration of all possible configurations is possible with a branch-and-bound algorithm. In this way, it is ensured that the global optimum is found. If full enumeration of all possible monitoring networks is impossible, a search algorithm is applied to find a (sub)-optimal solution.

Contents

1 **Introduction** 3
- 1.1 Spatial-(temporal) interpolation . . . 3
- 1.2 Optimal experimental design . . . 4
- 1.3 Optimization of monitoring networks . . . 5
- 1.4 Objectives of the thesis . . . 5
- 1.5 Outline of the thesis . . . 6

2 **Kriging and thin plate splines for mapping climate variables** 7
- 2.1 Introduction . . . 8
- 2.2 Material and Methods . . . 8
  - 2.2.1 Data sets . . . 8
  - 2.2.2 Interpolation techniques . . . 9
  - 2.2.3 Comparison of interpolation techniques . . . 15
- 2.3 Results . . . 15
- 2.4 Discussion . . . 16

3 **Global optimization problems in optimal design of experiments in regression models** 21
- 3.1 Introduction . . . 22
- 3.2 Theory of optimal design of experiments . . . 22
  - 3.2.1 Criteria . . . 25
  - 3.2.2 Properties of optimal design problems . . . 25
- 3.3 Searching the optimal design . . . 28
  - 3.3.1 Analytical results . . . 28
  - 3.3.2 Special algorithms for optimal designs . . . 29
  - 3.3.3 Combinatorial optimization of the optimal design . . . 31
  - 3.3.4 General purpose optimization . . . 31
- 3.4 Conclusions . . . 33

4 **Optimization of monitoring networks for the estimation of the semivariance function** 35
- 4.1 Introduction . . . 36
- 4.2 Theory . . . 37
  - 4.2.1 The semivariance function . . . 37

4.2.2 Optimal experimental design for estimation of the semivariance function . . . 38
4.3 Case study . . . 39
4.4 Results . . . 40
4.4.1 Visualization of the problem . . . 41
4.4.2 Results of the branch-and-bound algorithm . . . 41
4.4.3 Robustness analyses of the optimal design . . . 43
4.5 Discussion and Conclusions . . . 44

5 Optimization of a monitoring network for SO_2 45
5.1 Introduction . . . 46
5.2 Locally weighted regression . . . 46
5.2.1 Model formulation . . . 46
5.2.2 Optimal design for locally weighted regression . . . 48
5.2.3 Case study . . . 49
5.2.4 Specification of different design criteria . . . 50
5.2.5 Optimal reduction of a monitoring network . . . 52
5.3 Results and discussion . . . 53
5.3.1 Maps of interpolated SO_2 values with locally weighted regression . . . 53
5.3.2 Comparison of search algorithms . . . 54
5.3.3 Optimized monitoring networks . . . 56
5.4 Conclusions . . . 57

6 Optimization of a monitoring network for groundwater level 61
6.1 Introduction . . . 62
6.2 The data set . . . 63
6.3 Spatial-temporal modelling of groundwater level . . . 64
6.3.1 Elimination of seasonality . . . 65
6.3.2 Spatial-temporal variability . . . 66
6.4 Reduction of the number of measurements . . . 69
6.4.1 Reduction of frequency within a piezometer . . . 69
6.4.2 Reducing the monitoring network from a spatial-temporal perspective . . . 72
6.5 Discussion and conclusion . . . 73

7 General discussion and conclusions 75

Appendix: Branch-and-bound algorithm 79

Bibliography 85

Samenvatting 93

Curriculum vitae 97

Chapter 1

Introduction

During the last decades much attention is paid to environmental research. As an important determinant for the quality of life, the environment is worth to be investigated thoroughly. Environmental problems can occur at a small scale, such as a pollutant highly concentrated around a source or at a global scale, like global warming due to increased carbon dioxide concentration and a possible reduction of the ozone layer in the stratosphere. Environmental research is often related to the influence of human beings on environmental processes. Empirical observations are necessary to test theories about possible trends and spatial and/or temporal variability. At this stage statistics is brought into play, both for analysing the data and for determining the (optimal) sampling design. The various environmental problems yield a broad range of statistical problems and challenges. If only few data are available the proper selection of a statistical method becomes more important to avoid misleading results. In this thesis, three different statistical subjects are discussed and applied to real-world problems: spatial-(temporal) interpolation, optimal experimental design and optimization of monitoring networks.

1.1 Spatial-(temporal) interpolation

Spatial interpolation techniques are often applied to make maps of continuous variables for fields like meteorology, agriculture, hydrology and environmental engineering. Sparsely distributed point observations in an area are interpolated to a regular grid of points. Nowadays, many interpolation techniques are available (Cressie, 1991). The choice for an appropriate interpolation technique is increasingly important when data are scarce. When the variable of interest is sampled sparsely, covariables can be helpful to improve prediction accuracy.

The geostatistical prediction approach (kriging) is a popular method among different interpolation techniques. The origins of kriging can be found in Krige (1951). Kriging has been further developed by Matheron (1963), who gave it a better mathematical foundation. Geostatistical methods are optimal in the sense that it yields unbiased predictions and minimizes the variance of the prediction errors. Additional theoretical background and different forms of kriging can be found in Chilès and Delfiner (1999). In this thesis we will mainly concentrate on different forms and extensions of kriging for spatial-(temporal) prediction. Besides, two other interpolation techniques are applied: thin plate splines and locally weighted regression. Kriging and thin plate splines are formally alike, but practically very different (Cressie, 1991). In Chapter 2 of this thesis it is shown that the two methods can benefit from each other. It is shown how additional information can be included to improve prediction accuracy. Locally weighted regression (Chapter 5) estimates the trend surfaces by nonparametric regression. It has the advantage that no estimation is needed for trend and spatial variability, as in geostatistics.

Modelling spatial-temporal processes can be done in different ways. It is possible to extend time series models to regionalized time series models, e.g. Pfeifer and Deutsch (1980) and Knotters (2001). Hutchinson (1995) gives an application of the use of splines in stochastic spatial-temporal weather models. Chapter 6 of this thesis applies a geostatistical approach (e.g. Rouhani and Hall, 1989; Heuvelink *et al.*, 1997; Cressie and Huang, 1999).

1.2 Optimal experimental design

Parameter estimation provides a link between data and models. A well-designed experiment is an efficient method to collect data to estimate the model parameters as good as possible. The basic idea of the theory of optimal experimental design is that variances of parameter estimates depend upon the experimental design and can be minimized. The theory of optimal experimental design was originally developed by Kiefer (1959), followed by books of Fedorov (1972) and Silvey (1980). More recently Atkinson and Donev (1992) showed statistical aspects and relevance for practical applications. For nonlinear models the optimal design depends on values of model parameters and is therefore called a locally optimal design, i.e. depending on the parameter values of the model. In more simple cases the optimal design can be calculated analytically, but for other more restricted optimization problems algorithms have to be developed to reach it (Chapter 3). Müller (1998) showed how design criteria for the spatial configuration of sampling points can be derived from the theory of optimal experimental design. It is important to note that optimal experimental design can only be applied when a priori knowledge of the model exists (model-based). If investigations are still in the exploratory phase, designs have to be chosen which are robust and allow for model building and model validation. The case studies elaborated in this thesis are all after the exploratory phase (Chapters 4, 5 and 6).

1.3 Optimization of monitoring networks

Monitoring networks are intended to collect empirical observations of environmental processes in water, air and/or soil. Designers of environmental monitoring networks face several problems. An important problem is that there are usually many objectives rather than just one. Therefore, a formulation of an optimization criterion is difficult. Optimization criteria can be based on deterministic properties, like maximizing the minimum distance between observation locations (Müller, 1998), on stochastic properties like minimizing the kriging variance (e.g. McBratney and Webster, 1981; Van Groenigen, 1999) or minimizing the estimation variance of parameters of a certain model (Müller, 1998). Other problems which monitoring network designers have to face are sampling constraints. These can be financial restrictions, physical limitations like buildings (Van Groenigen and Stein, 1998) or wishes of (local) authorities. This thesis focusses on adaptation of existing monitoring networks that are designed in the past, and current opinions and constraints may have changed so that adaptations of the existing monitoring is desirable.

Given an optimization criterion, algorithms are needed to find the optimal configuration of sampling points in space. Van Groenigen (1999) developed a general optimization algorithm, called Spatial Simulated Annealing and shows how this algorithm can solve several sampling problems with different criteria. In this thesis a combinatorial approach is applied. If a limited number of candidate sampling points is considered, the problem can be solved by full enumeration (Chapter 4). If larger problems are considered, search algorithms have to be applied (Chapter 5).

For optimizing environmental monitoring networks not only the question of "where" to measure has to be answered, the frequency of sampling is also important. In Chapter 6 of this thesis a geostatistical approach for optimizing a spatial-temporal monitoring network is applied.

1.4 Objectives of the thesis

The objective of this thesis is to develop and apply statistical methods for interpolation and for optimizing monitoring networks from a model-based perspective. Given models for spatial-(temporal) interpolation of environmental phenomena, criteria for optimizing monitoring networks are formulated. The arising optimization problem has to be solved by development and application of optimization tools. This leads to three research aims of this thesis:

- Study statistical methods for spatial-(temporal) interpolation.
- Actual optimization of environmental monitoring networks for different criteria.
- Development and application of algorithms necessary to optimize the monitoring networks.

1.5 Outline of the thesis

This thesis can be considered as a collection of 5 papers which can be read to a great extent independently. In this section an outline is given of problems addressed in Chapters 2-6.

Chapter 2. The problem considered in this chapter was raised by the International Maize and Wheat Improvement Center (CYMMIT). One likes to know, how information from meteorological stations and elevation in the Jalisco State in Mexico can be used to generate climate maps. Four forms of kriging and three forms of thin plate splines are considered to solve this problem. The use of elevation of the area as covariable is important to improve prediction accuracy. The geostatistical approach introduced in this chapter is also applied in the Chapters 4 and 6.

Chapter 3. This chapter is an overview of the theory of optimal experimental design from a perspective of global optimization. It shows how various applications of optimal design of experiments determine the structure of corresponding (challenging) global optimization problems. Different ways of solving the various global optimization problems are shown. Algorithms developed in this chapter are used for the optimization of monitoring networks in the next chapters.

Chapter 4. Three new aspects concerning the use of optimal experimental design for estimation of the semivariance function are added in this chapter. The first aspect is a visualization of a simple adjustment of a monitoring network. Secondly, a branch-and-bound algorithm is applied to calculate an exact optimal configuration of monitoring sites. Finally, a robustness study of the optimal monitoring design against misspecified parameter values and model choice is given.

Chapter 5. In this chapter a methodology is developed to reduce an existing SO_2 network, part of the National Institute of Public Health and the Environment in the Netherlands (RIVM). A criterion based on locally weighted regression is formulated with various weight functions. Several measuring objectives are investigated. Because full enumeration of all possible monitoring networks is technically impossible, search algorithms are developed to find (sub)-optimal solutions.

Chapter 6. In the previous chapters problems are discussed from a spatial perspective. In Chapter 6 we look at a spatial-temporal monitoring network, which was found in the measuring of groundwater level in the Veluwe area. We focus on the reduction of the number of measurements of groundwater level of the monitoring network, both in reducing the measuring frequency at piezometers and removal of piezometers. This reduction is based on a geostatistical spatial-temporal model.

The main conclusions are summarized in Chapter 7.

Chapter 2

Kriging and thin plate splines for mapping climate variables

E.P.J. Boer, K.M. de Beurs and A.D. Hartkamp

International Journal of Applied Earth Observation and Geoinformation (2001) **3(2)**: *146-154*

Four forms of kriging and three forms of thin plate splines are discussed in this paper to predict monthly maximum temperature and monthly mean precipitation in Jalisco State of Mexico. Results show that techniques using elevation as additional information improve the prediction results considerably. From these techniques, trivariate regression-kriging and trivariate thin plate splines performed best. Results of monthly maximum temperature are much clearer than results of monthly mean precipitation, because the modelling of precipitation is more troublesome due to higher variability in the data and their non-Gaussian character.

2.1 Introduction

Statistical interpolation techniques are commonly applied in Geographical Information Systems (GIS). Data collected on a sparse network of measurement points are interpolated to a regular grid of points. Burrough and McDonnell (1998, p.158) show a table with characteristics of ten classes of interpolation techniques. In papers published recently (Goodale *et al.*, 1998; Dirks *et al.*, 1998, Pardo-Igúzquiza, 1998a; Goovaerts, 2000), a comparison is made between several of those interpolation techniques. An important question is often how additional information can be used to increase the prediction accuracy. In this paper, several ways of including additional information classified into two widely used classes of interpolation techniques - thin plate splines and kriging - will be discussed.

Climate variables provide an essential input for crop growth simulation models. Climate maps (surfaces) can be generated from a network of measurement stations - with measurements of precipitation, temperature, solar radiation, etc.- through interpolation. Accuracy of predictions of weather conditions at interpolated sites is important for crop growth simulation. Rosenthal *et al.* (1998), for example, state that the greater variability in their crop growth simulation results is most likely due to the relatively coarse grid for spatial interpolation of precipitation.
The International Maize and Wheat Improvement Center (CIMMYT) aims to improve productivity and sustainability of smallholder maize and wheat systems in developing countries. Crop growth simulation models are used to evaluate the opportunities and limitations of these production systems (e.g. Hartkamp *et al.*, 1998). Climate maps can provide essential input to crop models. The long-term monthly mean precipitation and long-term monthly maximum temperature, measured over a sparse network of climate stations in the Jalisco State of Mexico, will be used as a case study in this paper. A Digital Elevation Model (DEM) can be used as additional information to increase the prediction accuracy of the climate maps (De Beurs, 1998).

The main purpose of this study is to find an optimal way of including elevation data of the area into the interpolation techniques to increase the prediction accuracy of the climate maps. In total, four forms of kriging and three forms of thin plate splines will be presented.

2.2 Material and Methods

2.2.1 Data sets

Two data sets are considered in this paper. The first data set consists of long-term ($\geq$ 19 years, from 1940-1990) monthly mean precipitation values at 193 measurement stations. The second set consists of 136 long-term ($\geq$ 19 years, from 1940-1990) monthly maximum temperature data. These data were ex-

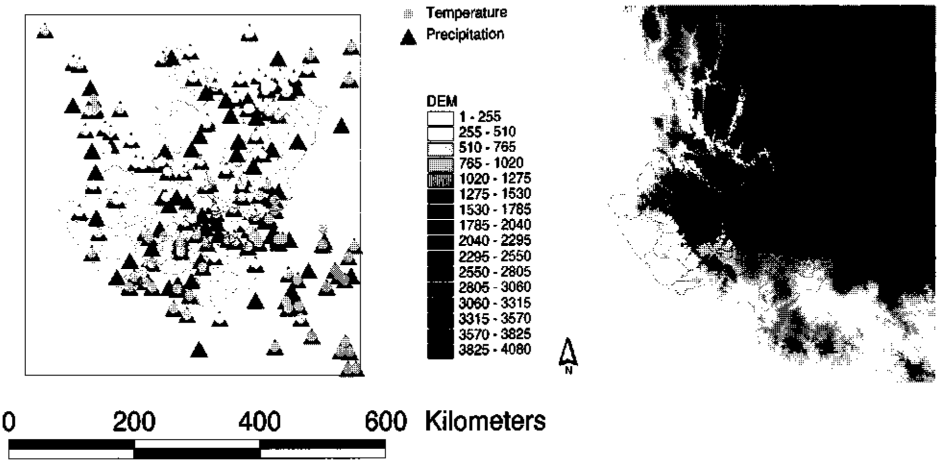


Figure 2.1: *The meteorological stations (IMTA, 1996) and a DEM (USGS, 1997) of Jalisco State, situated Northwest of Mexico-City.*

tracted from ERIC (Extractor Rápido de Información Climatológica; IMTA, 1996) for a square area (600 km x 600 km, called D), covering the state of Jalisco, Mexico. In this study, only the months April, May, August and September are considered, because for these months the correlation coefficient between elevation and precipitation is greater than 0.5. Figure 2.1 shows the measurement stations and a Digital Elevation Model (DEM) of the area.

Figure 2.2 shows scatterplots of long-term monthly maximum temperature (T_{max}) and long-term monthly mean precipitation (P_{mean}) against elevation for August. The correlation between T_{max} and elevation is -0.7 for April and May and -0.9 for August and September. For P_{mean} these values are 0.6 (April and May), -0.5 (August) and -0.6 (September). The scatterplot of P_{mean} shows statistically less attractive features.

2.2.2 Interpolation techniques

Interpolation techniques can be divided into techniques based on deterministic and stochastic models. Kriging technique is based on stochastic models while the method of thin plate splines is a deterministic interpolation technique with a local stochastic component. It is well known that under certain conditions these two interpolation techniques are equivalent to one another (Kent and Mardia, 1994). In this paper, however, at least the function for modelling the spatial correlation is chosen differently. The modelling of the trend and the neighbourhood used for prediction can differ too.

Let the actual meteorological measurements be denoted as $z(s_1)$, $z(s_2)$, ..., $z(s_n)$, where $s_i = (x_i, y_i)$ is a point in D, x_i and y_i are the coordinates of point s_i and

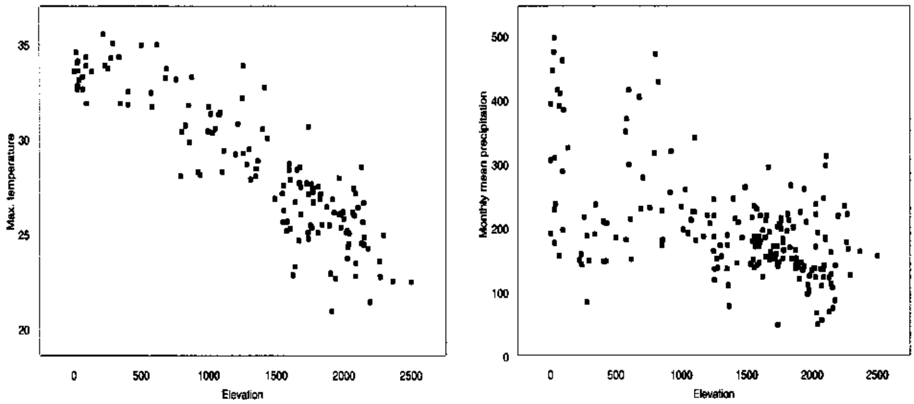


Figure 2.2: *Scatterplots for T_{max} and P_{mean} against elevation for August.*

n is equal to the number of measurement points. The elevation at a point s in the area D will be denoted as $q(s)$. A measurement considered as realization of a stochastic variable will be denoted by a lower case. Therefore, $q(s)$ can be considered as a realization of stochastic variable $Q(s)$ (i.e. outcome of a set of stochastic variables that have some spatial locations and whose dependence on each other is specified by some probabilistic mechanism).

Bivariate thin plate spline

Wahba (1990) described the theory of thin plate splines. In the case of bivariate thin plate splines, the measurements $z(s_i)$ are modelled as:

$$z(s_i) = f(s_i) + \epsilon(s_i), \qquad i = 1, ..., n \tag{2.1}$$

where f is an unknown deterministic smooth function and $\epsilon(s_i)$ are random errors. Commonly, it is assumed that $\epsilon(s_i)$ are realizations of zero mean and uncorrelated random errors.

The function f can be estimated by minimizing

$$\sum_{i=1}^{n} [z(s_i) - f(s_i)]^2 + \lambda J_2(f) \tag{2.2}$$

where $J_2(f)$ is a measure of smoothness of f, calculated by means of the following integral:

$$J_2(f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left\{ \left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right\} dxdy. \tag{2.3}$$

and λ is the so-called smoothing parameter which regulates the trade-off between the closeness of the function to the data and the smoothness of the function. The smoothing parameter λ can be estimated by generalized cross validation.

The minimization problem (2.2) is solved with $\hat{f}$ as the linear combination

$$\hat{f}(s) = \sum_{j=1}^{3} a_j \phi_j(s) + \sum_{i=1}^{n} b_i \Psi(h_i) \tag{2.4}$$

where $s = (x, y)$; ϕ_j are polynomials, $\phi_1(s) = 1, \phi_2(s) = x$ and $\phi_3(s) = y$; $\Psi(h) = h^2 \ln(h)$ and $h_i = \sqrt{(x - x_i)^2 + (y - y_i)^2}$ is the Euclidean distance between s and s_i. The coefficients b_i are restricted to satistify the boundary conditions given by:

$$\sum_{i=1}^{n} b_i \phi_j(s_i) = 0, \qquad j = 1, ..., 3 \tag{2.5}$$

The coefficients a_j and b_i in formula (2.4) can be calculated by solving a linear system of order n.

Partial thin plate spline

The bivariate thin plate spline model can be enlarged to a partial thin plate spline model by incorporating additional information (in our case elevation denoted as q) into (2.1). The measurements are modelled in the following way, on the basis of scatterplots in Figure 2.2:

$$z(s_i) = g(s_i) + \beta_1 q(s_i) + \beta_2 q^2(s_i) + \epsilon(s_i), \qquad i = 1, ..., n \tag{2.6}$$

where the function $f(s) = g(s) + \beta_1 q(s) + \beta_2 q^2(s)$ is the function to be estimated, $g(s)$ being an unknown smooth function and β_1 and β_2 are parameters with unknown value. The function g and the parameters β_1 and β_2 can be estimated by minimizing:

$$\sum_{i=1}^{n} \left[z(s_i) - g(s_i) - \beta_1 q(s_i) - \beta_2 q^2(s_i)\right]^2 + \lambda J_2(g). \tag{2.7}$$

giving the same solution structure as for bivariate thin plate splines.

Trivariate thin plate spline

Hutchinson (1998) shows that there is another way of incorporating the covariable elevation into bivariate thin plate splines. Namely, replacing the bivariate function $f(s_i)$ in (2.1) by a trivariate function $f(s_i, q_i)$:

$$z(s_i, q_i) = f(s_i, q_i) + \epsilon(s_i, q_i), \qquad i = 1, ..., n \tag{2.8}$$

where q_i is the elevation on the location s_i. The function $J_2(f)$ is enlarged by several terms (see Wahba, 1990) and the minimization problem is solved in the same way as for bivariate thin plate splines. The solution can be written as:

$$\hat{f}(s,q) = \sum_{j=1}^{4} a_j \phi_j(s,q) + \sum_{i=1}^{n} b_i \Psi(h_i) \tag{2.9}$$

where ϕ_j are polynomials, $\phi_1(s,q) = 1$, $\phi_2(s,q) = x$, $\phi_3(s,q) = y$ and $\phi_4(s,q) = q$; and $\Psi(h) = h$. The Euclidean distance h_i between (s,q) and (s_i,q_i) is calculated by $h_i = \sqrt{(x-x_i)^2 + (y-y_i)^2 + (q-q_i)^2}$. Scaling becomes important (Hutchinson, 1998), because x, y and q can be expressed in different units and the scale of variation can vary in different directions. We followed the suggestion of Hutchinson (1998) to calculate the generalized cross validation (GCV) on different scales of elevation and select the scaling with the lowest value of the GCV.

Ordinary kriging

The principles of ordinary kriging are well explained elsewhere (Isaaks and Srivastava, 1989; Cressie, 1991; Wackernagel, 1995). The measurements are modelled in the following way:

$$z(s_i) = f(s_i) + \epsilon(s_i), \qquad i = 1, 2, ..., n \tag{2.10}$$

where, in this case, $f(s_i)$ are considered as realizations of a random function F in point s_i, which may contain a deterministic function $\mu(s) = \mathrm{E}\{F(s)\}$ to model possible trends; $\epsilon(s_i)$ are realizations of zero mean and uncorrelated random errors. The trend $\mu(s)$ is assumed to be equal to an unknown constant μ.

The spatial correlation between the measurement points can be quantified by means of the semivariance function:

$$\gamma(s,h) = \frac{1}{2}\mathrm{var}[Z(s) - Z(s+hu)] \tag{2.11}$$

where h is the Euclidean distance between two points and u is a vector of unit distance ($\|u\| = 1$) and γ is independent of u (isotropy). Assume that the trend is constant and $\gamma(s,h)$ is independent of s. A parametric function is used to model the semivariance for different values of h. In this research, the spherical model - c Sph(a) - is used

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ c\left\{\frac{3}{2}\left(\frac{h}{a}\right) - \frac{1}{2}\left(\frac{h}{a}\right)^3\right\}, & 0 < h \leq a \\ c, & h > a \end{cases} \tag{2.12}$$

where c is the scale parameter of the semivariance function and a is a parameter which determines the so-called range of spatial dependence. The random errors (and/or the spatial nugget random function) have a variance c_0. For the stochastic variable Z the following semivariance function is used: c_0 Nug(0) + c Sph(a).

The interpolated value at an arbitrary point s_0 in D is the realization of the (locally) best linear unbiased predictor of $F(s_0)$ and can be written as weighted sum of the measurements

$$\hat{f}(s_0) = \sum_{i=1}^{n} w_i z(s_i) \tag{2.13}$$

where the weights w_i are derived from the kriging equations by means of the semivariance function; n is the number of measurement points within a radius from point s_0 (in this study we have taken a radius with an Euclidean distance of 240 km). The parameters of the semivariance function and the nugget effect can be estimated by the empirical semivariance function. An unbiased estimator for the semivariance function in point h is half the average squared difference between paired data values.

$$\hat{\gamma}(h) = \frac{1}{2n(h)} \sum\nolimits' [z(s_i) - z(s_j)]^2 \tag{2.14}$$

where the prime in $\sum'$ means that it is summed over all (i, j) for which the Euclidean distance $\|s_i - s_j\|$ is equal to h; the number of pairs with this property is denoted by $n(h)$.

Ordinary cokriging

Cokriging makes use of different variables, modelled as realizations of stochastic variables. In this study, elevation - Q(s) - of the area D is used as covariable to predict values of T_{max} and P_{mean}. The spatial dependence is characterized by two semivariance functions $\gamma_{zz}(s, h), \gamma_{qq}(s, h)$ and the cross-semivariance function:

$$\gamma_{zq}(s, h) = \frac{1}{2} \mathrm{E}\left\{[Z(s) - Z(s + hu)][Q(s) - Q(s + hu)]\right\} \tag{2.15}$$

where u is a vector of unit distance ($\|u\| = 1$) and γ_{zq} is independent of u (isotropy).

To ensure that the variance of any possible linear combination of the two stochastic variables is positive, a so-called linear model of coregionalization is applied. This model implies that each semivariance and cross-semivariance function must be modelled by the same linear combination of semivariance functions (Isaaks and Srivastava, 1989). Furthermore, the matrix of coregionalization should be positive semi-definite. A nested semivariance function is used with a nugget and two spherical semivariance functions with different ranges. The cross-semivariance function can be estimated by the empirical cross-semivariance function

$$\hat{\gamma}_{zq}(h) = \frac{1}{2n(h)} \sum\nolimits' [z(s_i) - z(s_j)][q(s_i) - q(s_j)] \tag{2.16}$$

where $n(h)$ is the number of data pairs where both variables are measured at an Euclidean distance h.

The interpolated value at an arbitrary point s_0 in D is the realization of the (locally) best linear unbiased predictor of $F(s_0)$ and can be written as weighted sum of the measurements:

$$\hat{f}(s_0) = \sum_{i=1}^{m_1} w_{1i} z(s_i) + \sum_{j=1}^{m_2} w_{2j} q(s_j) \tag{2.17}$$

where m_1 is the number of measurements of $Z(s)$ taken within a radius (of 240 km) from s_0 (out of the modelling data set), m_2 is the number of meteorological stations within a radius of 240 km from s_0 (out of the modelling and validation set). The weights w_{1i} and w_{2j} can be determined using the two semivariance functions and the cross-semivariance function.

Regression-kriging

Odeh *et al.* (1995) compared, among other techniques, three forms of regression-kriging (comparable with kriging with external drift). The idea of regression-kriging, in this paper, is that we characterize the trend component $\mu(s)$ of the model for the random function $F(s)$ as an unknown linear combination of known functions (regression model). In ordinary kriging the trend component is modelled as constant; in the usual form of universal kriging the trend component is modelled as a polynomial of a certain degree. In our application the trend is modelled as:

$$\alpha + \beta_1 q(s) + \beta_2 q^2(s) \tag{2.18}$$

where $q(s)$ is in our case a realization of the elevation now used in a regression equation.

The interpolated value at location s_0 can be calculated by a linear combination of the regression model and a weighted sum (ordinary kriging) of regression residuals $z^*(s_i) = z(s_i) - \hat{\alpha} - \hat{\beta}_1 q(s_i) - \hat{\beta}_2 q^2(s_i)$.

$$\hat{f}(s_0) = \hat{\alpha} + \hat{\beta}_1 q(s_0) + \hat{\beta}_2 q^2(s_0) + \sum_{i=1}^{n} w_i z^*(s_i) \tag{2.19}$$

The difficulty of this form of regression-kriging, and of universal kriging in general, is that the parameters of the regression model and the parameters of the semivariance function of the spatial correlated regression residuals should be estimated simultaneously (Laslett and McBratney, 1990). Under the assumption of normality, the parameters can be estimated by restricted maximum likelihood (REML), which is one of the techniques to estimate the parameters of the regression model and the parameters of the semivariance function simultaneously (Gotway and Hartford, 1996).

Trivariate regression-kriging

Finally, trivariate regression-kriging will be introduced. Trivariate regression-kriging is a form of regression-kriging, where trivariate ordinary kriging is applied on the regression residuals. The trend is chosen equally as in trivariate thin plate splines. The interpolated value at a location s_0 can be calculated by:

$$\hat{f}(s_0) = \hat{\alpha} + \hat{\beta}_1 x + \hat{\beta}_2 y + \hat{\beta}_3 q + \sum_{i=1}^{n} w_i z^*(s_i) \tag{2.20}$$

The weights w_i are determined by the semivariance function, which is a function of the Euclidean distance between two points (s_i, q_i) and (s, q). The units are of different order and scaling becomes important, the same scaling is used as for trivariate thin plate splines. In this case, REML is not applied because of limitations of the software used. The residual semivariance function is now estimated from the OLS regression residuals.

2.2.3 Comparison of interpolation techniques

To compare the interpolation techniques, the original data set is divided into a modelling data set and a validation set of 25 measurement points. The 25 points are not chosen randomly, but are selected by the authors, so that the area is still reasonably covered by measurement points. Five validation sets are chosen from each data set. Each validation set contains different measurement points from the original data sets. Predictions on the locations of the validation points - $\hat{z}(s_i)$ - and the measured values at these locations - $z(s_i)$ - are compared by two criteria: the Mean Square Error (MSE) and the Maximal Prediction Error (MPE).

$$\mathrm{MPE} = \max_{i=1,\ldots,n_v} \mid \hat{z}(s_i) - z(s_i) \mid \tag{2.21}$$

$$\mathrm{MSE} = \frac{1}{n_v} \sum_{i=1}^{n_v} [\hat{z}(s_i) - z(s_i)]^2 \tag{2.22}$$

where $n_v (= 25)$ is the number of validation points.

2.3 Results

The automatic calculation procedure of thin plate splines allows a straightforward analysis of these techniques. There is no need for any prior estimation of the spatial dependence of measurement points. ANUSPLIN (Hutchinson, 1997) is used to perform the analyses. For trivariate thin plate splines it is useful to optimize the elevation scale (Hutchinson, 1998). Therefore the square root generalized cross validation for trivariate thin plate splines is determined at different scales of elevation (metre, decameter, hectometre and kilometre). Decameter appears to be the optimal scaling for T_{max} and kilometre for P_{mean}.

This way of scaling is found to be sufficient, because no major differences between the GCV of two successive scales are found. Trivariate regression-kriging is applied with the same scaling as trivariate thin plate splines.

The semivariance functions for ordinary kriging are estimated by weighted least squares with GSTAT (Pebesma and Wesseling, 1998). For cokriging the semivariance functions, by means of the linear model of coregionalization, are estimated by COREG (Bogaert *et al.*, 1995). The residual semivariance function for trivariate regression-kriging is estimated in a relatively simple way. First the trend is estimated by ordinary least squares (OLS), followed by the estimation of the spatial variability of the regression residuals. For regression-kriging, where the semivariance function depends only on x and y, the parameters of the regression model and the parameters of the semivariance function are estimated simultaneously by the REML option of PROC MIXED in SAS (Littell *et al.*, 1996). Figures 2.3 and 2.4 show some examples of fitted semivariance functions (with models and parameter values) for T_{max} and P_{mean} for cokriging and trivariate regression-kriging.

Only the recorded elevations of the meteorological stations are used for interpolation with ordinary cokriging. We used a DEM of the area but prediction accuracy increased substantially as just the recorded elevation at the point to be predicted (validation point) was available, as for all other interpolation techniques. Tables 2.1 and 2.2 show the results of all 7 interpolation techniques for 5 validation sets for T_{max} and P_{mean}, respectively.

The results of Table 2.1 and 2.2 demonstrate the benefit of using the covariable elevation. Especially for T_{max} the differences of interpolation with elevation and without elevation are convincing. This is due to the high correlation between elevation and T_{max}. Comparing regression-kriging, cokriging, trivariate regression-kriging, trivariate thin plate splines and partial thin plate spline for T_{max} shows an advantage for the two interpolation techniques which made use of three-dimensional coordinates (trivariate). The differences between the results of the interpolation techniques are less clear for P_{mean}. Only for validation sets 1 and 2 did the trivariate interpolation techniques perform more accurately with respect to the MSE.

The prediction results of P_{mean} for validation set 1 for August and September are very poor for bivariate thin plate splines and partial thin plate splines. This is mainly caused by two validation points in the South-East of the area, which have a large prediction error. Probably, this is caused by a local trend at adjacent measurement stations.

2.4 Discussion

In this paper 7 interpolation techniques are discussed, 5 including and 2 excluding elevation as additional information. The two techniques excluding elevation

Table 2.1: *Results of the Mean Square Error (MSE) and the Maximum Prediction Error (MPE) for 5 validation sets (***v1-v5***) of long-term monthly maximum temperature (T_{max}). The values with the lowest MSE and MPE of the 7 interpolation techniques are written in italic.*

Interpolation technique		MSE				MPE			
		apr	may	aug	sep	apr	may	aug	sep
ordinary kriging	**v1**	10.3	9.8	8.2	8.4	6.1	5.8	5.2	5.4
regression-kriging		5.0	4.9	4.0	3.4	5.0	4.9	4.2	4.2
cokriging		5.7	5.5	3.9	4.1	5.5	5.6	4.3	4.5
trivariate regression-kriging		*4.1*	*4.0*	*2.9*	*2.9*	4.6	4.9	*3.8*	*3.7*
bivariate thin plate splines		10.5	9.9	8.8	8.3	6.1	5.7	5.8	5.2
trivariate thin plate splines		4.8	4.5	3.1	3.0	*4.5*	*4.4*	4.3	4.1
partial thin plate splines		4.8	5.2	3.4	3.7	5.1	5.2	3.9	4.4
ordinary kriging	**v2**	7.2	5.9	3.9	3.9	6.9	6.7	5.3	5.6
regression-kriging		4.5	4.2	2.0	2.0	5.9	5.7	3.3	3.7
cokriging		4.3	3.7	1.9	*1.8*	5.5	5.4	*3.0*	*2.9*
trivariate regression-kriging		*3.5*	*3.5*	2.1	2.0	*4.9*	*4.7*	3.3	3.2
bivariate thin plate splines		8.4	7.3	4.9	5.0	6.6	6.4	5.6	6.0
trivariate thin plate splines		3.8	4.0	*1.8*	*1.8*	5.4	5.5	3.5	3.4
partial thin plate splines		4.1	3.8	2.2	2.1	5.3	5.0	3.7	3.7
ordinary kriging	**v3**	5.9	5.6	5.0	5.2	5.7	*5.1*	5.4	5.3
regression-kriging		5.0	4.3	*1.7*	*1.8*	5.3	5.2	3.7	*3.9*
cokriging		6.1	5.8	2.0	2.1	5.9	6.0	4.4	4.5
trivariate regression-kriging		*4.5*	*4.0*	1.8	1.9	*5.2*	5.3	3.8	4.0
bivariate thin plate splines		7.5	7.3	5.0	5.2	6.6	5.9	5.2	5.1
trivariate thin plate splines		5.8	5.3	2.1	2.3	5.3	6.1	4.3	4.4
partial thin plate splines		7.1	6.2	*1.7*	2.0	6.4	6.2	*3.5*	4.5
ordinary kriging	**v4**	7.4	7.2	5.0	5.4	6.4	6.6	6.4	6.8
regression-kriging		3.9	4.2	2.1	1.9	*3.7*	*4.0*	4.1	4.2
cokriging		5.1	5.0	2.3	2.6	5.6	5.9	5.1	5.6
trivariate regression-kriging		*3.4*	*3.7*	*1.5*	*1.8*	3.9	*4.0*	2.9	3.5
bivariate thin plate splines		7.5	7.6	5.8	6.1	6.5	6.6	6.4	6.8
trivariate thin plate splines		3.8	*3.7*	1.6	*1.8*	4.3	4.2	*2.7*	*3.1*
partial thin plate splines		3.9	4.1	2.0	2.2	*3.7*	4.1	3.9	4.4
ordinary kriging	**v5**	6.8	6.4	5.7	5.4	6.8	7.0	5.7	5.5
regression-kriging		3.8	2.4	1.2	1.1	4.4	3.8	2.8	2.4
cokriging		4.6	4.3	2.0	1.7	5.5	5.0	3.3	2.8
trivariate regression-kriging		*2.0*	*2.0*	1.3	1.2	*3.5*	3.6	*2.2*	*2.1*
bivariate thin plate splines		7.4	7.1	8.3	7.5	6.7	7.0	6.8	6.2
trivariate thin plate splines		2.6	3.6	1.8	1.7	5.3	5.6	2.8	2.6
partial thin plate splines		2.6	2.3	*1.1*	*1.0*	*3.5*	*3.4*	2.8	2.7

Table 2.2: *Results of the Mean Square Error (MSE) and the Maximum Prediction Error (MPE) for 5 validation sets (***v1-v5***) for long-term monthly mean precipitation (P_{mean}). The values with the lowest MSE and MPE of the 7 interpolation techniques are written in italic.*

Interpolation technique	MSE				MPE			
	apr	may	aug	sep	apr	may	aug	sep
ordinary kriging **v1**	36.7	62.4	1251.9	1123.6	*26.2*	*22.4*	84.8	80.6
regression-kriging	36.1	64.4	1182.1	1111.2	26.7	22.7	89.2	78.7
cokriging	35.3	58.1	1132.8	820.0	26.4	23.4	77.3	*59.5*
trivariate regression-kriging	*33.2*	*50.6*	*728.4*	520.1	26.4	22.6	62.2	66.5
bivariate thin plate splines	44.1	78.5	3562.6	3418.2	*26.2*	22.9	172.5	195.3
trivariate thin plate splines	33.7	50.8	753.6	*515.9*	26.4	*22.4*	*58.9*	63.6
partial thin plate splines	36.9	67.4	3456.6	3101.4	26.9	22.9	183.6	169.5
ordinary kriging **v2**	17.5	*158.5*	4715.1	4261.5	14.9	44.9	256.2	208.4
regression-kriging	13.7	168.0	4712.9	4175.3	10.9	43.9	252.4	203.5
cokriging	19.3	172.3	4687.3	4062.4	13.6	*42.5*	249.3	203.1
trivariate regression-kriging	17.6	170.7	*4479.5*	*3554.5*	14.0	44.8	260.6	194.4
bivariate thin plate splines	13.8	170.2	4802.2	3834.3	12.8	46.0	250.3	201.4
trivariate thin plate splines	*12.0*	168.7	4492.2	3623.4	*10.5*	44.7	255.0	*191.4*
partial thin plate splines	13.3	166.8	5319.8	4037.6	10.9	43.8	*248.7*	198.6
ordinary kriging **v3**	7.8	55.4	1804.2	1206.7	6.8	17.6	93.4	114.6
regression-kriging	3.5	51.9	1683.5	*971.0*	4.2	17.5	*74.9*	*104.8*
cokriging	5.8	*48.3*	*1652.7*	1012.4	5.7	19.7	82.6	115.1
trivariate regression-kriging	6.7	49.3	1787.4	1252.6	5.6	17.8	95.2	138.8
bivariate thin plate splines	6.4	55.2	1762.8	1264.2	6.2	19.6	88.1	138.2
trivariate thin plate splines	*3.0*	48.5	1665.3	1230.0	4.3	17.8	90.0	145.9
partial thin plate splines	3.4	52.5	1720.1	1397.2	*3.9*	*17.1*	87.9	140.7
ordinary kriging **v4**	17.4	60.3	2477.1	*3150.5*	14.3	25.4	165.6	138.0
regression-kriging	9.5	44.9	*2467.8*	3321.0	11.2	18.7	*162.0*	*129.6*
cokriging	13.6	59.6	2537.4	3517.8	*10.0*	18.5	174.9	148.9
trivariate regression-kriging	13.7	47.0	3252.8	4038.0	12.2	17.3	200.1	195.1
bivariate thin plate splines	15.9	71.6	3069.9	4029.3	14.0	25.7	190.0	157.0
trivariate thin plate splines	9.5	45.1	3696.9	4137.1	11.1	*16.4*	217.3	196.6
partial thin plate splines	*9.1*	*42.4*	2858.4	3767.8	11.4	18.4	182.3	139.4
ordinary kriging **v5**	8.6	61.5	2193.9	2873.6	8.6	*17.2*	161.1	164.2
regression-kriging	*6.8*	68.7	1891.4	2291.8	8.3	24.4	135.2	135.1
cokriging	12.2	91.1	2276.2	2538.7	10.0	27.7	153.8	148.0
trivariate regression-kriging	8.4	*58.5*	2170.4	2723.6	8.3	23.0	149.0	132.5
bivariate thin plate splines	7.6	84.3	*1607.1*	*1984.9*	*7.0*	21.3	*83.3*	132.1
trivariate thin plate splines	*6.8*	64.2	2419.2	2923.4	8.7	25.5	136.6	131.8
partial thin plate splines	*6.8*	67.7	1724.2	2050.3	7.9	24.5	88.2	*129.8*

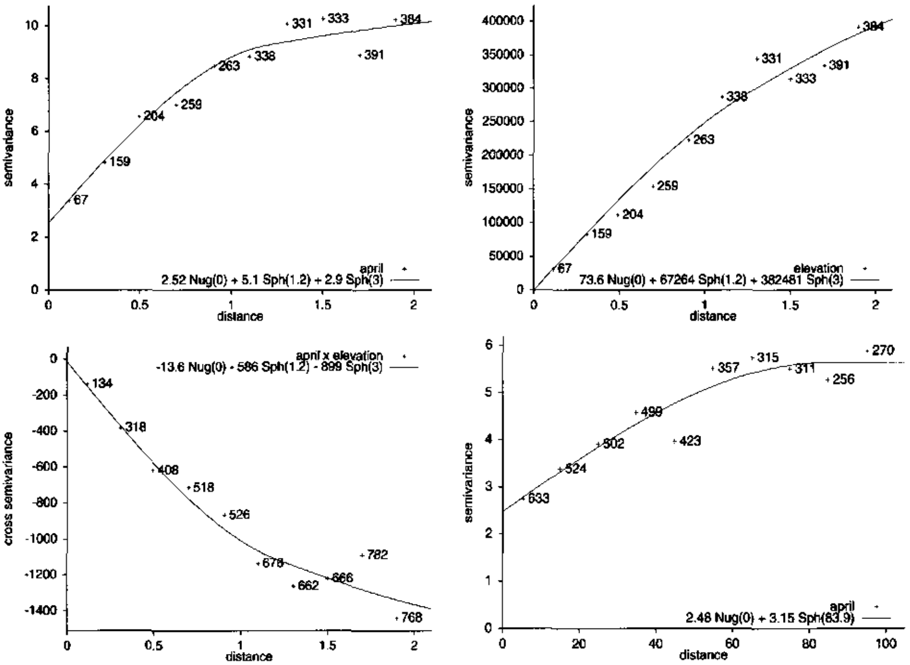


Figure 2.3: *Estimated semivariance functions for cokriging and trivariate regression-kriging for T_{max}, April, validation set 1. Upper left: T_{max}; upper right: elevation; lower left: cross semivariance function between T_{max} and elevation; lower right: residual semivariance function T_{max} for trivariate regression-kriging.*

perform, especially for T_{max} considerably less accurately than the techniques including elevation. The MSE and MPE values are lower when elevation is used as additional information for prediction, especially when the correlation between the two variables is high. From the techniques which include elevation, trivariate thin plate splines and trivariate regression-kriging seem to perform best. Cokriging appears to be the most time-consuming interpolation technique to implement. Therefore, in this case study, cokriging seems not preferable. The main reason for cokriging having relatively poor prediction results is the fact that in that case a linear relation is assumed between climate variable and elevation. The other techniques (including elevation) used in this paper, do not assume such relation because regression models with more regressors and trivariate techniques are used.

The results of T_{max} are much clearer to interpret than the results of P_{mean}. There is much more variability in the P_{mean} predictions resulting from a higher variability in the data. The correlation between the climate variable and elevation is lower, which causes smaller differences between including and excluding

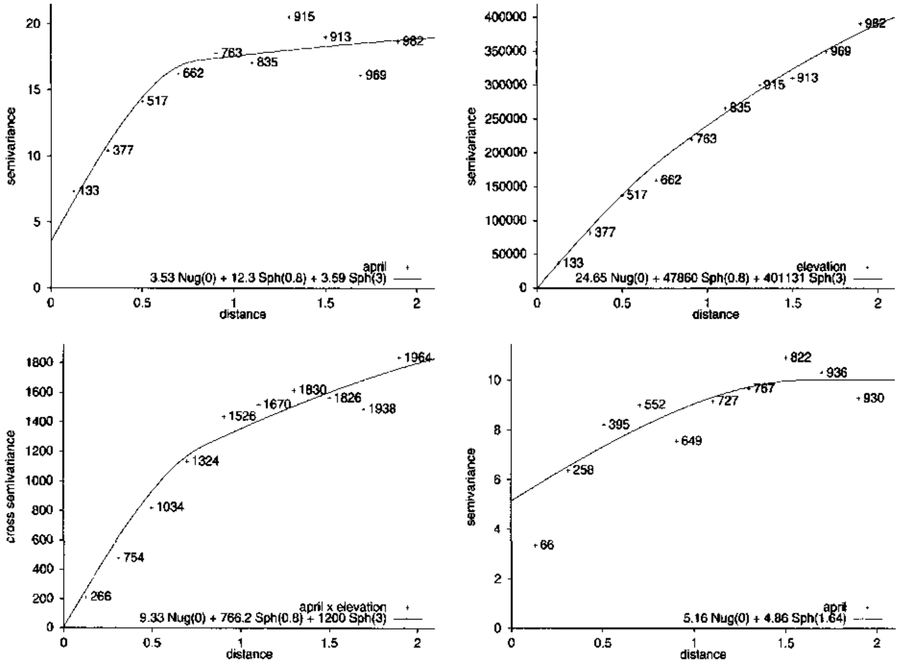


Figure 2.4: *Estimated semivariance functions for cokriging and trivariate regression-kriging for P_{mean}, April, validation set 1. Upper left: P_{mean}; upper right: elevation; lower left: cross semivariance function between P_{mean} and elevation; lower right: residual semivariance function P_{mean} for trivariate regression-kriging.*

elevation. In general, precipitation data are clearly non-Gaussian. Although a transformation can be considered, this has been reported to have disadvantages for local estimation (Roth, 1998).

Beek *et al.* (1992) stress the importance of interpolation techniques for crop growth simulation. In this paper more advanced forms of kriging and thin plate splines are applied. Especially, the trivariate forms of kriging and thin plate splines performed well. The main advantage of thin plate splines over kriging is the operational simplicity of this technique, which can be very important from a practical point of view. The kriging procedure requires more effort and experience. For T_{max} the predictions results of trivariate regression-kriging are slightly more accurate compared to the results of trivariate thin plate splines, but the differences are small. Within the geostatistical framework trivariate regression-kriging, as described in this paper, seems to be an attractive option.

Acknowledgements

The authors thank A.C. van Eijnsbergen and A. Stein for fruitful discussions at various stages of this work.

Chapter 3

Global optimization problems in optimal design of experiments in regression models

E.P.J. Boer and E.M.T. Hendrix

Journal of Global Optimization (2000) **18:** *385-398*

In this paper we show that optimal design of experiments, a specific topic in statistics, constitutes a challenging application field for global optimization. This paper shows how various structures in optimal design of experiments problems determine the structure of corresponding challenging global optimization problems. Three different kinds of experimental designs are discussed: discrete designs, exact designs and replicationfree designs. Finding optimal designs for these three concepts involves different optimization problems.

3.1 Introduction

In many fields of sciences, experiments are done in order to estimate parameters of regression models. Optimal experimental designs can be used to maximize the precision of the least squares estimator, given the total number of observations. The theory of optimal experimental design has been explained (among others) in the monographs of Fedorov (1972), Silvey (1980) and Pukelsheim (1993). Atkinson and Donev (1992), Atkinson (1996) and Müller (1998) show the usefulness of optimal experimental designs in a more practical setting. Given the total number of observations, the optimal design is determined by the design space (experimental region), the regression model and the optimality criterion. Searching for these optimal designs yields challenging optimization problems (Zhigljavsky, 1991), which has resulted in a large number of publications (among others: Welch, 1982; Gaffke and Mathar, 1992; Jones and Wang, 1999). In this paper it is shown how general and more specific properties of experimental design problems result in properties of optimization problems for three different kinds of experimental designs. Important properties of optimal experimental designs are discussed and it is indicated how these properties can be helpful by solving the optimization problems for finding the optimal design.

This paper considers optimal experimental design in the context of regression models. Let

$$Y_i = \eta(x_i, \theta) + E_i, \quad x_i \in X \subset \mathbb{R}^k \tag{3.1}$$

be a (statistical) regression model with a regression function η and i.i.d. zero-mean error terms E_i. The unknown θ is a parameter vector with m elements, $\theta^T = (\theta_1, \ldots, \theta_m) \in \Omega \subset \mathbb{R}^m$. Further we assume that η is a twice differentiable continuous function.

3.2 Theory of optimal design of experiments

A concept of an experimental design in regression analysis, frequently used in literature, is that of a so-called *discrete design*. A discrete design ϵ is written as:

$$\epsilon = \begin{pmatrix} x_1 & x_2 & \ldots & x_r \\ p_1 & p_2 & \ldots & p_r \end{pmatrix} \tag{3.2}$$

where p_i indicates measurement weight at support point x_i, $i = 1, 2, \ldots, r$, $r \geq m$. The weights sum to unity: $\sum_{i=1}^{r} p_i = 1$, $p_i \geq 0$. The support points are chosen from the design space X; $x_i \in X$. The design space X may have dimension ≥ 1, which means that also spatial problems could be considered (Müller, 1998). From an optimization point of view, for a given number of support points r, we would like to find the best values for p_i and x_i (in a sense to be specified). Notice however, that in some situations this number r is not known beforehand.

A more practical definition of an experimental design is that of a normalized exact design. In this design, for all p_i holds that p_iN is integer, where N is the maximum number of observations allowed in the experiment. An *exact design* (not normalized) $\epsilon(N)$ is usually written as follows:

$$\epsilon(N) = \begin{pmatrix} x_1 & x_2 & \dots & x_r \\ n_1 & n_2 & \dots & n_r \end{pmatrix} \tag{3.3}$$

where n_i is the number of replications at each support point, $\sum_{i=1}^{r} n_i = N$. An exact design becomes discrete by using $p_i = n_i/N$. It is worth noticing, that searching for an exact design (in practice all designs are exact) results in a mixed continuous/integer optimization problem. These problems are in general hard to solve.

In spatial problems, but also in other problems, observing in a point of the design space is often restricted to a certain number of replications. If the number of replications is restricted to one (*replicationfree design*), the observations have a minimal distance between each other. In this case, the design space X is often (see e.g. Fedorov, 1989) considered as (a grid of) Q candidate points or possible measurement points (observations). The design problem becomes a combinatorial problem of selecting N observations from Q candidate points. The solution of the problem will give an exact design with only one replication at each support point ($N = r$).

Optimality of a design depends on the function $\eta(x, \theta)$ with parameter vector θ, under consideration. In many cases, research focuses on models which are linear in the parameters; then $\eta(x, \theta)$ can be written as $\theta^T f(x)$. Moreover performance depends on a specific criterion which is a function of the so-called information matrix. Experiments which contain a lot of information enlarge the precision of the estimation of the parameters of the model. If a linear model with the usual regression assumptions of independent errors and constant variance is studied, the information matrix for a discrete design is given by:

$$M[\epsilon] = \sum_{i=1}^{r} p_i f(x_i) f^T(x_i) \tag{3.4}$$

The inverse of the information matrix (M^{-1}, for exact designs the variance-covariance matrix of $\hat{\theta}$) is helpful to represent the variance of the predictor $\eta(x, \hat{\theta})$ on the design space X by means of the standardized variance function. The standardized variance function (under the usual statistical assumptions) is defined as follows:

$$d(x, \epsilon) = \mathrm{var}[\eta(x, \hat{\theta})] = f^T(x) M^{-1} f(x) \tag{3.5}$$

This standardized variance function makes the design problem easier to understand from a graphical point of view. Note that for linear models the standardized variance function is independent of θ, because f does not depend on the parameter vector θ. Figure 3.1 shows the standardized variance functions of two designs ($N = 3$) for a simple linear model, $\eta(x, \theta) = \theta_1 + \theta_2 x$, $x \in [-1, 1]$. The designs are chosen as follows:

$$\epsilon_1 = \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}, \qquad \epsilon_2 = \begin{pmatrix} -1 & 1 \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

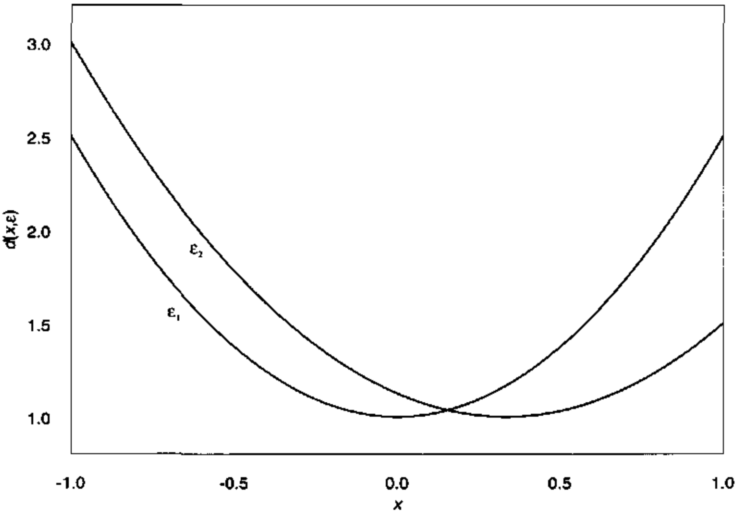


Figure 3.1: *Standardized variance functions for the two normalized exact designs* ϵ_1 *and* ϵ_2 *for* $N = 3$.

The two designs are almost the same, only the measurement at $x = 0$ in design ϵ_1 is moved to the right end of the interval for design ϵ_2. Figure 3.1 shows the result of this movement, as the standardized variance at the right end of the interval is lowered. The maxima of $d(x, \epsilon)$ are found, for any design, at the extreme points of the design space. In Section 3.2.2 we will come back to this.

The theory of optimal experimental designs can be extended to nonlinear models by considering the Taylor series expansion (Atkinson and Donev, 1992). In this case, for $f^T(x)$ the vector of partial derivatives is used

$$f^T(x) = \begin{bmatrix} \frac{\partial \eta(x,\theta)}{\partial \theta_1} & \frac{\partial \eta(x,\theta)}{\partial \theta_2} & \cdots & \frac{\partial \eta(x,\theta)}{\partial \theta_m} \end{bmatrix} \tag{3.6}$$

Optimal designs for nonlinear models are called *locally optimal designs* because $f(x)$ depends on values of θ, so local with respect to parameter values of θ. This is confusing given another interpretation of the terminology of locally optimal solutions in global optimization. The information matrix in the nonlinear case is denoted by $M[\theta, \epsilon]$.

3.2.1 Criteria

Many criteria for optimal designs are functions of the information matrix, say $\phi(M[\epsilon])$. The most popular criterion is *D-optimality*, which minimizes the generalized variance of the parameter estimates. This corresponds to minimizing the value of the determinant of the variance-covariance matrix $M^{-1}[\epsilon]$. When interest is focused on estimation of a subset of elements of θ, the criterion is written as D_s. A design which minimizes the maximum of the standardized variance function over the design region X is called a *G-optimal* design. The criteria are determined by minimizing the following functions:

$$\text{D-optimality:}\quad \det(M^{-1}[\epsilon]) \tag{3.7}$$

$$\text{D}_\text{s}\text{-optimality:}\quad \det(M^{11}[\epsilon]) \tag{3.8}$$

$$\text{G-optimality:}\quad \max_{x \in X} d(x, \epsilon) \tag{3.9}$$

where $M^{11}[\epsilon]$ is the s x s submatrix of $M^{-1}[\epsilon]$ with rows and columns corresponding to the s selected elements of θ. The exact designs ($N = 3$) obtained from ϵ_1 and ϵ_2 presented in Figure 3.1 are G-optimal and D-optimal designs respectively.

3.2.2 Properties of optimal design problems

The optimization problem for a discrete design can be considered as choosing the best x_i and p_i. A problem in using general purpose optimization methods is that the number of support points is not known beforehand (Jones and Wang, 1999). The optimal design problem becomes even more difficult when an exact design is needed, which results into a mixed continuous/integer optimization problem. Boer *et al.* (2000) show that this problem can not be solved very easily, because of local optima. Some important properties (theorems) from the optimal experimental design theory can assist in solving the optimization problems. We will present them without going very much into detail giving the reader a flavour of the existing theory. The properties will be ordered from general properties of models and criteria to more specific cases.

- The criterion function $\phi(M[\epsilon])$ has certain properties that capture the idea of whether the information included in matrix M is large or small. If a design ϵ^* is better than ϵ, the information included in matrix $M[\epsilon^*]$ is considered larger than that in $M[\epsilon]$ in a certain ordering. A reasonable criterion of the information matrix is that the value of a criterion function is non-decreasing (monotonic) when measurements are removed from an existing design. An application of this property in a branch-and-bound algorithm for optimal replicationfree designs is discussed in Section 3.3.3.

- Although, the number of support points r of a design is variable in the optimization problem considered here, certain bounds can be given. These bounds can be derived by looking at some basic properties of the information matrix M (Fedorov, 1972):

1. For any design ϵ: $M[\epsilon]$ is positive-semidefinite.
2. If $r < m$ then $\det(M[\epsilon]) = 0$, i.e. $M^{-1}[\epsilon]$ does not exist.
3. For any compact design space X, the set $\{M[\epsilon];\ \epsilon \text{ is discrete }\}$ is convex.

From property 2 it is clear that the number of support points should at least be equal to the number of parameters ($r \geq m$). Property 3 leads together with Carathéodory's Theorem (see Silvey, appendix 2, 1980) to an upper bound for the minimum number of support points. This upper bound is equal to $\frac{1}{2}m(m+1)+1$. For D-optimality this can be strengthened to $\frac{1}{2}m(m+1)$. Thus, for certain criteria $\phi(M)$ there exists an optimal design with at least m and at most $\frac{1}{2}m(m+1)+1$ support points. These bounds for the number of support points are especially useful for general purpose optimization (Section 3.3.4).

- The most celebrated theorem in optimal design of experiments is undoubtedly the Equivalence Theorem of Kiefer & Wolfowitz (1960). This theorem states that the following characterizations of an optimal discrete design ϵ^* are equivalent.

$$(i)\ \text{design } \epsilon^* \text{ is D-optimal} \tag{3.10}$$

$$(ii)\ \epsilon^* \text{ minimizes } \max_x d(x,\epsilon) \text{ or } \epsilon^* \text{ is G-optimal} \tag{3.11}$$

$$(iii)\ \max_x d(x,\epsilon^*) = m \tag{3.12}$$

For a discussion and proof, see (among others) Silvey (1980). This theorem was first derived by Kiefer and Wolfowitz (1960) for linear models, but White (1973) showed that it can be extended to nonlinear models. Note that this theorem holds for discrete designs, not for all exact designs. Figure 3.1 shows an example of (normalized) exact designs where the D- and G-optimal designs are different.

This theorem gives the opportunity to calculate D-optimal discrete designs by means of properties of G-optimal (discrete) designs. For G-optimality the maximum of the standardized variance matrix is minimized. It is known - see (*iii*) of the Equivalence theorem - that as long as this maximum is larger than m, the design is not G-optimal and thus not D-optimal. By putting (additional) weight at point x^* where the maximum of $d(x,\epsilon)$ is reached, the standardized variance at point x^* can be lowered (see Figure 3.1). This concept is used in the development of an algorithm (Fedorov, 1972), which will be elaborated further in Section 3.3.2.

- Optimal *exact* designs are often difficult to calculate because the number of replications at each support point should be integer. In the special case of $r = m$ and D-optimality, the number of replications should be chosen as equal as possible. Rasch (1990) shows this by rewriting the information matrix $M[\theta, \epsilon]$ in the following matrix notation:

$$M[\theta,\epsilon] = G^T(\theta,\epsilon(N))\ \mathcal{N}\ G(\theta,\epsilon(N)) \tag{3.13}$$

where

$$G^T(\theta,\epsilon(N)) = \begin{bmatrix} f(x_1) & f(x_2) & \ldots & f(x_r) \end{bmatrix}$$

and

$$\mathcal{N} = \mathrm{diag}(n_1, n_2, \ldots, n_r)$$

Minimizing $\det(M^{-1}[\epsilon])$ means maximizing $\det(M[\epsilon])$. Now

$$\left|G^T(\theta,\epsilon(N))\ \mathcal{N}\ G(\theta,\epsilon(N))\right| = \left|G^T(\theta,\epsilon(N))\right|\ |\mathcal{N}|\ |G(\theta,\epsilon(N))| \tag{3.14}$$

G is independent of n_i and $|\mathcal{N}| = \prod n_i$ is maximized when the values of n_i are as equal as possible.

- Figure 3.1 illustrates the standardized variance functions for two designs for the easy case of simple linear regression. The maxima of the standardized variance functions can be found in the extreme points of the design space, due to the convexity of these functions. As long as $f(x)$ is linear in x, the standardized variance function $d(x,\epsilon) = f^T(x)M^{-1}f(x)$ is a convex quadratic function. If $f(x)$ is nonlinear in x, the resulting standardized variance function is not quadratic. Consider the following quasi-linear (linear in parameters) model:

$$\eta(x,\theta) = \theta_1 + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_1 x_2 \tag{3.15}$$

where x_1 and x_2 can be chosen from $X = [-1,1]^2 \subset \mathbb{R}^2$. A D-optimal design for this function is equal to:

$$\epsilon_3 = \begin{pmatrix} 1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix} \tag{3.16}$$

where the first row contains the coordinates of x_1 and the second row the coordinates of x_2. It can be shown that this design is D-optimal by calculating the standardized variance function for design ϵ_3.

$$d(x,\epsilon_3) = 1 + {x_1}^2 + {x_2}^2 + {x_1}^2{x_2}^2 \tag{3.17}$$

A plot of this standardized variance function on a unit square is given in Figure 3.2. Note that this design is indeed D-optimal, because the maximum on the unit square is equal to the number of parameters (see Equation (3.12); Müller, 1998).

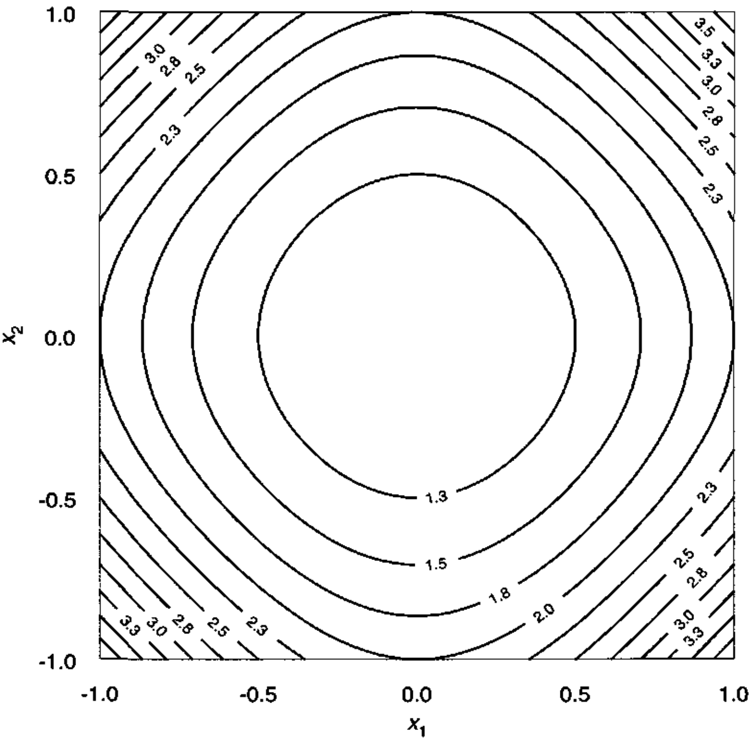


Figure 3.2: *Contour map of the standardized variance function of* $\eta(x, \theta) = \theta_1 + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_1 x_2$ *for design* ϵ_3 *on a unit square.*

3.3 Searching the optimal design

In this section different ways of finding optimal designs are discussed. The properties of the design problem for a certain model and criterion can be extended to a complete analytical solution for a specific design problem (Section 3.3.1). However, most problems are too complex to find an analytical solution. Therefore, Fedorov introduced an algorithm to find the optimal solution (Section 3.3.2). In Section 3.3.3 a combinatorial optimization algorithm is outlined, for the special case of a design space consisting of a finite set of candidate points. Finally, it is illustrated how general purpose optimization algorithms will perform for these kind of problems (Section 3.3.4).

3.3.1 Analytical results

Many examples of analytical derivations of optimal designs can be found in literature (e.g. Fedorov, 1972; Rasch, 1990; Vila, 1991). An illustrative example of an analytical solution of a design problem is given by Boer *et al.* (2000). An exact D-optimal two-point design for the Michaelis-Menten function ($r = m = 2$) can be found by minimizing

$$\phi(M^{-1}[\epsilon]) = K(x_1, x_2, n_1, n_2) = \frac{(1+\beta x_1)^4\ (1+\beta x_2)^4}{\alpha^2\ (x_1 - x_2)^2\ {x_1}^2\ {x_2}^2\ n_1\ n_2} \tag{3.18}$$

It is obvious that n_1 and n_2 should be chosen as equal as possible, as already has been shown by the third theoretical property in Section 3.2.2. The choice of x_1 and x_2 is more complicated. For $x \in [0, x_u],\ x_u > 0$ it can be derived that the following exact design is D-optimal

$$\begin{pmatrix} \frac{x_u}{2+x_u\beta} & x_u \\ n_1 & n_2 \end{pmatrix} \tag{3.19}$$

with $n_1 + n_2 = N$. For $N = 2n$, n_1 equals n_2 and for $N = 2n + 1$ choose $n_1 = n, n_2 = n + 1$ or $n_1 = n + 1, n_2 = n$ (Ermakov and Zhigljavsky, 1987).

3.3.2 Special algorithms for optimal designs

Because an analytical solution can not be found for every design problem, some specific algorithms have been constructed to find the optimal solution. Although more algorithms are available, we restrict ourself to the V-(Fedorov, 1972) algorithm, which can be described as follows:

1. Given start design ϵ_0, stopping criteria, $s = 0$,
   r_0 the number of support points of ϵ_0.

2. Determine:

$$M[\epsilon_s] = \sum_{i=1}^{r_s} p_{is} f(x_{is}) f^T(x_{is}).$$

3. Calculate $D[\epsilon_s] = M^{-1}[\epsilon_s]$.

4. Now $d(x, \epsilon_s) = f^T(x) D[\epsilon_s] f(x)$.

   Determine:

$$\delta_s = \max_x d(x, \epsilon_s) - m.$$

$$\text{a point } x_s^* \in \arg\max_x d(x, \epsilon_s).$$

5. The step-size: $\alpha_s = \delta_s / (\delta_s + (m-1))m$.

6. ϵ_{s+1} is calculated by:

   a)
   Recalculate all the weights $(i = 1, 2, \ldots, r_s)$ of ϵ_s in the following way:

   $$p_{i(s+1)} = p_{is}(1 - \alpha_s).$$

   b)
   Add x_s^* to design ϵ_s with weight α_s, update r_{s+1}. If $x_s^* \in \epsilon_s$, update ϵ_{s+1}.

7. check stopping criteria, $s := s + 1$ and go to 2.

This algorithm is mainly based on the properties of the Equivalence Theorem. It is known that a D-optimal (discrete) design minimizes the maximum of the standardized variance function. This algorithm puts (additional) weight on the value of a point x^* (step 6) where the standardized variance function reaches its maximum (step 4), as long as that maximum is larger than the number of parameters considered. Note that step 4 implies a global optimization problem. Jones and Wang (1999) mention some pros and cons of this algorithm. The main advantage of this algorithm is that the number of support points does not have to be fixed beforehand. Further, it is important that the algorithm ensures convergence to the optimal design under some conditions. The main disadvantage is that the algorithm may be very slow for some problems (Atkinson and Donev, 1992). This is mainly caused by the fact that after introduction, a (possibly non-optimal) support point does not disappear, probably only its weight decreases. The following example is given as an illustration of the V-algorithm.

Consider the following model

$$\eta(x, \theta) = \theta_1 + \theta_2 x + \theta_3 x^2 \tag{3.20}$$

with $X = [0, 1]$. Figure 3.3 shows the standardized variance function of the following (not optimal) design, which is used as a start design for the V-algorithm.

$$\epsilon_0 = \begin{pmatrix} 0 & 0.2 & 1 \\ 0.3333 & 0.3333 & 0.3333 \end{pmatrix}$$

After one iteration of the V-algorithm the design becomes as follows:

$$\epsilon_1 = \begin{pmatrix} 0 & 0.2 & 1 & 0.5378 \\ 0.2492 & 0.2492 & 0.2492 & 0.2523 \end{pmatrix}$$

The standardized variance function of design ϵ_1 is graphically represented in Figure 3.3. We restrict ourselves to one iteration. The final result of the algorithm converges to the D-optimal design, which is:

$$\epsilon^* = \begin{pmatrix} 0 & 0.5 & 1 \\ 0.3333 & 0.3333 & 0.3333 \end{pmatrix}$$

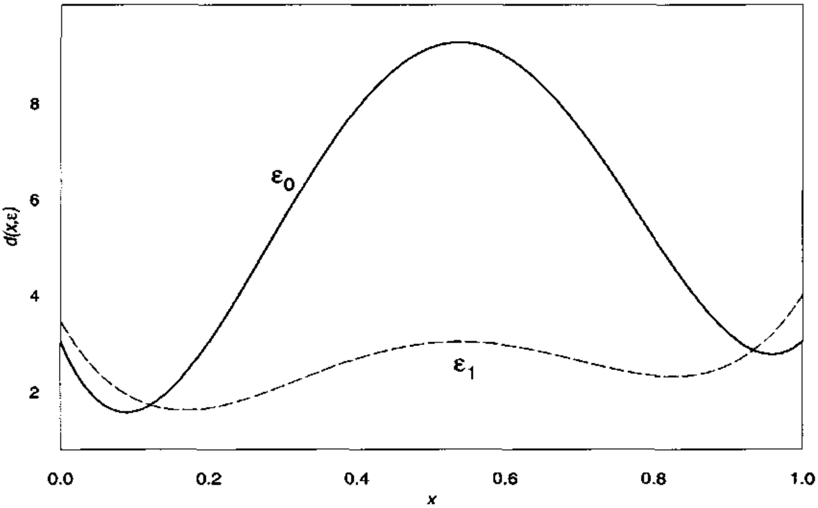


Figure 3.3: *Illustration of the V-algorithm for model* (3.20) *with start design* ϵ_0

3.3.3 Combinatorial optimization of the optimal design

The optimal design problem becomes a combinatorial optimization problem, when the design space is restricted to a finite discrete set of Q candidate (design) points, $B_Q = \{x_1, x_2, \ldots, x_Q\}$. Rasch *et al.* (1997) show some algorithms for selecting N design points out of B_Q. The calculation time of full enumeration of this problem was reduced considerably by applying a branch-and-bound algorithm (for a description, see appendix of this thesis). This fast branch-and-bound algorithm is based on the fact that the criterion function $\phi(M[\epsilon])$ is monotonic (see Section 3.2.2). The drawback of this procedure is that the number of candidate points is restricted to about 30. Figure 3.4 gives an impression of this combinatorial optimization problem for one of the examples used by Rasch *et al.* (1997).

Müller and Pázman (1998) constructed an algorithm to find optimal designs with more candidate points. The algorithm makes use of a corresponding information matrix, which approximates the information matrix for exact designs. Promising results are shown for a spatial example of Fedorov (Fedorov, 1989) of a 20 x 20 point grid.

3.3.4 General purpose optimization

The difficulty in finding optimal designs with general purpose optimization procedures is that the number of support points is not known beforehand. We saw already that there can be given certain bounds for the number of support points in Section 3.2.2. However, it would be preferable when a general purpose

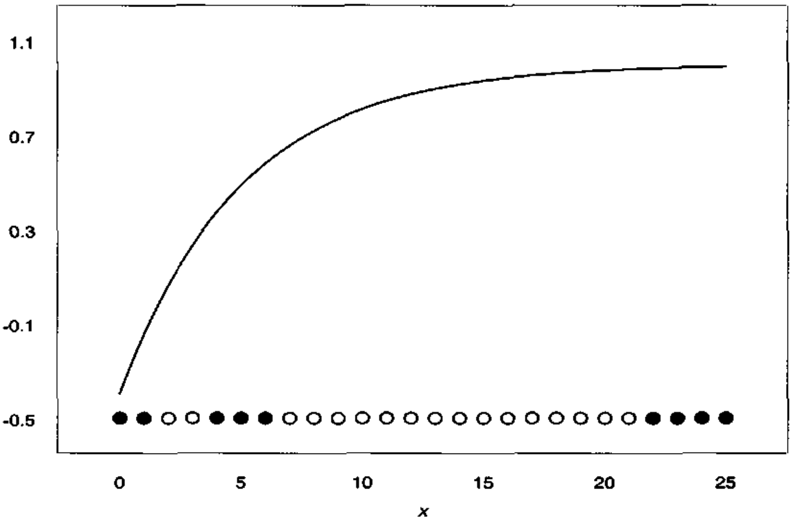


Figure 3.4: *A plot of the function* $1 - 1.4e^{-0.2x}$ *with the corresponding 9-point replicationfree D-optimal design selected from 26 candidate points.*

optimization procedure does not depend on the number of support points. Boer *et al.* (2000) illustrate with the Michaelis-Menten function, that the mixed continuous/integer programming problem can be rewritten into a fully continuous nonlinear programming problem, formulated as follows:

$$\begin{aligned} &\min\{K(x'_1, x'_2, \ldots, x'_N)\} \\ &\text{under the condition :} \\ &x_l \leq x'_1 \leq x'_2 \leq \ldots \leq x'_N \leq x_u \end{aligned} \tag{3.21}$$

where K is equal to a certain criterion, x_l is the lower bound and x_u the upper bound of the one dimensional design space. In this case, x'_i are (single) measurement points in the design space.

In the paper of Boer *et al.* (2000) it is shown that a (sub)-D-optimal design with 6 and 4 replications at the two support points is a local minimum of the continuous optimization problem. Figure 3.5 illustrates this by changing the value of variable x'_6 of the D-optimal design ($x'_1, \ldots, x'_5 = 28.32$, and $x'_6, \ldots, x'_{10} = 1440$) from the lower to the upper bound of the design space. In this way many local minima may appear.

Jones and Wang (1999) argue that general optimization procedures are more efficient than special algorithms like the V-algorithm. They use global optimization methods, because the criterion considered has several local optima. They discuss two well-known stochastic global optimization methods: multistart local search and simulated annealing. For the last it is suggested to stop

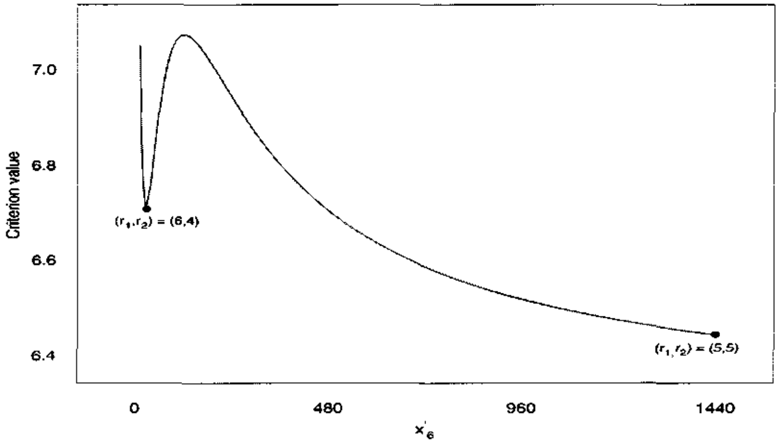


Figure 3.5: *Non-convexity of the continuous NLP formulation of the D-optimal design problem.*

the annealing procedure at a certain point and then continue the search by an effective local search procedure.

3.4 Conclusions

This paper shows how the structure of the design space, model and criterion in optimal design of experiments problems determines the structure of corresponding challenging global optimization problems. Three different kinds of experimental designs are discussed: discrete designs, exact designs and replicationfree designs. Finding the optimal designs for these three concepts involves different optimization problems.

Discrete design problems are most easy to solve. There are many examples of a complete analytical derivation of the optimal design, without using optimization methods. However, if an analytical solution is not available optimization methods are needed. Fedorov (1972) proposes a specific algorithm which ensures convergence to the optimal discrete design, but may be very slow for some problems. General purpose optimization does often not work adequately, because the number of support points is often not known beforehand and local minima may occur (Jones and Wang, 1999).

Exact design problems are hard to solve, because finding the optimal design implies solving a mixed continuous/integer optimization problem. Boer *et al.*

(2000) show that a fully continuous formulation of the problem results in many local minima. An other interesting approach to find exact designs is to construct an exact design from the optimal discrete design with a certain rounding method (Pukelsheim and Rieder, 1992; Gaffke and Heiligers, 1995). It can be shown that the criterion values of these exact designs have a limited loss of efficiency compared to the criterion values of optimal exact designs.

If the design space is restricted to a set of candidate points, combinatorial optimization can be applied to find the optimal solution. Rasch *et al.* (1997) show a branch-and-bound algorithm (full enumeration) for this, based on the fact that every reasonable design criterion is monotonic. A same kind of algorithm, in this case for maximum entropy sampling, can be found in Ko *et al.* (1995). For both articles, full enumeration is only applicable when the number of candidate points is restricted. Larger problems have to be solved with search algorithms (Fedorov, 1989; Müller, 1998).

Up to now no specific global optimization algorithms have been developed in the field of optimal experimental designs. In our opinion, optimal design of experiments constitutes a challenging application field for global optimization.

Acknowledgements

We would like to thank A.C. van Eijnsbergen and D.A.M.K. Rasch for reading and giving helpful comments on earlier versions of the manuscript. We are grateful for the useful remarks of our recently deceased colleague J.M. ten Vregelaar.

Chapter 4

Optimization of monitoring networks for the estimation of the semivariance function

E.P.J. Boer, E.M.T. Hendrix and D.A.M.K. Rasch

MODA 6 - Advances in Model-Oriented Design and Analysis (2001), 21-28

The optimal adjustment of an existing monitoring network for estimation of the semivariance function by means of optimal design of experiments is discussed. The difference between neglecting and including correlation between point pairs, from which the semivariance function is estimated, is visualized for a simple adjustment of a monitoring network. A branch-and-bound algorithm is applied to calculate an exact optimal configuration of monitoring sites (design). For a case study it is shown that the optimal design is robust against misspecified parameter values and model choice.

4.1 Introduction

A problem frequently encountered in practice is the adjustment of an existing monitoring network. Monitoring networks have to be enlarged or reduced, mostly depending on the budgets of the research. One of the biggest problems in designing or adjusting a monitoring network is that the objective of measuring is often ambiguous. A clear example of this can be found in geostatistical analyses. In general, the semivariance function (characterizing spatial continuity) is estimated for kriging (spatial prediction), but the optimal monitoring network for the estimation of the semivariance function is not the same as the optimal network for kriging (Zimmerman and Homer, 1991). In this paper the focus is on the objective of designing an optimal monitoring network for the estimation of the semivariance function. The criterion for optimality, used in this paper, is based on the classical theory of optimal design of experiments (Kiefer, 1959; Fedorov, 1972).

Early discussion of optimal design for estimation of the semivariance function can be found in Russo (1984), followed by a paper of Warrick and Myers (1987). They suggest a criterion which attempts to modify the spatial configuration of points in such a way that distances between point pairs (lags) are as much as possible equally distributed among the several distance classes. The results of this approach hardly depend on the model of the semivariance function.

Zimmerman and Homer (1991) applied the classical theory of optimal experimental design for the estimation of the semivariance function. The design problem consists of adding q new sites, to an existing monitoring network $B_F = \{s_1, ..., s_n\}$, from a set $B_P = \{s_{n+1}, ..., s_{n+Q}\}$ of potential sites. This approach is strongly model-based, which means that caution is needed for model errors (De Gruijter and Ter Braak, 1990). Therefore, investigating the robustness of optimal designs against misspecified parameter values and model choice is important.

In this paper we would like to add three new aspects to the use of optimal designs for the estimation of the semivariance function. In the first place, a visualization of a simple adjustment of a monitoring network is given. Secondly, a branch-and-bound algorithm is applied to calculate an exact optimal configuration of monitoring sites, given a certain criterion and a set of possible monitoring sites. Finally, a robustness study of the optimal design against misspecified parameter values and model choice is done. All is elaborated for a case study introduced by Cressie *et al.* (1990).

4.2 Theory

4.2.1 The semivariance function

A number of papers have been written about design aspects when a certain semivariance function is assumed (e.g. Cressie *et al.*, 1990). However, usually the semivariance function needs to be estimated; which can be considered as a first step in geostatistical analyses.

In geostatistical analyses the random variable Z at a site s in a domain $D \subset \mathbb{R}^2$ can be described as (Cressie, 1991)

$$Z(s) = \mu(s) + \delta(s) \tag{4.1}$$

where $\mu(s)$ is the deterministic part of $Z(s)$, $\delta(s)$ is a spatially dependent zero-mean stochastic process. This paper focuses on the optimal estimation of the unknown spatial correlation of $\delta(s)$. The spatial correlation between points can be quantified by means of the semivariance function:

$$\gamma(h) = \frac{1}{2}\mathrm{var}[\delta(s_1) - \delta(s_2)] \tag{4.2}$$

where it is assumed that the variance of the differences depends only on the (Euclidean) distance $h = \parallel s_1 - s_2 \parallel$ between sites s_1 and s_2. The most common applied parametric semivariance function, $\gamma(h, \theta)$ with unknown parameter vector $\theta = (\theta_1, \theta_2, \theta_3)^T$, is the spherical semivariance function:

$$\gamma_S(h,\theta) = \begin{cases} 0, & h = 0 \\ \theta_1 + \theta_2\left\{\frac{3}{2}\left(\frac{h}{\theta_3}\right) - \frac{1}{2}\left(\frac{h}{\theta_3}\right)^3\right\}, & 0 < h \leq \theta_3 \\ \theta_1 + \theta_2, & h > \theta_3 \end{cases} \tag{4.3}$$

Figure 4.1 shows a spherical function with corresponding interpretation of parameters of semivariance functions. We refer to Isaaks and Srivastava (1989) for the exponential and Gaussian semivariance function.

The semivariance can be estimated from pairs $\{h_k, \hat{\gamma}_k\}$, derived by measuring for every point pair (s_i, s_j)

$$h_k = \parallel s_i - s_j \parallel, \quad k = 1, \ldots, \frac{1}{2}n(n-1) = N.$$

and

$$\hat{\gamma}_k = \frac{1}{2}[\hat{\delta}(s_i) - \hat{\delta}(s_j)]^2, \quad i < j; \qquad i, j = 1, \ldots, n. \tag{4.4}$$

$\delta(s)$ can be estimated by detrending the data by means of median polish (Cressie, 1991). All the pairs $\{h_k, \hat{\gamma}_k\}$ can be displayed as a scatter plot (variogram cloud). Müller (1999) showed that a direct fit of a parametric semivariance function is

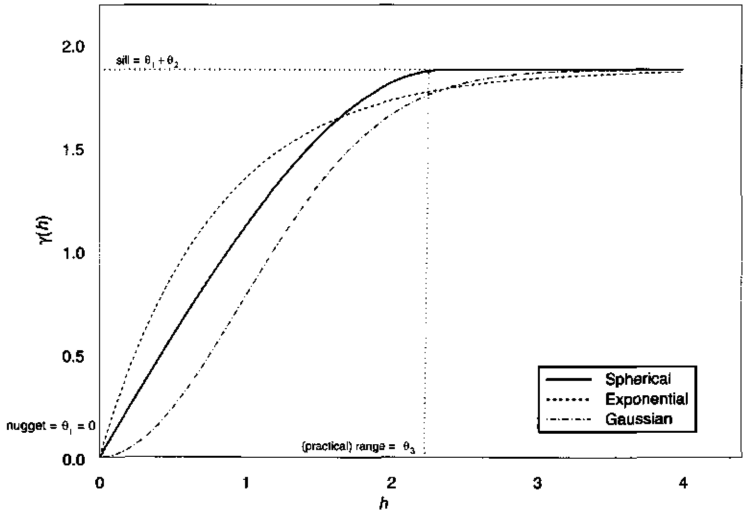


Figure 4.1: *Three frequently used semivariance functions with interpretation of the parameters of the semivariance functions.*

often preferable. The observations of the semivariance function (Equation 4.4) are often correlated, as from n monitoring sites come N observations of the semivariance function. Therefore, generalized least-squares estimation (GLS) is used to estimate (iteratively) $\theta = (\theta_1, \theta_2, ..., \theta_m)$, i.e.

$$\hat{\theta}_{(j)} = \operatorname{Arg}\min_{\theta_{(j)}} [\hat{\gamma} - \gamma(\theta_{(j)})]^T \, \Sigma^{-1}(\theta_{(j-1)}) \, [\hat{\gamma} - \gamma(\theta_{(j)})], \; j = 1, 2, \ldots \quad (4.5)$$

where $\hat{\gamma} = (\hat{\gamma}_1, ..., \hat{\gamma}_N)^T$, $\gamma(\theta) = [\gamma(h_1, \theta), ..., \gamma(h_N, \theta)]^T$ and $\Sigma(\theta)$ is the covariance matrix of $\hat{\gamma}$. If a Gaussian random field is assumed, Cressie (1985) describes how the entries of $\Sigma(\theta)$ can be calculated (see also: Müller, 1998; Section 2.4).

4.2.2 Optimal experimental design for estimation of the semivariance function

Optimal experimental design can be used to maximize the precision of the estimation of parameters of regression models, given the number of allowed observations. The most important references can be found in Fedorov (1972), Silvey (1980) and Atkinson and Donev (1992). Zimmerman and Homer (1991), Müller and Zimmerman (1999) and Bogaert and Russo (1999) showed how the theory of optimal experimental design can be applied to the estimation of the semivariance function. Müller and Zimmerman (1999) clarify the two major differences with the standard optimal design methods for nonlinear regression. The first difference is that the observations of the semivariance function are a result of the spatial configuration of monitoring sites. This means that adding one monitoring site yields n additional pairs of points from which the semivariance function is estimated. Secondly, in optimal design for nonlinear regression models it is usually assumed that the observations are uncorrelated. This is hardly the case for empirical observations of the semivariance function.

Many criteria for optimal designs are functions of the so-called information matrix. Let the monitoring network (design) be denoted as $\xi_n = (s_1, ..., s_n)$, which define $h_1, ..., h_N$. The information matrix corresponding to $\hat{\theta}_{\mathrm{GLS}}$ (Equation 4.5) and a monitoring network ξ_n is equal to

$$M[\theta, \xi_n] = J_\theta^T \, \Sigma^{-1}(\theta, \xi_n) \, J_\theta \tag{4.6}$$

where

$$J_\theta = \begin{bmatrix} \partial\gamma(h_1, \theta)/\partial\theta_1 & \cdots & \partial\gamma(h_1, \theta)/\partial\theta_m \\ \vdots & \ddots & \vdots \\ \partial\gamma(h_N, \theta)/\partial\theta_1 & \cdots & \partial\gamma(h_N, \theta)/\partial\theta_m \end{bmatrix} \tag{4.7}$$

Two ways of calculating the information matrix can be found in literature. One where $\Sigma^{-1}(\theta, \xi_n)$ is approximated by ignoring the off-diagonal elements (Zimmerman and Homer, 1991), mainly in view of the computational advantages. In this way, the correlation between point pairs is neglected. The other way can be found in Müller and Zimmerman (1999) and Bogaert and Russo (1999), where the whole $N \times N$ matrix $\Sigma^{-1}(\theta, \xi_n)$ is included in the calculations.

The so-called D-optimality is equal to the minimization of the determinant of the inverse of the information matrix:

$$\text{D-optimality: } \operatorname{Arg}\min_{\xi_n} \det(M^{-1}[\theta, \xi_n]) \tag{4.8}$$

Optimizing according to (4.8) results in an optimal design for a certain parameter vector θ. Because of this dependence on the parameter values of the semivariance function, the optimal design is a so-called 'locally' optimal design. Note that the word 'locally' is used in another context as in (global) optimization.

4.3 Case study

In this paper we make use of a case study introduced by Cressie *et al.* (1990). This case study considers the Utility Acid Precipitation Study Program (UAPSP) monitoring network located in the eastern and midwestern U.S.A. Annual acid-deposition levels were measured in 1982 and 1983 in a network of 19 U.S. sites ($B_F = \{s_1, ..., s_{19}\}$). There are 11 potential sites ($B_P = \{s_{20}, ..., s_{30}\}$) available for enlarging the monitoring network with one or more sites. The optimal selection of q out of $Q = 11$ sites results in a combinatorial optimization problem,

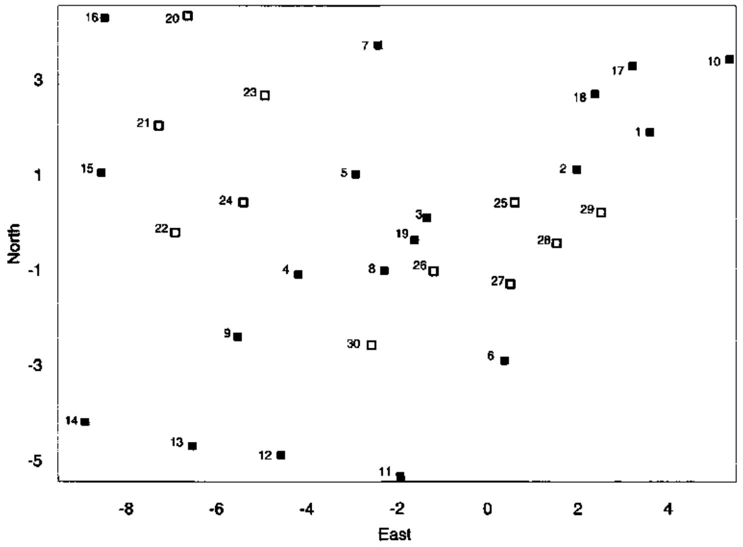


Figure 4.2: *The 19 sites of the UAPSP monitoring network (■) and the 11 potential sites (□). Units are in 100 miles.*

which can be solved by full enumeration with a branch-and-bound algorithm (see Rasch *et al.*, 1997 and appendix of this thesis). Figure 4.2 shows the 19 sites of the existing network and the 11 potential sites.

Zimmerman and Homer (1991) calculate the semivariance function by means of median polish, for the data set considered. The estimated parameter values of a spherical semivariance function were: nugget = 0, range = 236.2 miles and sill = 1.875 (see Figure 4.1 for a plot of this function). We will use this estimated semivariance function as a (basic) semivariance function.

$$\gamma_S(h) = \begin{cases} 1.875\left\{\frac{3}{2}\left(\frac{h}{2.362}\right) - \frac{1}{2}\left(\frac{h}{2.362}\right)^3\right\}, & 0 < h \leq 2.362 \\ 1.875 & h > 2.362 \end{cases} \tag{4.9}$$

4.4 Results

This section of results is split up in three subsections. In the first place, some visualizations will be presented of the problem considered in Zimmerman and Homer (1991). The figures show the difference between neglecting and including correlation between point pairs. Secondly, some results of the branch-and-bound algorithm are discussed. Finally, we will consider the robustness of optimal designs against misspecified parameter values and choice of the model.

4.4.1 Visualization of the problem

The problem of where to add one additional monitoring site to the 19 existing monitoring sites can be visualized by a contourplot of criterion values on a fine regular grid of points. The criterion value at a certain grid point is calculated by supposing that the additional monitoring site was located on that grid point. The contourplot shows where interesting locations are to add one additional monitoring site, given the existing monitoring network and a certain network. This plot shows where interesting locations are to add one additional monitoring site, given the existing monitoring network and a certain criterion. In Figure 4.3 such contourplots are presented for the estimation of the semivariance function written in Equation 4.9 in a D-optimal way, neglecting and including correlation between point pairs respectively.

Figure 4.3 shows on comparing with Figure 4.2 that site 26 is the optimal choice to add, when correlation between point pairs is neglected. This corresponds with the result of Zimmerman and Homer (1991). Note that there are relatively few observations in the interval $h = (0, 2.362)$, which is the most interesting part of the semivariance function. Clusters of points result in more observations of the semivariance function at short distances. This principle comes clearly back in Figure 4.3. The optimal site to add, changes to site 25 for calculations including the correlation between point pairs, see Figure 4.3. The surface of the criterion value is less smooth than in the case of neglecting correlation and it tends less to clustering of monitoring sites.

4.4.2 Results of the branch-and-bound algorithm

Adding only one site to the existing network, can be solved by selection of the site which lowers the criterion value the most. The branch-and-bound algorithm, described in the appendix of this thesis, becomes useful when more than one site has to be added to the existing network. To test the branch-and-bound algorithm, 25 sites will be selected from the combined set of 30 monitoring sites: $B_F \cup B_P$. The number of combinations for this problem is equal to 142506.

For both neglecting and including correlation between point pairs, sites with the numbers $\{7, 10, 11, 14, 23\}$ have to be removed out of $B_F \cup B_P$ to obtain an optimal monitoring network of 25 sites. Surprisingly, the computation time for including correlation is less than for neglecting correlation between point pairs, 1054 and 4934 seconds (Pentium II, 266 MHz) respectively. This is due to the fact that including correlation needs less calls (bounding more efficient) of the recursive branch-and-bound algorithm (2885 and 15710 calls). A simple drop algorithm, sequential removal of one site, finds the optimal solution too. However, these kinds of heuristic algorithms can never guarantee that the solution is optimal.

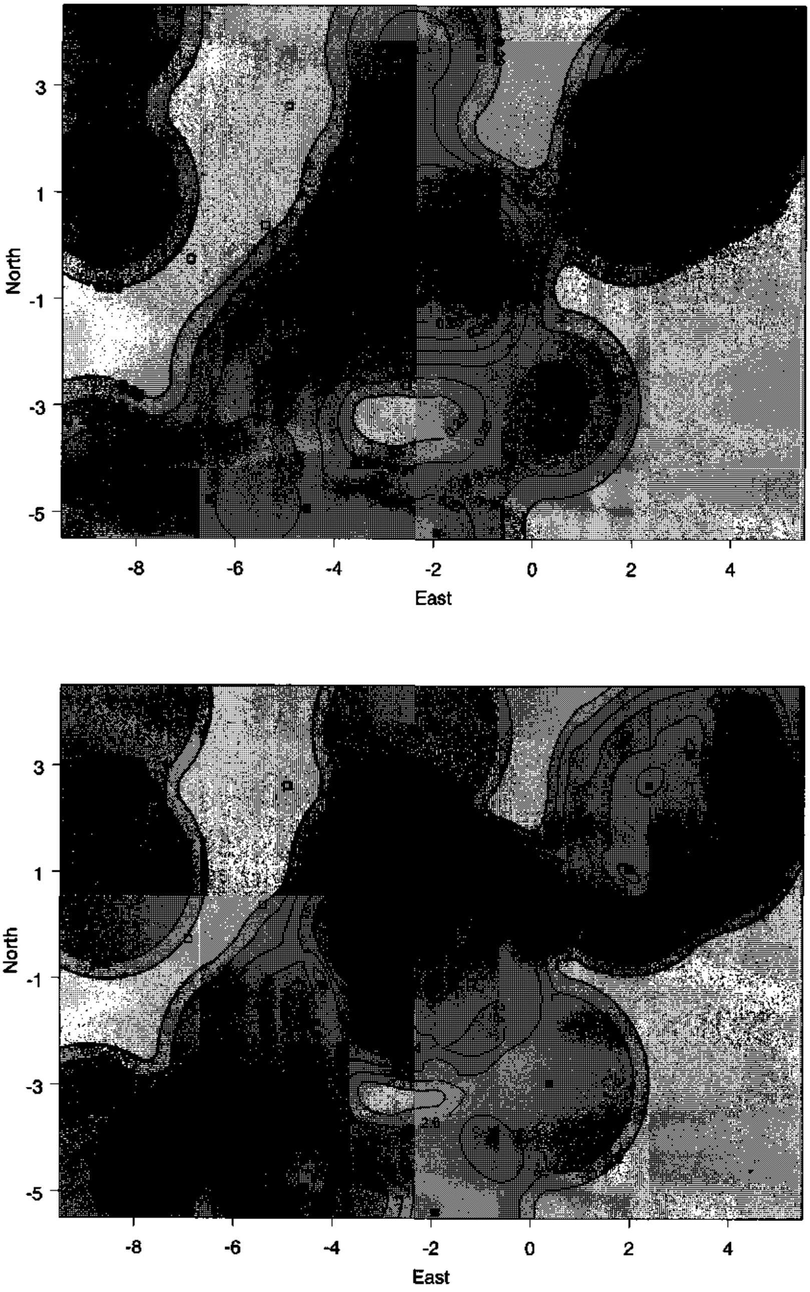


Figure 4.3: *Contourplots of the criterion value (D-optimality) for the estimation of the semivariance function (Equation 4.9) on a fine grid of points, supposing that an additional monitoring site was located on a grid point. Neglecting correlation between point pairs in the upper graph, including correlation in the lower graph.*

Table 4.1: *Optimal combination of 5 sites and robustness ratio for different parameter values of the spherical semivariance function; including correlation between point pairs. The upper table* (**A**) *for Equation (4.9). Lower table* (**B**) *for including nugget effect in Equation 4.9,* $\theta_1 + \theta_2 = 1.875$ *and* $\theta_3 = 2.362$.

A	$\frac{1}{2}\theta_3$	θ_3	$2\theta_3$
$\frac{1}{2}\theta_2$	{2,3,7,9,10}, 1.00	{6,7,8,9,10}, 1.00	{3,4,6,8,11}, 0.83
θ_2	{2,3,7,9,10}, 1.00	{6,7,8,9,10}, 1.00	{3,4,6,8,11}, 0.83
$2\theta_2$	{2,3,7,9,10}, 1.00	{6,7,8,9,10}, 1.00	{3,4,6,8,11}, 0.83

B	θ_3	$2\theta_3$	$3\theta_3$
$\theta_1 = 0$	{6,7,8,9,10}, 1.00	{3,6,7,8,11}, 0.95	{3,4,6,7,11}, 0.92
$\theta_1 = 0.75$	{6,7,8,9,10}, 1.00	{3,6,7,8,11}, 1.00	{3,4,8,9,10}, 0.90
$\theta_1 = 1.5$	{6,7,8,9,10}, 1.00	{3,6,7,8,11}, 1.00	{3,4,7,9,10}, 0.91

4.4.3 Robustness analyses of the optimal design

An optimal monitoring network depends on preliminary estimates of the parameter values of the semivariance function. If these preliminary estimates are too different from the true parameter values θ, the optimized monitoring network will not be optimal for the parameter values which have to be estimated (true parameter values). Robustness analysis is applied to see how the optimal designs will change at deviating values of the parameters. The optimal design of the preliminary estimated parameters can be compared with optimal designs which correspond to parameter values deviating from these values. Let ${\xi_n}^+$ and ${\xi_n}^*$ be optimal designs with corresponding preliminary parameter values θ^+ and deviating parameter values θ^*. A robustness ratio shows how much the criterion values of these designs differ from each other. The robustness ratio is defined as follows:

$$\text{robustness ratio} = \frac{K(\theta^*, {\xi_n}^*)}{K(\theta^*, {\xi_n}^+)} \tag{4.10}$$

where K is criterion function (4.8), which has to be minimized.

The robustness analysis is applied to the situation of adding 5 from the 11 potential sites to the existing monitoring network of 19 sites. In Table 4.1 some results are presented for calculations including correlation between point pairs. In the upper table (A) are shown optimal designs with corresponding robustness ratios for values deviating from $\theta_2 = 1.875$ and $\theta_3 = 2.362$ of Equation 4.9. The lower table (B) shows the result of a robustness analysis including a nugget effect (an additional parameter θ_1), where the value of the sill is kept equal ($\theta_1 + \theta_2 = 1.875$ and $\theta_3 = 2.362$). For the case of neglecting correlation, the optimal designs were all equal to {6,7,8,9,10} for all different combinations of parameter values presented in Table 4.1.

Although the parameter values of the preliminary estimate (by Zimmerman and Homer, 1991) were changed considerably, the optimal combination of sites remains constant for many different parameter values. Furthermore, if another combination of sites was preferred the criterion values did not differ much.

So far, only the parameter values of the spherical semivariance function were changed in our robustness analyses. The question raises how robust is the optimal design against another model choice. Therefore, we calculated the optimal designs for all three functions plotted in Figure 4.1, where the nugget effect was left out of the models. It turned out that the optimal choice of 5 points was equal for all three semivariance functions.

4.5 Discussion and Conclusions

Adding only one site to a monitoring network can be easily solved by a simple algorithm. Visualizations of this problem show the difference between neglecting and including correlation between point pairs. Figure 4.3 shows that extreme clustering of sampling points occurs only when correlation between point pairs is neglected (Van Groenigen and Stein, 1998).

Applying a branch-and-bound algorithm (Rasch *et al.*, 1997) works well for small cases, as considered in this paper. However, when the size of the combinatorial problem increases it will be soon too large to solve within a reasonable computation time. Heuristic search algorithms are needed to solve larger problems. For this specific case study, we found that a simple drop algorithm delivers the optimal design for many cases. Only in a few cases the drop algorithm was trapped in a local minimum. For this case study, the branch-and-bound algorithm is efficient for calculating the optimal design including the correlation between point pairs.

Optimal experimental design for the estimation of the semivariance function is based on preliminary estimates of parameters of the semivariance function and model choice. Therefore, a robustness study against misspecified parameter values and model choice is advisable. Although, the parameter values and the shape of the semivariance function are changed considerably the optimal monitoring designs and the criterion values do not differ much from each other. The dependence of optimal designs on θ seems not to be a major problem. In our opinion, the application of a more complex criterion such as an averaged D-optimality, is not necessary.

Chapter 5

Optimization of a monitoring network for SO_2

E.P.J. Boer, A.L.M. Dekkers and A. Stein

accepted for publication in the Journal of Environmental Quality

In this study, we develop and apply a methodology to reduce an existing monitoring network to find an optimal configuration of a smaller network. We use a criterion based on locally weighted regression with two different weight functions. The methodology is applied to the Dutch national SO_2 network and offers the possibility to include different politically relevant options in the model by weight criteria. Because full enumeration of all monitoring networks is impossible, a combinatorial search algorithm is applied to find a (sub)-optimal solution.

5.1 Introduction

In environmental monitoring programs, optimization of the monitoring network is often an important issue. Most networks for air pollution were designed in the past, the underlying physical processes and emissions may have changed so that collected data somehow do not answer the necessary questions. Several papers deal with optimization of monitoring networks (e.g., Caselton and Zidek, 1984; Warrick and Myers, 1987; Cressie *et al.*, 1990; Pardo-Igùzquiza, 1998b; Van Groenigen and Stein, 1998; Fedorov *et al.*, 1999). The goal of optimizing an environmental monitoring network is, in many cases, related to the accuracy of maps and/or reduction in costs.

Adaptation of an existing monitoring network is often done under constraints of authorities and sometimes based on expert judgment without any formal mathematical criteria. An example of a scientific criterion is the maximization of the minimum distance between monitoring stations, so that the stations are as evenly spread as possible over the area of interest (Müller, 1998). Other criteria are based on geostatistics such as minimization of the kriging variance (Cressie *et al.*, 1990). If little is known about the structure of the stochastic process that underlies the monitoring data, locally weighted regression (Cleveland, 1979) is an appropriate interpolation technique. This flexible method allows the characterization of trends by using simple local models. The variance of estimates resulting from this method can be used to formulate a criterion for optimization.

The aim of this study is to further explore the possibilities of locally weighted regression for the optimization of a monitoring network (Müller, 1995; Fedorov *et al.*, 1999). We show that it leads to a flexible criterion for adaptation of an existing monitoring network. This flexibility is expressed in the application of this interpolation technique and by incorporating different design criteria, which can easily include external (policy) constraints. The criterion is applied to a case study in the Netherlands, where the number of SO_2 monitoring stations has to be decreased by about 60%. The combinatorial optimization problem of selecting the optimal monitoring network is solved by search algorithms.

5.2 Locally weighted regression

5.2.1 Model formulation

Let $\{s_1, s_2, ..., s_n\}$ be locations, $s_i = (x_i, y_i)$, of n monitoring stations in a region D of the monitoring network ξ_n, with corresponding observations $\{z(s_1), z(s_2), ..., z(s_n)\}$ on a certain point of time. The observations are modelled by

$$z(s_i) = \eta(s_i) + \epsilon_i \tag{5.1}$$

where $\eta(s)$ for $s \in D$ is a smooth function and ϵ_i are independent, identically distributed, zero-mean observation errors. A flexible collection of functions consists of those functions that can be approximated locally by a polynomial expression. Consider a location $s^* \in D$, and let $\eta_L(s, \beta(s^*))$ be the local approximation of $\eta(s)$, where $\beta(s^*)$ is a vector with parameters of a locally spatial trend around $s^* = (x^*, y^*)$. The smaller the distance between s and s^* the better the approximation.

For local approximation, we consider the following polynomial expression of order p,

$$\eta_L(s_i, \beta(s^*)) = \sum_{k+j \leq p} \beta_{jk}(x_i - x^*)^j (y_i - y^*)^k + r_i \tag{5.2}$$

which can be considered as a two-dimensional Taylor expansion, where r_i is the remainder term of the approximation.
For an isotropic field Equation (5.2) can be simplified to

$$\eta_L(s_i, \beta(s^*)) = \sum_{j=0}^{p} \beta_j(s^*) h^j + r_i \tag{5.3}$$

where $h = \parallel s - s^* \parallel$. For every estimation location s^*, the vector $\beta(s^*)$ has to be estimated. Weighted least squares is applied with decreasing weights as the distance between s_i from s^* increases. Let an estimator $\hat{\beta}(s^*)$ be defined as

$$\hat{\beta}(s^*) = \arg \min_{\beta(s^*)} \sum_{i=1}^{n} \lambda(s_i, s^*)[z(s_i) - \eta_L(s_i, \beta(s^*))]^2 \tag{5.4}$$

where λ is a weight function depending on observation locations s_i and estimation location s^*. The use of a weight function corresponding to the model assumption that points close to s^* plays a larger role in the determination of $\hat{\eta}_L(s_i, \beta(s^*))$ than points further away.

In principle, many weight functions can be considered. In this study, the choice is restricted to two weight functions mentioned by Müller (1998). The first is the so-called tricube weight function, defined as

$$\lambda_t(s, s^*) = \begin{cases} \left(1 - \left(\frac{h}{h_f}\right)^3\right)^3 & , 0 \leq h \leq h_f \\ 0 & , \text{otherwise} \end{cases} \tag{5.5}$$

where h_f is a smoothing parameter, which determines the neighbourhood of an estimation location s^* and $h = \parallel s - s^* \parallel$. The McLain function is the second choice

$$\lambda_m(s, s^*) = \frac{e^{\frac{-h^2}{h_f^2}}}{h^2 + 0.5} \tag{5.6}$$

where weights are only determined within a fixed neighbourhood with range h_f. Outside this neighbourhood, $\lambda_m(s, s^*) = 0$.

Given the polynomial expression, the weight function and the smoothing parameter h_f the estimation of $z(s^*)$ is equal to:

$$\hat{z}(s^*) = \hat{\eta}(s^*) = \hat{\beta}_0(s^*) \tag{5.7}$$

as can be seen from Equations (5.2) and (5.3).
The smoothing parameter h_f is estimated by calculating cross validation values for a range of smoothing parameter values. The optimal smoothing parameter is chosen as that value for h_f which minimizes the following expression:

$$CV(h_f) = \sum_{i=1}^{n} \{z(s_i) - \hat{z}(s_i)_{(-i)}\}^2, \tag{5.8}$$

where $\hat{z}(s_i)_{(-i)}$ is the local fit at s_i, without using $z(s_i)$ for the estimation, the "leaving-out-one method". The influence of the choice of the weight function will be shown in Section 5.3.

5.2.2 Optimal design for locally weighted regression

Estimation variances of locally weighted regression parameters are used as a basis for a criterion for optimizing a monitoring network. These estimation variances are estimated on a set of locations where estimations are required: $\{s_1^*, s_2^*, ..., s_q^*\}$. Given a weight function and a value of the smoothing parameter h_f, the classical theory of optimal design of experiments is applicable to every single estimation location. This allows formulation of a criterion for optimizing the monitoring network $\xi_n = \{s_1, s_2, \ldots, s_n\}$.

Let F_j' be the design matrix in the case of an isotropic field and let $p = 1$ for the order of Equation (5.3). Then

$$F_j' = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ h_{1j} & h_{2j} & \ldots & h_{nj} \end{pmatrix} \tag{5.9}$$

where $h_{ij} = \|s_i - s_j^*\|$. If Λ is diag(λ), $\lambda = [\lambda(h_{1j}), \lambda(h_{2j}), \ldots, \lambda(h_{nj})]$ and the so-called information matrix is $M_j^{-1}(\xi_n) = F_j'\Lambda F_j$, which depends on the design ξ_n, then

$$\hat{\beta}(s_j^*) = M_j^{-1}(\xi_n) F_j' \Lambda z. \tag{5.10}$$

Since

$$\hat{z}(s_j^*) = \hat{\beta}_0(s_j^*) \tag{5.11}$$

only the first element of the parameter vector is required for estimation of $z(s_j^*)$. The estimated variance of $\hat{\beta}_0(s^*)$ then equals

$$\hat{\text{var}}(\hat{\beta}_0(s_j^*)) = \hat{\sigma}^2 \left(M_j^{-1}(\xi_n)\right)_{11} \tag{5.12}$$

where $\left(M_j^{-1}(\xi_n)\right)_{11}$ is the upper left element of the matrix $M_j^{-1}(\xi_n)$ and σ^2 is the residual variance.

A criterion for assessing the performance of a network design ξ_n can be derived from (5.12) by calculating a weighted sum over a set of q estimation locations $\{s_1^*, \ldots, s_q^*\}$:

$$\phi(\xi_n) = \sum_{j=1}^{q} w_j \left(M_j^{-1}(\xi_n)\right)_{11}. \tag{5.13}$$

The value of $\phi(\xi_n)$ is used as criterion for obtaining an optimal monitoring network ξ_n. The inclusion of the weights w_j allows some locations to be considered more important than others, because of external or political reasons (see Section 5.2.4). An optimal design ξ_n *is the design that minimizes the value of* $\phi(\xi_n)$ in Equation (5.13).

5.2.3 Case study

In 1993, the RIVM (National Institute of Public Health and the Environment in the Netherlands) measured SO_2 at 74 measuring stations of the Dutch National Air Quality Monitoring Network (Doesburg, *et al.*, 1994). Figure 5.1 shows the original monitoring stations with their annual average concentrations (μg m^{-3}). To reduce the expense of data collection this network was reduced to 29 stations, a reduction of 60%. This reduction is feasible because SO_2 concentrations are decreasing and the political pressure to maintain an expensive monitoring network is decreasing. In addition, deterministic models are available that allow the reliable calculation of SO_2 concentrations (Bleeker and Den Hartog, 1995). However a total abandoning of the entire network is not possible because of national and European regulations.

Figure 5.1 shows that both the annual average SO_2 concentrations and spatial variability of these values (i.e. relatively large differences in concentrations on short distance) are higher in the South-West of the Netherlands compared with the North of the Netherlands. Annual mean concentrations at 74 stations have an average of 10.2 μg m^{-3}, whereas the minimum and maximum values are 4.2 μg m^{-3} and 28.1 μg m^{-3}, respectively, and the variance is 26.6 $\mu g^2\ m^{-6}$. A geostatistical approach toward optimization (Van Groenigen, 1999) was not an attractive option, because of difficulties in modelling the trend and because

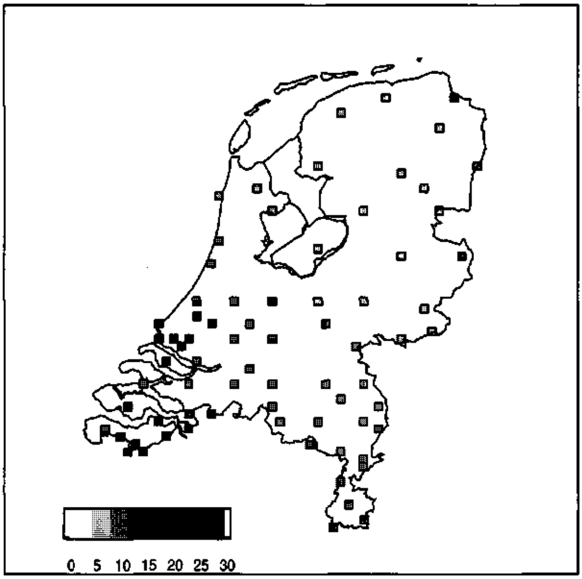


Figure 5.1: *Monitoring network of 74 stations in the Netherlands in 1993 coded with the annual average SO_2 concentration ($\mu g/m^3$).*

of the high spatial variability in the South-West compared with the North of the Netherlands. Therefore, the trend surface is estimated by a nonparametric regression technique: locally weighted regression, which was described in Section 5.2.1.

5.2.4 Specification of different design criteria

The formulated criterion for the optimization of a monitoring network (5.13) is used for both weight functions in (5.5) and (5.6), to reduce the number of monitoring stations from 74 to 29. Given a particular choice for a weight function, the smoothing parameter h_f has to be determined. By a reduction, however, the optimal smoothing parameter increases since the number of monitoring stations decreases. A small monitoring network leads to high variability of estimated values of the smoothing parameter. It is possible, however, to determine optimal smoothing parameters for several monitoring designs given the desired number of stations n. The average of the smoothing parameters thus obtained can be considered as an optimal smoothing parameter for a monitoring network of 29 stations.

Also, the number and coordinates of estimation locations have to be determined with corresponding weights (w_j). Three sets of weights are used in this paper in

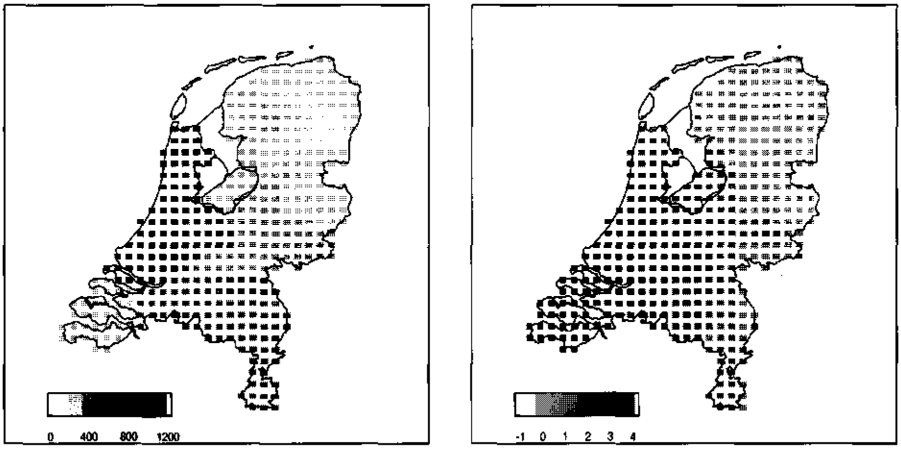


Figure 5.2: *Estimation locations with corresponding weights based on population density at each location in number of inhabitants per km^2 (criterion II, left). Estimation locations with weights based on residuals of locally weighted regression (criterion III, right).*

order to accommodate external information or political regulations into criterion (5.13). The following criteria are considered:

- Criterion I puts equal weights at every estimation location.

- Criterion II puts weights according to the population density in the different provinces. It is an example of including political regulations into the optimization of a monitoring network.

- Criterion III puts weights in such a manner that it reflects the differences in spatial variability, that is, the greater variability in SO_2 values in the South-West compared with the North of the Netherlands on short distances. The idea behind these weights is that in the North fewer monitoring stations are required than in the South-West because of the relatively constant values of the annual average SO_2 concentrations. The weights are based on mean residual values of locally weighted regression at the measurement locations in a neighbourhood with range h_f for every grid location. In this case, for $p = 1$, isotropic, tricube weight function with $h_f = 93$.

Figure 5.2 shows estimation locations with corresponding weights for Criteria II and III.

5.2.5 Optimal reduction of a monitoring network

The reduction of an existing monitoring network is a combinatorial optimization problem (Ko *et al.*, 1995). For the case study we have to select 29 stations from a full set of 74 possible monitoring stations. This combinatorial optimization problem yields a large number of possible combinations (3×10^{20}). It is impossible to enumerate these in full to find the best solution in terms of criterion (5.13) and the different design criteria. Therefore, we developed a sequential search algorithm to find (sub)-optimal solutions (Section 5.2.5). In Section 5.2.5 this algorithm is improved by using combinatorial solutions of sub-problems.

Sequential search algorithm

A combination of a drop- and an add algorithm is applied as search algorithm. The drop algorithm sequentially removes stations from a set of monitoring stations ξ_k until n monitoring stations are left. Mostly the algorithm will be started at $k = N$, where N is the number of stations in the original monitoring network. The set of removed points is defined as ξ^+_{N-k}, which is used later on in the drop-add algorithm. The drop algorithm is described as follows (Rasch *et al.*, 1997):

drop algorithm

1. Let $\xi_k = \{s_1, s_2, \ldots, s_k\}$ and $\xi^+_{N-k} = \{s_{k+1}, s_{k+2}, \ldots, s_N\}$.
2. Determine $\phi^* = \min \phi(\xi_k \backslash \{s_t \in \xi_k\})$ over $t = 1, \ldots, k$.
3. Drop the point s_t corresponding to ϕ^* from the design ξ_k;
   Add the point to ξ^+_{N-k}; $k := k - 1$; Renumber set ξ_k.
4. If $k = n$, STOP
   Otherwise go to 2.

The add algorithm can be formulated in a manner analogous to the drop algorithm (Rasch *et al.*, 1997). The idea of the drop-add algorithm is that some points can be exchanged after running the drop algorithm. The number of additional points added to, or dropped from, a design is called m. The drop-add algorithm consists of three steps. First an (n-m)-point design is obtained with the drop algorithm, followed by an addition of $2 \times m$ points and finally deletion of m points, so that an n-point monitoring design is obtained. This is repeated until no improvements are found.

Search algorithm with combinatorial analysis

Although a full enumeration of all combinations is impossible, smaller combinatorial sub-problems can be solved. Sequential search algorithms add or drop one point at each iteration. It may be worthwhile to consider dropping combinations of points. Rasch *et al.* (1997) describe a branch-and-bound algorithm that solves combinatorial problems for optimal designs in regression analysis. With some adaptations, the branch-and-bound algorithm can be applied to the optimization problem discussed in this paper. The branch-and-bound algorithm is part of the last step of the drop-add algorithm, which is the final dropping of m points.

5.3 Results and discussion

5.3.1 Maps of interpolated SO_2 values with locally weighted regression

As pointed out in Section 5.2.1, three choices have to be made in order to apply locally weighted regression. For the order p of the polynomial expression, $p = 1$, both isotropic and anisotropic are decided upon first, being a compromise between computational ease and flexibility (Müller, 1998). A higher order of the polynomial expression did not improve the prediction accuracy. The McLain and the tricube weight function - Equations (5.5) and (5.6) - were used as a weight function. Given a weight function, the unknown smoothing parameter h_f is estimated by cross validation. Figure 5.3 shows the cross validation results for $p = 1$, both isotropic and anisotropic for both weight functions and these weight functions under the optimal smoothing parameter. A simplification of the problem to an isotropic approach results in higher values of the cross validation. In the anisotropic case, a small range of a neighbourhood h_f is preferred for both weight functions. However, h_f can not be chosen too small because at least three measurement points have to be available in the neighbourhood. Therefore, a range of a neighbourhood of 70 km ($h_f = 70$) is chosen for both weight functions. Furthermore, Figure 5.3 shows that cross validation values for the McLain weight function are smaller for every h_f (separately for isotropic and anisotropic). This indicates that locally weighted regression with the McLain weight function produces estimations with a higher prediction accuracy. However, caution is required because the results are only based on observations made at 74 monitoring stations. Maps of interpolated SO_2 values are shown in Figure 5.4.

The maps of interpolated SO_2 values in Figure 5.4 show a clear difference. The surface of interpolated SO_2 values using the tricube weight function is smoother than the map obtained by using the McLain weight function, as a result of different shapes of the weight functions (Figure 5.3, under). The McLain weight is almost an exact interpolator, because the point closest to s^* gets a large weight compared with the weights of points further away.

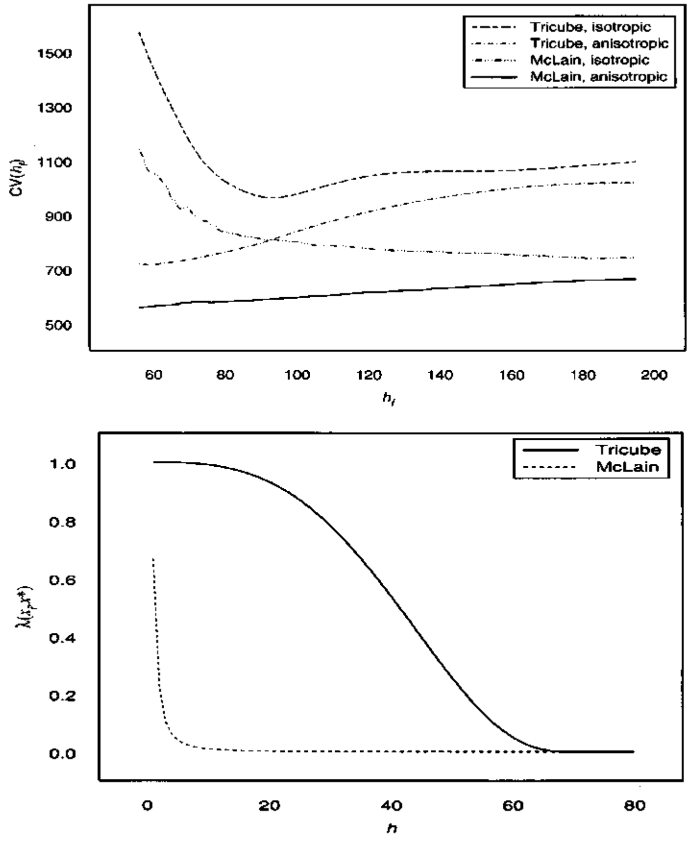


Figure 5.3: *Plots of cross validation values ($N = 74$) for a range of smoothing parameters for the tricube and the McLain weight function both isotropic as anisotropic(upper). The weight functions with the optimized smoothing parameters (under).*

5.3.2 Comparison of search algorithms

In Section 5.2.5 three sequentially search algorithms are introduced: the drop algorithm, the drop-add algorithm and the drop-add algorithm with combinatorial analysis. In this paragraph algorithms will be compared, because of computational reasons the calculations are restricted to the isotropic case. A smoothing parameter of 175 (km) for the tricube weight function and 183 (km) for the McLain weight function is chosen as a test case.

Figure 5.5 shows how the values of $\phi(\xi_n)$ increase when monitoring stations are dropped sequentially from an existing monitoring network. The values of $\phi(\xi_n)$ are calculated for the tricube weight function with a smoothing parameter of 175 km with equal weights at the q estimation locations. The smaller the

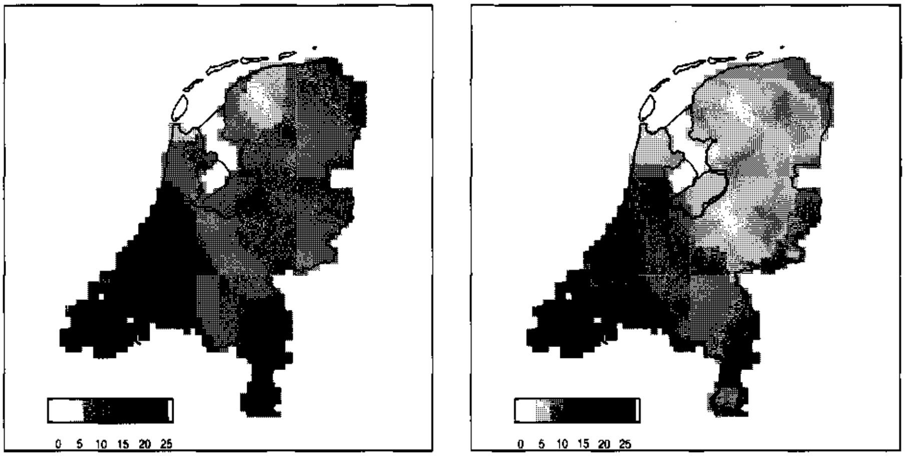


Figure 5.4: *Maps of interpolated* SO_2 *values* ($N = 74$) *of the annual average concentration of* SO_2 *by locally weighted regression,* $h_f = 70$. *Tricube weight function (left) and McLain weight function (right).*

monitoring network becomes, the larger the influence on the value of $\phi(\xi_n)$ of dropping one station from the monitoring design.

The drop algorithm is further refined to the drop-add algorithm and the drop-add algorithm with combinatorial analysis. The question that we would like to discuss in this paragraph is how far designs found by the sequential drop-add algorithm (A) differ from those found by the drop-add algorithm with combinatorial analysis (B). Therefore we introduce efficiency of a monitoring design by dividing the value of $\phi(\xi_n)$ of the result of algorithm B by $\phi(\xi_n)$ of the design found by algorithm A. These results are compared with values of $\phi(\xi_n)$ of designs found by maximizing the minimum distance (maxmin distance design). This is a very simple way of calculating an optimal design, which is comparable with common practice. Table 5.1 presents these efficiencies. Note that only values of $\phi(\xi_n)$ of designs calculated with the same weight function and criterion (I, II or III) can be mutually compared.

Table 5.1 shows that there is only a small difference between the drop-add algorithm with combinatorial analysis and the sequential drop-add algorithm. For the McLain weight function with criterion III, the value of $\phi(\xi_n)$ was even slightly higher with algorithm B (0.01%). The maxmin distance designs have considerably higher values of $\phi(\xi_n)$. The computation time for algorithm B is approximately 80 minutes (Pentium II, 266 MHz), about 200 times as much as for algorithm A.

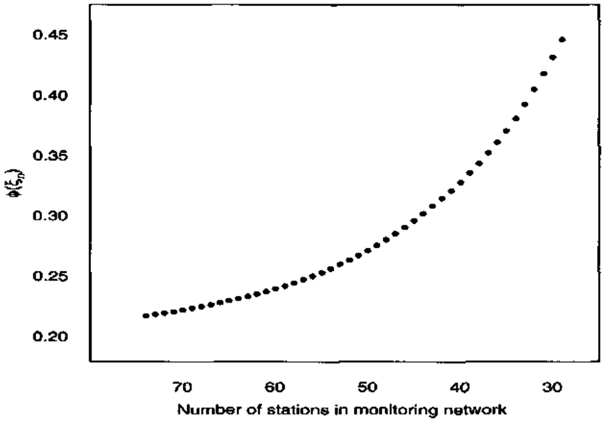


Figure 5.5: *Values of $\phi(\xi_n)$ of a monitoring network where sequentially monitoring stations are dropped from the original network of 74.*

5.3.3 Optimized monitoring networks

As already pointed out in Section 5.2.4, a smoothing parameter for a given weight function has to be chosen before starting the optimization. The optimal values for the smoothing parameters found in Section 5.3.1 were obtained using the full network of 74 stations. The values of the smoothing parameters will be larger for reduced monitoring networks. To obtain an optimal smoothing parameter for a monitoring network of 29 stations, the 74 stations are divided into three clusters and a random selection of 29 stations (in total) is chosen from these clusters. This is followed by the determination of the optimal smoothing parameter, (i.e. h_f with the lowest value of the cross validation). This is done 500 times, the average of the optimal smoothing parameters is considered as the optimal h_f for a monitoring network of 29 stations for all three criteria. A

Table 5.1: *Efficiency of monitoring designs of different algorithms of the sequential drop-add algorithm (A) and maxmin algorithm compared with the drop-add algorithm with combinatorial analysis (B). Values of $\phi(\xi_n)$ of designs found by algorithm B are divided by values of $\phi(\xi_n)$ found by other algorithms, presented as percentage.*

	Tricube		McLain	
criterion	algorithm A	maxmin	algorithm A	maxmin
I	99.95	91.56	100.00	84.67
II	99.77	86.49	100.00	77.77
III	99.68	83.89	100.01	77.89

smoothing parameter of 120 (km) for the tricube weight function and 136 (km) for the McLain weight function is obtained by this procedure.

Results of the drop-add algorithm (Section 5.2.5) when $n = 29$ and $m = 6$ for the three different design criteria are presented in Figure 5.6. A comparison of results for two weight functions shows that the McLain weight function tends to spread the monitoring stations more over the Netherlands than the tricube weight functions. Further, a clear influence of the different criteria (weights at estimation locations) can be seen in the configuration of monitoring stations, especially for the McLain weight function. Monitoring stations are moved from parts of the Netherlands with low weights to parts with high weights at estimation locations.

Choice of a weight function for locally weighted regression has a considerable influence on the final optimal monitoring network. The two weight functions considered in this paper have a different shape. The difference between weights put at points close to s^* and points further away is small for the tricube weight function compared with the McLain weight function. The McLain weight function puts relatively high weights at points close to s^* and weights diminish rapidly as the distance to s^* gets larger. Therefore the McLain weight function will tend to spread the monitoring stations as evenly as possible over the region.

The choice of the weight function is, in principle, arbitrary. An investigation with cross validation can help to choose the best concerning to the estimation accuracy. The smoothness of the surface of SO_2 values is also controlled by the weight function. If there is interest in the mean SO_2 concentration over the Netherlands a more smooth results will be desirable (tricube weight function). If the maximum SO_2 concentration is of interest, a weight function as the McLain weight function will be preferred.

In this study we used locally weighted regression for interpolation and optimization of an existing monitoring network. Other interpolation techniques such as stratified kriging could be used, wherein the South-West of the Netherlands can be considered as a separable stratum. However, optimizing a monitoring network for the entire country can be a problem, because there are not enough monitoring stations in a stratum to fit a semivariance function to use for kriging. Furthermore, it would be difficult to optimize a monitoring network near the boundary between two strata.

5.4 Conclusions

The reduction of a monitoring network is investigated in this study. Design of experiments for locally weighted regression is used to reduce the number of stations in an optimal way. Different criteria are formulated on the basis of different objectives for monitoring SO_2, these are reflected in the weights assigned at estimation locations. We make the assumption throughout this

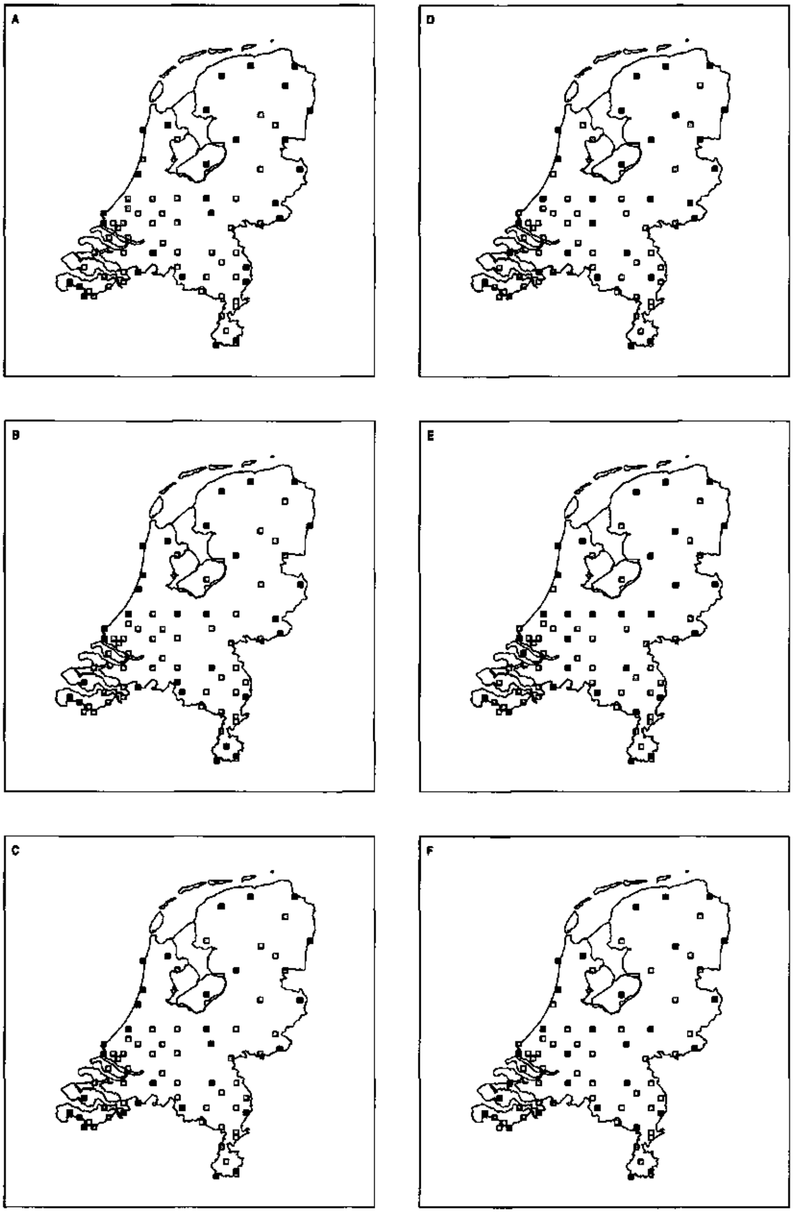


Figure 5.6: *Monitoring networks (black squares) obtained from the network of 74 monitoring stations by the drop-add algorithm* ($n = 29, m = 6$)*. Tricube weight function,* $p = 1$*, anisotropic and* $h_f = 120$ *(km), for the three design criteria in Section 5.2.4, (A-C). Same for McLain weight function,* $h_f = 136$ *(km) (D-F).*

study that σ is constant over the whole country. Possibly, if more data are available, then a locally estimated σ - $\sigma(s)$ - could be included into criterion (5.13).

Two search algorithms are used: a sequential drop-add algorithm (A) and a drop-add algorithm with combinatorial analysis (B). Algorithm B requires more computation time than algorithm A, with a slight improvement of efficiency (Table 5.1).

Different design criteria, distinguished by the choice of the weights at estimation locations, result in different optimal monitoring networks. A precise formulation of the monitoring objective is necessary to make sure that the optimized monitoring network is indeed adequate. Table 5.1 shows that the values of $\phi(\xi_n)$ of maxmin distance designs can be considerably greater than those of optimized designs. Although the objective of a monitoring network is difficult to formulate in practice, it should receive more attention.

Acknowledgements

The authors would like to thank the Air Research Laboratory (LLO) of the RIVM for providing the data. They also thank A.C. van Eijnsbergen and E.M.T Hendrix for giving helpful suggestions for this paper.

Chapter 6

Optimization of a monitoring network for groundwater level

Many spatial-temporal environmental processes are followed by monitoring networks. One example is a monitoring network for groundwater level. At a number of piezometers the groundwater level is measured in a certain frequency in time. Such a monitoring network in the Veluwe area of the Netherlands is taken as a case study. This paper will focus on a possible reduction of the number of measurements at this monitoring network without losing much information about the groundwater level at the different piezometers. The investigations of a reduction of the number of measurements is based on a spatial-temporal model. This model consists of three components: the average groundwater level at a piezometer, the seasonal component at a piezometer and a spatial-temporal error-term. By means of a simulation study different frequencies of measuring are tested at two different piezometers. Furthermore, investigations are done to see how different spatial-temporal monitoring strategies are performing on the basis of a nonseparable spatial-temporal semivariance function. For this study, results show that a reduction of the number of measurements is possible.

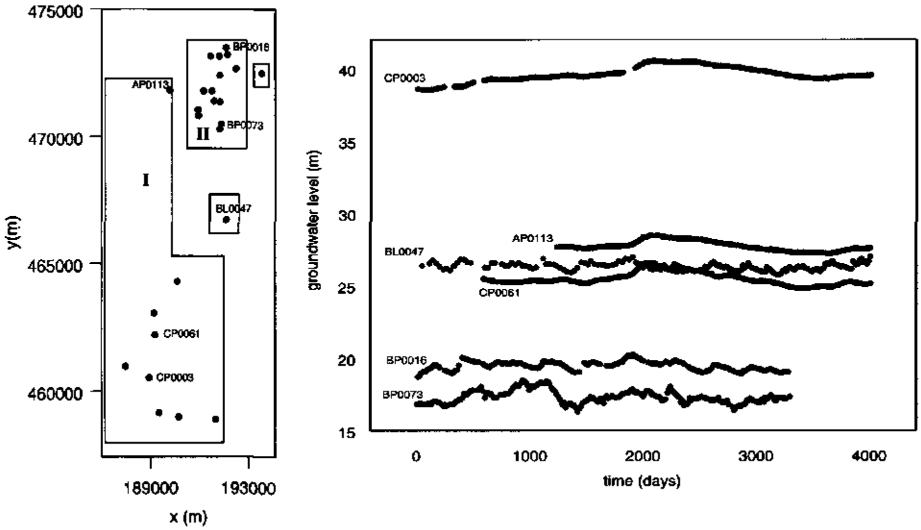
6.1 Introduction

Monitoring of environmental processes often results in large data sets which are characterized by temporal and spatial heterogeneity. In practice, the question arises whether the sampling effort is proportional to the need of information of the environmental processes. To answer that question it is necessary to precisely formulate the goal of monitoring. Goals include the risk of crossing a threshold in a lead-pollution study (Van Groenigen *et al.*, 1997), a time trend estimation involving a surface temperature field (Sølna and Switzer, 1996) and prediction in space and time for environmental and agricultural related phenomena (Stein *et al.*, 1998). Monitoring networks often have a plural goal of measuring. Different goals of monitoring require the formulation of different criteria. These may include minimization of the average prediction error variance, minimizing the maximum prediction error variance or minimization of the variance of regression parameter estimates. Evaluating these criteria for the existing monitoring network can result into the conclusion that the monitoring network has to be enlarged or reduced.

In recent years, many papers have been published on either spatial-temporal modelling (e.g. Heuvelink *et al.*, 1997; Wikle *et al.*, 1998; De Cesare *et al.*, 2001) or optimal spatial designs (e.g. Müller, 1998; Fedorov *et al.*, 1999; Van Groenigen, 1999; Prakash and Singh, 2000). However, only few publications concern (model-based) spatial-temporal design of environmental monitoring networks. Stein *et al.* (1998) design an optimal monitoring network for estimation of the spatial-temporal semivariance function by simulated annealing. Wikle and Royle (1999) study the benefit of mobile monitors. They show that sampling plans can be improved if allowed to change with time. Unfortunately, not in every situation the monitor can be moved. In this study we consider a monitoring network where spatial locations of monitors are fixed and only the frequency of measuring can be changed. In particular we will look at the consequences of reducing a monitoring network for groundwater level data.

Groundwater is an important source for drinking-water in many areas of the world. Research investigates both its quantitative and qualitative properties. Rouhani and Hall (1989) and Stein (1999) extended the spatial geostatistical analysis of groundwater data in spatial-temporal geostatistical analysis. Use of temporal information results into more accurate maps. Another example of spatial-temporal modelling of groundwater can be found in D'Agostino *et al.* (1998). These papers focus on modelling characteristics of groundwater, but do not pay attention to sampling design.

The purpose of this study is to address the problem of a reduction of a monitoring network, based on a spatial-temporal model. In the Veluwe area in the Netherlands, groundwater is extracted to serve as drinking-water. We will use a part of its monitoring network as a case study. We focus on groundwater level, which a Dutch public utility is monitoring using piezometers. The utility suspects that the number of measurements of groundwater level is too high, because



(A) (B)

Figure 6.1: *Locations of piezometers and indication of clustering - I, II and two separate piezometers - (A) and groundwater level fluctuations from January 1983 to December 1993 for the six coded piezometers (B).*

they find approximately the same time series at different piezometers. In this paper we look at consequences, in terms of mean squared and absolute errors and kriging variances, of a reduction of this monitoring network for groundwater level. The paper is organized as follows. Section 6.2 shows some characteristics and details of the available data set. Section 6.3 deals with spatial-temporal modelling of groundwater. Given this model, the consequences of a reduction of the monitoring network (both frequency of measuring and the number of piezometers) are described in Section 6.4. Finally, the results will be discussed in Section 6.5.

6.2 The data set

The data set consists of biweekly (approximately) measurements of groundwater levels on 25 locations in the Veluwe area in the middle of the Netherlands from 1 january 1983 to 31 december 1993. Figure 6.1 (A) shows these locations whereas Figure 6.1 (B) shows the time series at 6 of these locations. Time series at a location can be incomplete, as Figure 6.1 (B) indicates. Table 6.1 shows some summary statistics of the six coded piezometers.

The variability of groundwater level fluctuations can be very different at measuring locations, mainly as a consequence of differences in thickness of the un-

Table 6.1: *Summary statistics of six piezometers. Number of measurements (N), the average groundwater level ($Mean$), the minimum (Min) and maximum (Max) groundwater level and the standard deviation (Stddev.) of time series.*

Piezometer	N	$Mean$	Min	Max	$Stddev.$
CP0003	247	39.61	38.53	40.56	0.52
AP0113	232	27.61	26.78	28.47	0.40
BL0047	178	26.30	25.74	26.99	0.25
CP0061	255	25.31	24.14	26.56	0.51
BP0016	198	19.45	18.71	20.24	0.31
BP0073	211	17.30	16.29	18.45	0.45

saturated zone in the Veluwe area. A thick zone corresponds to a low influence of short-term fluctuations like wet and dry seasons and rainfall. Short-term rainfall events are buffered in the thick unsaturated zone and do not show up in the groundwater level (Gehrels, 1999).

The difference in variability is also shown in Figure 6.1 (B). Small scale variability on piezometers corresponds with a relatively thin unsatured zone and large scale variability on piezometers with a thick unsatured zone. Piezometer BL0047 seems to be an exception, possibly as a consequence of a different structure of the unsaturated zone (Dufour, 1998). This heterogeneity hampers the modelling of groundwater level. To reduce heterogeneity we propose clustering of piezometers based on the correlation between time series at piezometers. In this way two spatial clusters of piezometers are formed. Cluster I consists of points which are highly correlated (≥ 0.88) with piezometer CP0003. Cluster II consists of piezometers which are correlated (≥ 0.53) to piezometer BP0073. Two piezometers called BP0023 and BL0047, did not fit into both clusters and are considered separately. Clusters I and II are indicated in Figure 6.1 (A) and will be used for spatial-temporal modelling and investigation of a possible reduction of the measuring frequency and/ or piezometers. A reduction within clusters is a conservative reduction, because no information of other clusters is used.

6.3 Spatial-temporal modelling of groundwater level

To model groundwater levels we consider a spatial-temporal random field $Z(s,t)$ where $s=(x,y)$ is a location within the geographical area of interest and t is the time index. Let $Z(s,t)$ have the following decomposition:

$$Z(s,t)=\mu(s)+seas(s,t)+\epsilon(s,t) \qquad (6.1)$$

where $\mu(s)$ is the average groundwater level at location s, $seas(s,t)$ is the seasonal component and $\epsilon(s,t)$ is an error-term, with zero expectation and spatial-temporal correlated variance. In the next section it is shown how the seasonal component is estimated, followed by Section 6.3.2 where the spatial-temporal variability of the error-term is estimated.

6.3.1 Elimination of seasonality

The raw data are preprocessed in the following way. To eliminate the seasonality in the different time series we applied moving average estimation (Brockwell and Davis, 1991). For this procedure measurements have to be equally spaced in time. The time series in this study are not equally spaced in time, because measuring has not been done exactly every two weeks and because of missing values. To obtain equally spaced series we applied linear interpolation to biweekly spaced time series. The seasonal component is modelled by periodic regression (Batschelet, 1981). Batschelet also shows how to deal with moderately skew seasonal oscillations, as occur in this study.

The following technique is used to estimate the seasonal component with period $d = 2q = 26$ on each measurement location:

1. Apply a moving average estimation on every time point t at the different measurement locations s:

$$\begin{aligned}\hat{m}_t = (0.5z(s,t-q) + z(s,t-q+1) + \cdots + z(s,t+q-1) +\\ 0.5z(s,t+q))/d, \quad q < t \leq Q-q. \quad (6.2)\end{aligned}$$

   where Q is the number of time index t of the last measurement.

2. If data are available, calculate for each $k = 1, \ldots, d$ and for each year $j = 0, 1, \ldots, 10$

$$z^*_{kj} = z(s, k+jd) - \hat{m}_{k+jd}, \quad q < k+jd \leq n-q. \quad (6.3)$$

3. Estimate the parameters of the following nonlinear periodic regression function (Batschelet, 1981) through z^*_{kj}:

$$z^*_{kj} = A\cos(\omega k - \phi + \nu\cos(\omega k - \phi)), \quad k = 1, \ldots, d. \quad (6.4)$$

   with A the amplitude of the oscillation, $\omega = 2\pi/d$, ϕ determines the peak phase of the function and the parameter of skewness ν is limited to the interval: $-0.50 \leq \nu \leq 0.50$.

The seasonal component is characterized by the estimated nonlinear function at each measurement location. Estimated seasonal components are shown for piezometers CP0003 and BP0016 in Figure 6.2. The parameters of the nonlinear periodic regression function (6.4) are estimated by PROC NLIN in SAS

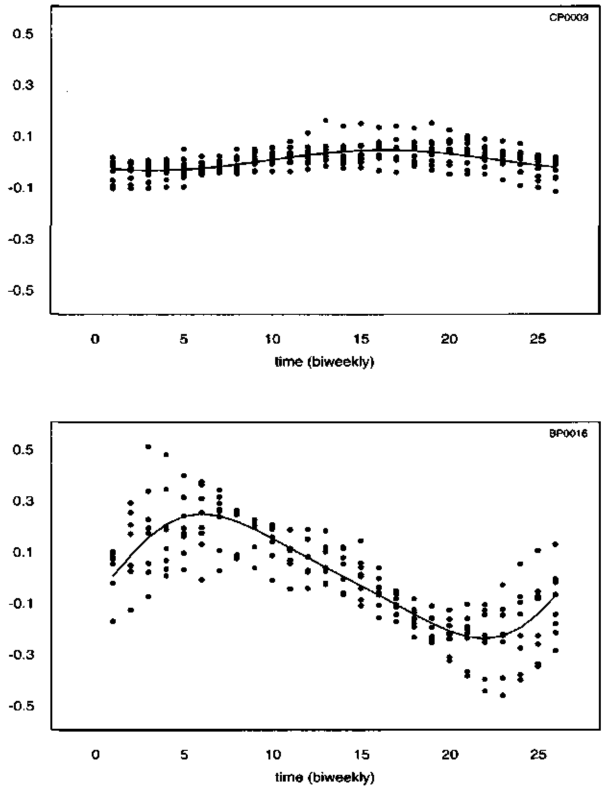


Figure 6.2: *The estimated seasonal component of piezometers BP0016 and CP0003*

(SAS/STAT, 1989). The corresponding estimates of the parameters are: $\hat{A} = 0.0392, \hat{\phi} = 3.9415$ with the skewness parameter ν set to 0, for piezometer CP0003; $\hat{A} = 0.2436, \hat{\phi} = 1.8098$ and $\hat{\nu} = 0.3806$ for piezometer BP0016.

After removal of the seasonal component the average groundwater level from each time series at each piezometer is estimated.

6.3.2 Spatial-temporal variability

Spatial-temporal variability is modelled by the error-term of Equation (6.1). The estimated errors can be calculated as residuals of observations minus the average groundwater level $\mu(s)$ and the seasonal component $seas(s,t)$. The error-term in Equation (6.1) can be correlated in both a spatial and a temporal direction. Estimation of the spatial-temporal variability can be done by an empirical estimation of the spatial-temporal semivariance function. For one single point pair $z(s_1, t_1)$ and $z(s_2, t_2)$ this can be written as:

$$\gamma(h, u) = \frac{1}{2}\mathrm{E}\{\epsilon(s_1, t_1) - \epsilon(s_2, t_2)\}^2 \tag{6.5}$$

where h is the Euclidean distance $h = \|s_1 - s_2\|$ and u the time interval $u = \|t_1 - t_2\|$. Because the spatial-temporal data sets have many observations and consequently many point pairs, the semivariance is estimated for an average distance in space and time. In this way, it is possible to fit a function through these points.

Cressie and Huang (1999) derive classes of nonseparable, spatial-temporal stationary covariance functions, which can be used to model spatial-temporal variability. In this study we apply one of these spatial-temporal covariance functions, formulated as a semivariance function:

$$\gamma(h; u|\theta) = \begin{cases} 0 & \text{if}|u| = |h| = 0 \\ c_0 + c(1 - \exp(-\alpha|u| - \beta^2|h|^2)) & \text{otherwise.} \end{cases} \tag{6.6}$$

where θ is a vector of parameters $\theta = (\alpha, \beta, c, c_0)$. In Cressie and Huang (1999) the equation contains an interaction term between spatial and temporal variability. We exclude this as it does not improve the fit of the function for the data in the case study.

For modelling temporal variability within a piezometer we use the spherical semivariance function, c_1 Sph(a_1):

$$\gamma(u) = \begin{cases} 0, & u = 0 \\ c_1\left\{\frac{3}{2}\left(\frac{u}{a_1}\right) - \frac{1}{2}\left(\frac{u}{a_1}\right)^3\right\}, & 0 \le u \le a_1 \\ c_1, & u > a_1 \end{cases} \tag{6.7}$$

and the Gaussian semivariance function, c_2 Gau(a_2):

$$\gamma(u) = c_2\left\{1 - e^{-\left(\frac{u}{a_2}\right)^2}\right\}, \quad u \ge 0 \tag{6.8}$$

A combination of these two semivariance functions gives a better fit for modelling the temporal semivariance.

Spatial-temporal variability cluster I

The result of an empirical spatial-temporal semivariance estimation in the case of cluster I can be found in Figure 6.3. It shows that the spatial variability is negligible compared to the temporal variability. An estimation, by PROC NLIN in SAS (SAS/STAT, 1989), of parameters of Equation (6.6) to model the spatial-temporal variability is equal to $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{c}, \hat{c_0}) = (0.0003, 0.0000, 1.0087, 0.0000)$. This numerical result confirms that spatial variability can be neglected, as $\hat{\beta} = 0.0000$. Therefore, we assume that spatial-temporal variability can be

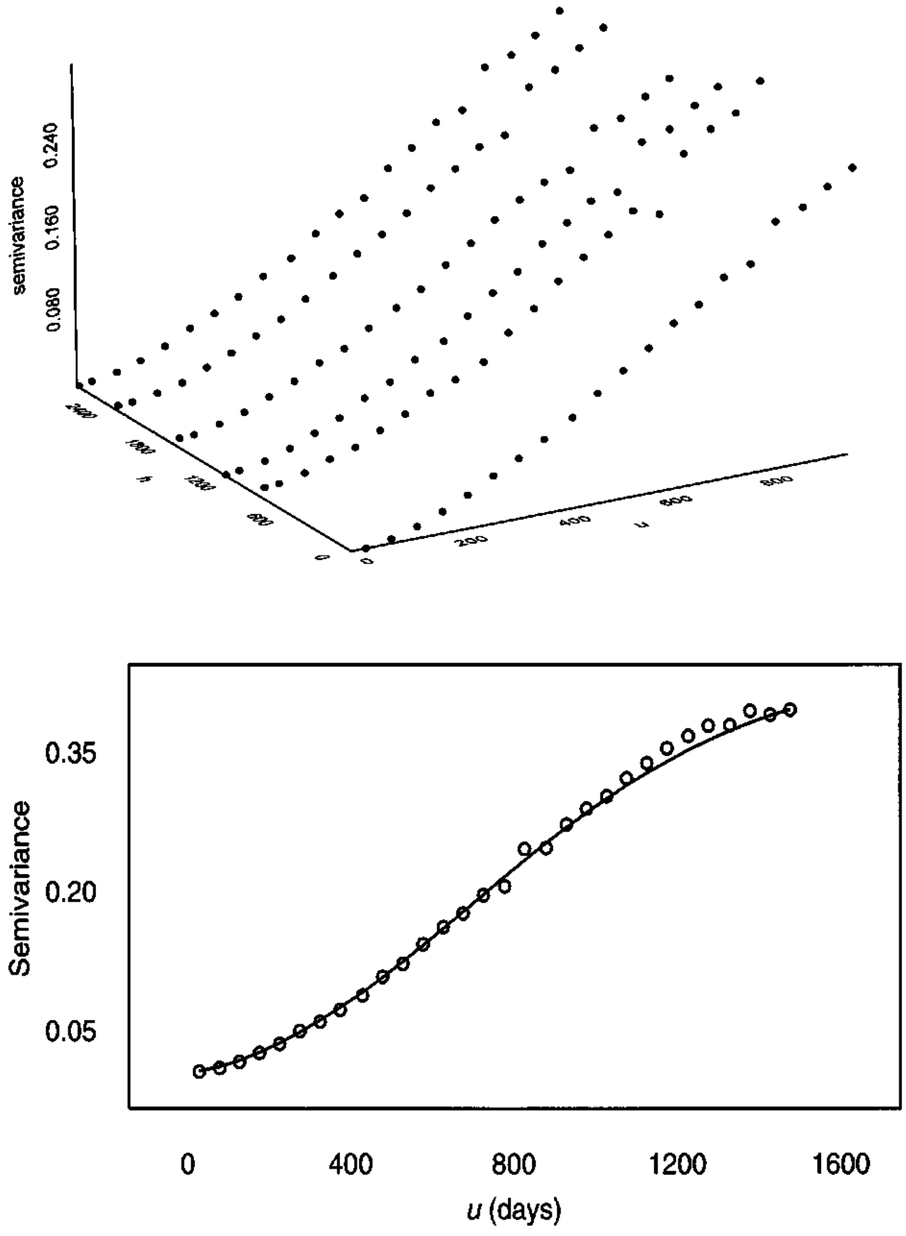


Figure 6.3: *The empirical spatial-temporal semivariance function (top) and the empirical and fitted temporal semivariance function (bottom) for cluster I.*

modelled by the temporal variability only. It turned out that the empirical semivariance could be modelled by a sum of a spherical and a Gaussian semivariance function. The empirical and fitted semivariance function - 0.0123 Sph(377.6) + 0.4199 Gau(968.8) - are shown in Figure 6.3.

Spatial-temporal variability cluster II

The analysis is repeated for cluster II. Figure 6.4 shows the empirical spatial-temporal semivariance function. Spatial variability is not negligible in this case as the empirical spatial-temporal semivariance increases along the *h*-axis. Equation (6.6) is used to model the spatial-temporal variability. The nugget (c_0) is difficult to estimate because of lack of spatial distance information. Therefore we estimated the nugget from the temporal variability. A fit of this function through the empirical semivariance function results into the following estimations of parameters: $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{c}, \hat{c_0}) = (0.00195, 0.00018, 0.1431,\ 0.0000)$. A plot of this function can be found in Figure 6.4.

6.4 Reduction of the number of measurements

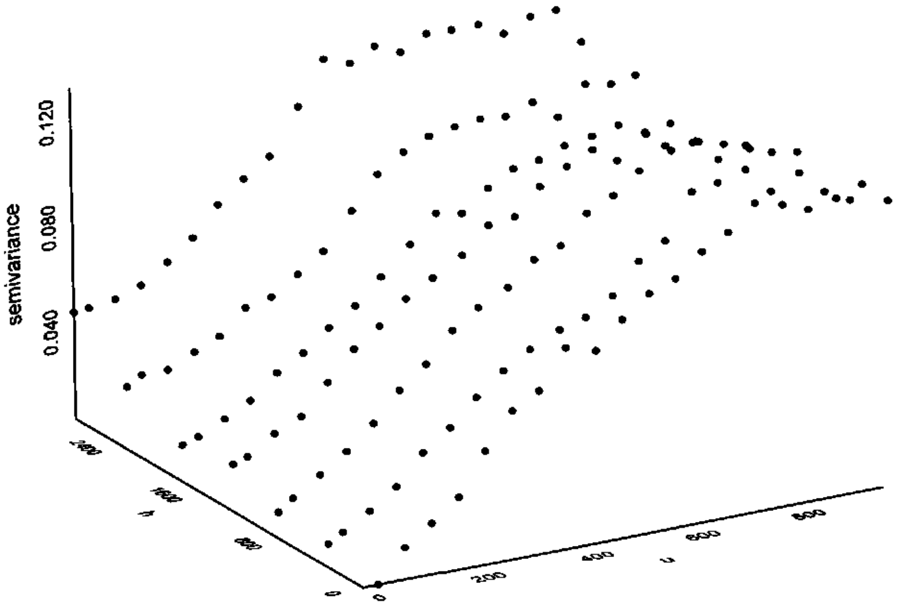
In this section two different ways of reducing the number of measurements of the monitoring network are discussed. In the first place, we concentrate on temporal variability and the consequences for the monitoring frequency within a piezometer. Therefore, information at other piezometers is not taken into account for the predictions. Secondly, the problem is discussed in a spatial-temporal perspective.

6.4.1 Reduction of frequency within a piezometer

For this investigation we selected two piezometers: CP0003 from cluster I and BP0073 from cluster II (see Figure 6.1). For each piezometer an unconditional simulation is done. The simulated time series are used to quantify consequences of a reduction of frequency of measuring. Weekly measures are simulated on the basis of estimated temporal semivariance functions. Prediction results of three monitoring frequencies (every two, four and eight weeks) are compared by means of a validation set out of the simulated data set. The whole procedure for one piezometer of estimating the Mean Squared (prediction) Error (MSE) and the Maximum Absolute prediction Error (MAE) at different frequencies is described as follows:

1. Estimate the temporal semivariance function of the error-term at the piezometer on the basis of the original data.

2. Simulate weekly errors with unconditional Gaussian simulation on the basis of the estimated semivariance function. Simulate groundwater levels by adding the estimated seasonal component and average value of groundwater level of the piezometer (from the original data).



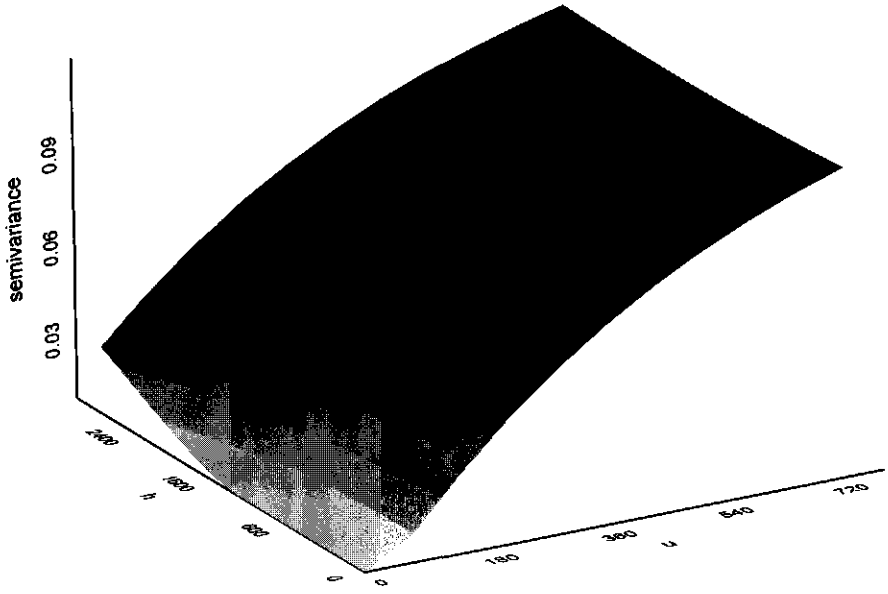


Figure 6.4: *The empirical spatial-temporal semivariance function (top) and the fitted function (bottom) for cluster II.*

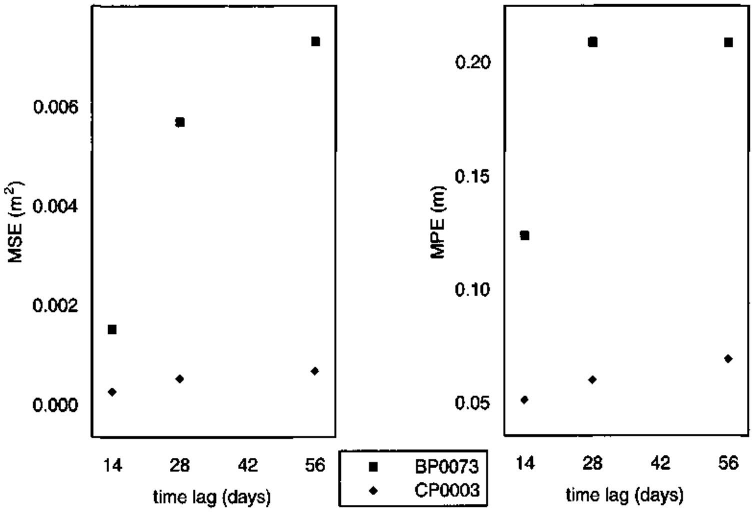


Figure 6.5: *The Mean Square Error (MSE) and Maximum Absolute Prediction Error (MAE) of prediction of 50 validation points at three different measuring frequencies for simulated time series of the piezometers BP0073 and CP0003.*

3. Make three sample data sets of three different monitoring frequencies (every two, four and eight weeks).
4. Sample a validation set of approximately 10% of the total simulated values.
5. Estimate for the three sample data sets the seasonal component and the average of the simulated values. Followed by the estimation of the semivariance function of the error-term.
6. Perform ordinary kriging of the errors at the validation locations by means of the three sample data sets and the estimated semivariance functions.
7. Predict for every frequency the groundwater level at the validation locations; predicted error + seasonal component + average value.
8. Compare simulated values with predicted values and calculate for every frequency the MSE and MAE.

Figure 6.5 shows MSE and MAE values of 50 validation points at three measuring frequencies by means of simulated time series at piezometers BP0073 and CP0003. The results are intuitive clear; the MSE and MAE will increase as the measuring frequency decrease. For CP0003 this increase in prediction error is low. A reduction of measuring frequency at piezometer CP0003 has little consequences for the prediction accuracy. On the principle that the spatial variability in cluster I is negligible, we conclude that the measuring frequencies of all piezometers in cluster I can be reduced.

Consequences of reducing the measuring frequency at BP0073 are more severe. MSE and MAE values are considerably higher at lower measuring frequencies. Short distance fluctuations will be lost at lower measuring frequencies. However, the spatial information (i.e. measurements at other piezometers in the neighbourhood) is neglected in this analysis. Therefore, we look at the spatial-temporal monitoring with the help of a spatial-temporal semivariance function.

6.4.2 Reducing the monitoring network from a spatial-temporal perspective

We first consider cluster I. The spatial variability is not modelled because it is negligible compared to the temporal variability. Strictly speaking, one piezometer suffices to characterize the time series of the nine piezometers, because of the high correlation between time series. However, a reasonable cover of the area of interest might still be desirable to monitor possible temporal changes in cluster I.

The correlation between the time series of the different piezometers in cluster II is much less than in cluster I. Therefore, the spatial variability can not be neglected as in cluster I. The spatial-temporal variability also shows that. To see if a reduction of the number of measurements in this cluster is possible we compare 9 different spatial-temporal monitoring strategies. The comparison of these 9 different strategies is based on the kriging variance, like many papers already did in the spatial setting (e.g. McBratney and Webster (1981), Van Groenigen (1999), Prakash and Singh (2000)).

Kriging variance is calculated on a grid of points. The points are at the 14 locations of cluster II every week for 10 years, in total 7294 points. The monitoring strategies are denoted as: S1,S2,...,S9. The monitoring strategies S1, S2 and S3 measure on every 14 locations every two, four and eight weeks respectively. Figure 6.1 shows that the locations in cluster II can easily be divided in pairs of two locations. For the strategies S4, S5 and S6, one of the locations of each pair is treated as in the S1, S2 and S3 strategy. At the other 7 locations, 6 measurements are done each year. The last three strategies (S7, S8 and S9) are equal to S4, S5 and S6 respectively, accept that the 6 yearly measures are reduced to none measures at these 7 piezometers. In Table 6.2 the total number of measurements, the average and the maximum kriging variance are presented for all 9 monitoring strategies.

Table 6.2 shows a number of comparisons between the 9 monitoring strategies. Strategies S1, S2 and S3 differ only in measuring frequency at the 14 locations. It shows that the average and maximal kriging variance approximately double if the measuring frequency is halved. Secondly, a reduction of the number of measurements at only 7 locations (comparison of S1, S4 and S7) is compared, whereby at the other 7 locations the frequency of measuring is every two weeks. We see that a reduction of the number of measurements at the 7 locations,

Table 6.2: *The number of measurements (N), the average kriging variance and the maximum kriging variance at a grid of points for 9 different spatial-temporal monitoring strategies.*

Monitoring strategy	N	Av. krig. var.	Max. krig. var.
S1	3654	0.0010	0.0020
S2	1834	0.0024	0.0039
S3	924	0.0051	0.0078
S4	2254	0.0010	0.0020
S5	1344	0.0021	0.0039
S6	889	0.0036	0.0078
S7	1827	0.0010	0.0021
S8	917	0.0025	0.0041
S9	462	0.0051	0.0078

to even none measurements by S7, hardly influence the average and maximal kriging variance. Furthermore, Table 6.2 shows that an alternating design, i.e. measuring at different times at the two groups of 7 piezometers, can result into lower average kriging variances. This can be seen from S3 and S6. However, such an alternating design might not be desirable because of practical reasons. On the basis of the results summarized in Table 6.2 we can conclude that the number of piezometers in cluster II can be reduced to 7, without losing much information of groundwater level in the area. This means a reduction of 50% of the number of measurements of S1.

6.5 Discussion and conclusion

This paper discusses the possible reduction of a monitoring network for measuring groundwater level. The temporal variability at the piezometers differs mainly as a consequence of the thickness of the unsatured zone. For a thick unsatured zone the temporal variability will be characterized by long term variability. At piezometers in an area with a relatively thin zone, the time series are characterized by short range temporal variability. This heterogeneity has to be taken into account. In this study, heterogeneity is reduced by clustering piezometers. These clusters are piezometers with correlated time series of groundwater level. This study shows that within a cluster, a reduction of the number of measurements is possible.

The spatial-temporal empirical semivariance function quantifies the spatial and temporal variability within a cluster. For cluster I, the temporal variability is the main source of spatial-temporal variability. This means that the time series at the different locations show a comparable behaviour in time. The amount of information of groundwater level will not decrease much if piezometers are removed. However, it can be politically desirable to cover the area with piezometers to see possible local changes in time of groundwater level. The measuring frequency at a piezometer can decrease, because the time series are fairly smooth. A same analysis for cluster II results into the conclusion that half of the piezometers can be removed, without losing much information of groundwater level in cluster II.

Acknowledgements

The author would like to thank A. Stein and B. van Putten for their useful comments on manuscripts of this paper. The data are kindly provided by NV NUON Water in cooperation with the engineering firm Witteveen+Bos.

Chapter 7

General discussion and conclusions

This work shows development and application of various statistical techniques to real-world case studies. The main subjects are spatial(-temporal) interpolation and optimization of monitoring networks. Optimal design of experiments is discussed and applied to a spatial setting. The objective formulated in Section 1.4 results into three aims of research. In this final chapter the results and conclusions of Chapters 2-6 are compared with the three formulated aims. Each of these aims will be repeated, followed by the main results and conclusions related to these aims:

Study statistical methods for spatial(-temporal) interpolation.

In this thesis three categories of spatial interpolation techniques are discussed: geostatistical techniques (kriging), thin plate splines and locally weighted regression. In Chapter 2 both several forms of kriging and of thin plate splines are discussed. Additional information (the introduction of a covariable) improves the prediction accuracy. Different ways of including additional information are compared on prediction accuracy. Three-dimensional approaches, i.e. additional information considered as a third dimension, appeared to have the best prediction accuracy for our case study. This approach is well-known in the application of thin plate splines, but less in a geostatistical framework. However, three-dimensional kriging outperformed (frequently applied) geostatistical methods as cokriging and regression-kriging. A nonparametric interpolation technique is used in Chapter 5. By applying locally weighted regression, the modelling of trend and spatial variability is avoided. This method is especially useful if stationarity is hard to assume.

Spatial modelling is still in development, spatial-temporal modelling is even more developing. An example in the recent past is the development of a class of nonseparable spatial-temporal semivariance functions. Chapter 6 shows an application of such a function. Note that these nonseparable spatial-temporal semivariance functions can also be used in Chapter 2 by the three-dimensional form of kriging. It is assumed that the semivariance function is stationary in space and time. To meet this assumption as much as possible, we proposed analyses for clusters of measuring points. Furthermore, the seasonality in time and levels at a location of measuring are removed for the estimation of the spatial-temporal variability.

Actual optimization of environmental monitoring networks for different criteria

Chapters 4, 5 and 6 are dealing with optimization of existing monitoring networks for several criteria. This means that data from the past can be used to model trend and spatial(-temporal) variability.

Chapter 4 shows that the theory of optimal design of experiments can be helpful to formulate a criterion for optimizing a monitoring network for estimation of the semivariance function. The practical use of this criterion is doubted because of extreme clustering of points. However, Chapter 4 shows that this problem can be solved, if correlation between empirical semivariances is taken into account.

In Chapter 5 a reduction of a monitoring network is investigated. A criterion based on prediction variances of locally weighted regression is used to reduce an existing monitoring network. The method described can easily be adapted for enlarging monitoring networks. Furthermore, Chapter 5 shows that statistical criteria can be combined with political wishes for monitoring.

In Chapters 4 and 5 the spatial configuration of measuring locations in a monitoring network is investigated for annual averages. Both frequency of measuring at a location and spatial configuration of locations are investigated in Chapter 6. The empirical spatial-temporal semivariance function shows the variability in space and time and is modelled by a nonseparable spatial-temporal semivariance function. This function is a useful tool for optimizing monitoring networks.

Development and application of algorithms necessary to optimize the monitoring networks

Chapter 3 focuses on global optimization in optimal design of experiments. Several ways of reaching the global optimum are discussed for various situations. An algorithm is developed which is useful if a limited number of candidate points for the design is considered. This branch-and-bound algorithm is described in the appendix of this thesis. It can be applied for optimal design of experiments in regression models, but is also applicable in other situations. Chapter 4 gives some results with this algorithm for optimizing a monitoring network for estimation of the semivariance function. In this case, full enumeration of all possible combinations of monitoring sites is possible. However, in many circumstances the number of combinations is too large to enumerate completely. Therefore, search algorithms are needed. Two examples are given in Chapter 5, one is a simple algorithm and the other is a more advanced algorithm. The computation time of the simple algorithm is much less than the more advanced algorithm. The additional effort in both developing the algorithm and computation time gives only a very small improvement of the value of the criterion.

Appendix: Branch-and-bound algorithm

In this thesis several combinatorial problems are solved by a branch-and-bound algorithm. The basic problem consists of the selection of n points out of N possible (or candidate) points optimizing a certain criterion. Optimality criteria in this thesis are all based on variances or prediction errors and optimization always means minimization. This appendix is based on Rasch *et al* (1997).

We consider the set

$$B = B_N = \{x_1, x_2, \ldots, x_N\},$$

where x_i are candidate points. Let X be a subset of B. The criterion value for design X is denoted by $\Phi(X)$. We will assume that leaving points out of a design will not decrease the criterion value Φ, i.e.

$$X \subseteq Y \subsetneq B \to \Phi(X) \geq \Phi(Y) \geq \Phi(B) \tag{1}$$

To find an optimal n-point design in B_N we only have to calculate the value of the criterion for all $\binom{N}{n}$ possible subsets with size n of B_N and select a subset which minimizes the outcome of the criterion Φ. If there are several subsets giving the same minimal outcome of the criterion, one of them can be selected from a practical point of view. This sounds easy but if N becomes large, even a high speed computer needs much time. Quick algorithms or theoretical results are needed to solve these problems. Here an example of a quick algorithm will be presented.

Fast enumeration

The problem described in the last paragraph implies in the crudest form the generation and evaluation of all possible $\binom{N}{n}$ designs. A straightforward implementation, which was used by Müller (1994) and Kraan (1995), typically required hours to generate and evaluate all designs for a specific test case in regression analysis. Focus in this appendix is on the acceleration of the enumeration of all n point designs. The first improvement was found by formulating a recursive procedure which leads to savings on the calculation time for the determination of the criterion value. The procedure is given as follows.

```
PROC(min, max, $l$, $Q$, $X_{l-1}$)
for $i$ = min to max do
    add $x_i$ to design $X_{l-1}$ resulting in a design $X_l$
    $Q^*$ = updated $Q$
    If $l = n$ then
        Calculate $\Phi(X_l)$
        If $X_l$ is the best design found, save it.
    Else
        PROC($i$+1, max+1, $l$+1, $Q^*$, $X_l$).
```

In procedure PROC, the variable X_l represents an l-point design (not complete if $l < n$), the variable Q represents all information which is needed to calculate the criterion function. In general this consists of a summation of operations on the points x_i of the design. Actually the sequentially updating of this information generates the savings in calculation time. Due to its recursive structure, it is hard to understand directly the consequences in terms of time and memory of calling the procedure PROC.

The procedure implicitly generates (part of) a tree of which the end nodes represent the n-point designs. Level l defines the levels of the tree. Therefore we apply the name *tree-algorithm* for generating the designs with the aid of procedure PROC.

Tree algorithm

0. Given N,n and $B_N = \{x_1, \ldots, x_N\}$
   Initiate $X_0 = \emptyset$, Q contains zeros only.
1. PROC(1,$N - n + 1$,1,Q,X_0)

On the first level, $l = 1$, the first point of the design is selected. In the example of Figure A these are the points 1,2 or 3. On the second level the next point is added to the design. After adding a point, its contribution is calculated and added to the information Q. Note that that the tree is asymmetric and its end nodes will represent all $\binom{N}{n}$ designs.

The gain in calculation time is considerable. One of the causes is that the elements of Q are not calculated only at the end nodes of the tree. At first sight the difference between the tree algorithm and a straightforward enumeration looks like a tradeoff between calculation time and computer memory.
Namely, at every call of the procedure, computer memory is occupied to store

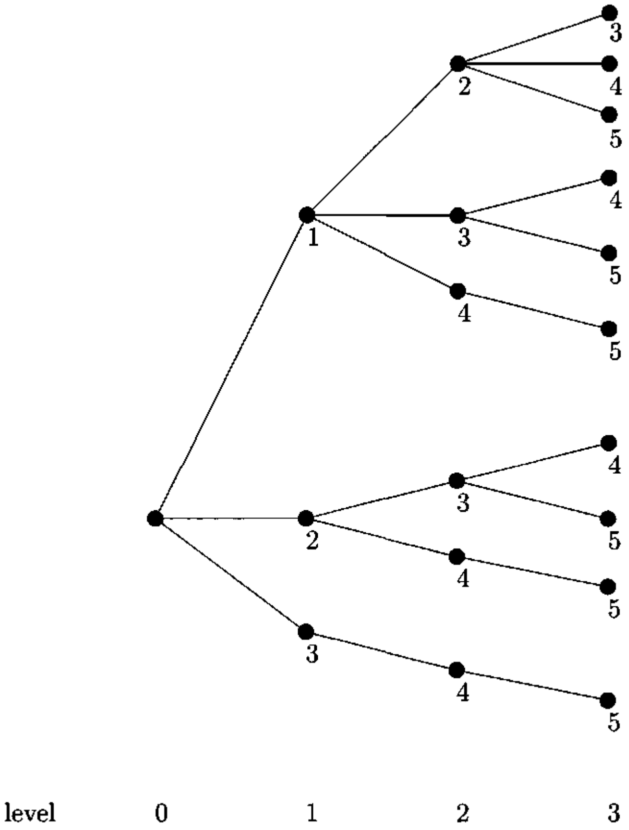


Figure A: *Tree algorithm for* $N = 5$, $n = 3$ *with* $B_N = \{1, 2, 3, 4, 5\}$

Q and X_l. The tree has

$$\sum_{l=0}^{n} \binom{N-n+l}{l} = \binom{N+1}{n}$$

nodes corresponding to calls of the procedure. When the tree is generated according to a level by level strategy (see Reingold *et al*, 1977), this leads to an explosion of required memory with increasing N. However, procedure PROC implies a depth first search strategy so that the number of nodes stored grows linear in n.

Adaptations of the algorithms can be found to decrease the calculation time. It is easier to enumerate the combinations which are left out of the design when

$$\binom{N+1}{n} > \binom{N+1}{N-n}, \text{i.e. } n > \frac{N}{2}.$$

Because we will use this concept further in the development, a specific algorithm is formulated which is called the *complement tree* algorithm. The corresponding recursive procedure is formulated as follows:

```
CPROC(min, max, $l$, $Q$, $X_{N-l+1}$)
for $i =$ min to max do
    drop $x_i$ from design $X_{N-l+1}$ resulting in a design $X_{N-l}$
    $Q^*$ = updated $Q$
    If $l = N - n$ then
        Calculate $\Phi(X_{N-l})$
        If $X_{N-l}$ is the best design found, save it.
    Else
        CPROC($i$+1, max+1, $l$+1, $Q^*$, $X_{N-l}$).
```

The complement tree algorithm starts with all candidate points $X_N = B_N$, and proceeds calculating with overcomplete designs until an n-points-design has been generated.

Complement tree algorithm

0. Given N,n and $B_N = \{x_1, \ldots, x_N\}$
   Initiate $X_N = B_N$, Q containing the information of the full design with all candidate points.
1. CPROC(1, $n+1$, 1, Q, X_N)

The points in Figure B correspond to the candidate points which are left out of the design. Notice that every node in the tree corresponds to an overcomplete design. By dropping a point from the design, the criterion value will not decrease. Leaving points out of a design never makes it better.

This monotonicity and the complement tree algorithm will be used to derive an even faster algorithm, the branch-and-bound algorithm. The concept of this algorithm is known in literature on combinatorial optimization (see Lawler *et al*, 1985). The idea of branching may be clear from the tree algorithm. The concept of bounding is based on the assumption that Φ is monotically non-decreasing (see equation (1)) when points are left out and is now explained with the example of Figure B.

Assume that a very good 3-point design { 1,3,5 } has been found for the example of Figure A or B. It has a criterion value of say 0.3. Now the complement tree algorithm is started to validate the optimality of the design found. At the first step of the iteration the point $x_1 = 1$ is left out of the design. If the criterion value of the overcomplete design { 2,3,4,5 } is larger than the best criterion value for a 3-point design which has already been found (0.3 in this case), then searching further from this node does not make any sense. In this case it is not necessary to call the procedure CPROC and start dropping other points, as the criterion value is not going to decrease and reach a value below 0.3. The cutoff of a branch of the tree in this way is called bounding. The best value found so

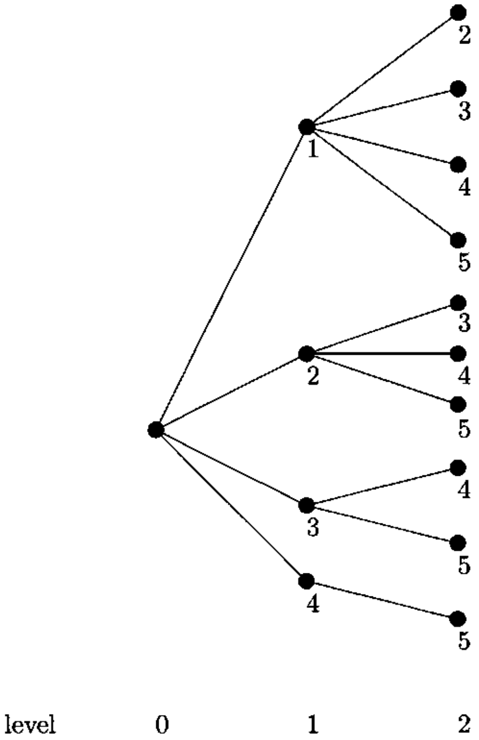


Figure B: *Complement tree algorithm for* $N = 5$, $n = 3$ *with* $B_N = \{1, 2, 3, 4, 5\}$

far is called the *(upper) bound.* The Branch-and-Bound procedure is formulated as follows:

```
BPROC(min, max, $l$, $Q$, $X_{N-l+1}$)
for $i =$ min to max do
    drop $x_i$ from design $X_{N-l+1}$ resulting in design $X_{N-l}$
    $Q^*$ = updated $Q$
    calculate $\Phi(X_{N-l})$
    If $l = N - n$ then
        If $\Phi(X_{N-l}) <$ bound then
            bound := $\Phi(X_{N-l})$
            save $X_{N-l}$
    Else
        If $\Phi(X_{N-l}) <$ bound then
            BPROC($i+1$, max+1, $l$+1, $Q^*$, $X_{N-l}$)
```

The branching is only continued, when the criterion value of the overcomplete design is smaller than the bound. The bound can either be initialized on infinity or be based on a n-point design which is found by another (search) algorithm.

The branch-and-bound algorithm is now formalized as follows:

Branch-and-Bound algorithm

```
0.    Given $N$,$n$ and $B_N = \{x_1, \ldots, x_N\}$
      Initiate the bound: bound = $\infty$
      $X_N = B_N$ , $Q$ contains the information of the
      full design with all candidate points.
1.    BPROC(1, $n+1$, 1, $Q$, $X_N$)
```

Implementation of this algorithm reduces the calculation time of an optimal design considerably. Within this branch-and-bound framework various strategies are possible, which may improve the performance. Due to the asymmetry of the tree, the larger parts can be found earlier in the tree. If bounding takes place there, larger parts of the tree are cut off. The bounding is more likely to take place when points which actually should be in the design, have a positive contribution to the criterion value, are dropped. This means that ordering the points in B_N according to their contribution to the criterion value, influences the performance of the algorithm. This sorting procedure can be performed at the root, but may also be repeated at other nodes. The sorting is only effective if a good bound (from a search algorithm) is available.

The computation time for finding an optimal design by full enumeration has been reduced considerably. However, the computation time for fixed n remains polynomial in N. For larger problems, e.g. when the candidate points are taken from a higher dimensional space, the search for the optimum may be very time consuming. Therefore algorithms have been developed for which the calculation time grows much less and which generate good but not necessarily optimal designs.

Bibliography

Atkinson, A.: 1996, The usefulness of optimum experimental designs, *Journal of the Royal Statistical Society B* **58**(1), 59–76.

Atkinson, A. and Donev, A.: 1992, *Optimum experimental designs*, Oxford University Press, Oxford.

Batschelet, E.: 1981, *Circular statistics in biology*, Academic Press, London.

Beek, E., Stein, A. and Janssen, L.: 1992, Spatial variability and interpolation of daily precipitation amount, *Stochastic Hydrology and Hydraulics* **6**, 304–320.

Bleeker, A. and Den Hartog, P.: 1995, *Rapportage besluiten luchtkwaliteit over het jaar 1993*, Staatuitgeverij, Den Haag.

Boer, E., Rasch, D. and Hendrix, E.: 2000, Locally optimal designs in non-linear regression: A case study of the Michaelis-Menten function, *in* N. Balakrishnan, S. Ermakov and V. Melas (eds), *Stochastic Simulation Methods*, Birkhäuser, Boston, chapter 11, pp. 177–188.

Bogaert, P., Mahau, P. and Beckers, F.: 1995, Cokriging software and source code, *Technical Report 12*, FAO, Rome.

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.

Brockwell, P. and Davis, R.: 1991, *Time Series: Theory and Methods*, Springer-Verlag, New York.

Burrough, P. and McDonnell, R.: 1998, *Principles of geographical information systems*, Oxford University Press, New York.

Caselton, W. and Zidek, J.: 1984, Optimal monitoring network designs, *Statistics & Probability Letters* **2**, 223–227.

Chilès, J. and Delfiner, P.: 1999, *Geostatistics: modeling spatial uncertainty*, John Wiley & Sons, New York.

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

Cressie, N.: 1985, Fitting variogram models by weighted least squares, *Journal of the International Association for Mathematical Geology* **17**, 563–586.

Cressie, N.: 1991, *Statistics for spatial data*, John Wiley & Sons, New York.

Cressie, N., Gotway, C. and Grondona, M.: 1990, Spatial prediction from networks, *Chemometrics and Intelligent Laboratory Systems* **7**, 251–271.

Cressie, N. and Huang, H.: 1999, Classes of nonseparable, spatio-temporal stationary covariance functions, *Journal of the American Statistical Association* **94**(448), 1330–1340.

D'Agostino, V., Greene, E., Passarella, G. and Vurro, M.: 1998, Spatial and temporal study of nitrate concentration in groundwater by means of coregionalization, *Environmental Geology* **36**(3-4), 285–295.

De Beurs, K.: 1998, Evaluation of spatial interpolation techniques for climate variables. Case study of Jalisco, Mexico, *Technical report*, Wageningen University, Sub-department of Mathematics.

De Cesare, L., Myers, D. and Posa, D.: 2001, Product-sum covariance for spatial-temporal modelling: an environmental application, *Environmetrics* **12**, 11–23.

De Gruijter, J. and Ter Braak, C.: 1990, Model-free estimation from spatial samples: a reappraisal of classical sampling theory, *Mathematical Geology* **22**(4), 407–415.

Dirks, K., Hay, J., Stow, C. and Harris, D.: 1998, High-resolution studies of rainfall on Norfolk Island. Part II: Interpolation of rainfall data, *Journal of Hydrology* **208**, 187–193.

Doesburg, M., Rentinck, E. and Swaan, P.: 1994, Landelijk Meetnet Luchtkwaliteit. Meetresultaten 1993. Deel 1-4, *Technical Report 722101010*.

Dufour, F.: 1998, *Grondwater in Nederland; onzichtbaar water waarop wij lopen*, NITG-TNO, Delft.

Ermakov, S. and Zhigljavsky, A.: 1987, *Matematitscheskaja teorija optimalnich experimentov*, Nauka, Moskva.

Fedorov, V.: 1972, *Theory of optimal experiments*, Academic Press, New York.

Fedorov, V.: 1989, Optimal design with bounded density: Optimization algorithms of the exchange type, *Journal of Statistical Planning and Inference* **22**, 1–13.

Fedorov, V., Montepiedra, G. and Nachtsheim, C.: 1999, Design of experiments for locally weighted regression, *Journal of Statistical Planning and Inference* **81**, 363–382.

Gaffke, N. and Heiligers, B.: 1995, Algorithms for optimal design with application to multiple polynomial regression, *Metrika* **42**, 173–190.

Gaffke, N. and Mathar, R.: 1992, On a class of algorithms from experimental design theory, *Optimization* **24**, 91–126.

Gehrels, J.: 1999, *Groundwater level fluctuations*, PhD thesis, Vrije Universiteit Amsterdam.

Goodale, C., Aber, J. and Ollinger, S.: 1998, Mapping monthly precipitation, temperature, and solar radiation for Ireland with polynomial regression and a digital elevation model, *Climate Research* **10**, 35–49.

Goovaerts, P.: 2000, Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall, *Journal of Hydrology* **228**, 113–129.

Gotway, C. and Hartford, A.: 1996, Geostatistical methods for incorporating auxiliary information in the prediction of spatial variables, *Journal of Agricultural, Biological, and Environmental Statistics* **1**, 17–39.

Hartkamp, A., Hoogenboom, G. and White, J.: 1998, Adaptation of the CROPGRO growth model to velvet bean as a green manure cover crop, *Agronomy Abstracts* **64**.

Heuvelink, G., Musters, P. and Pebesma, E.: 1997, Spatio-temporal kriging of soil water content, *in* E. Baafi and N. Schofield (eds), *Geostatistics Wollongong '96*, Vol. 2, Kluwer Academic Publishers, Dordrecht, pp. 1020–1030.

Hutchinson, M.: 1995, Stochastic space-time weather models from groundbased data, *Agricultural and Forest Meteorology* **73**, 237–264.

Hutchinson, M.: 1997, *ANUSPLIN version 3.2.*, Center for resource and environmental studies, The Australian National University, Canberra, Australia.

Hutchinson, M.: 1998, Interpolation of rainfall data with thin plate smoothing splines: II Analysis of topographic dependence, *Journal of Geographic Information and Decision Analysis* **2**, 168–185.

IMTA: 1996, *Extractor Rápido de Información Climatológica*, Instituto Mexicano de Technologiá del Agua, Morelos Mexico.

Isaaks, E. and Srivastava, R.: 1989, *Applied geostatistics*, Oxford University Press, New York.

Jones, B. and Wang, J.: 1999, Constructing optimal designs for fitting pharmacokinetic models, *Computational Statistics* **9**, 209–218.

Kent, J. and Mardia, K.: 1994, The link between kriging and thin plate splines, *in* F. Kelly (ed.), *Probability, Statistics and Optimisation*, John Wiley & Sons, Chichester, chapter 24, pp. 325–339.

Kiefer, J.: 1959, Optimum experimental design, *Journal of the Royal Statistical Society Series B* **21**, 272–319.

Kiefer, J. and Wolfowitz, J.: 1960, The equivalence of two extremum problems, *Canadian Journal of Mathematics* **12**, 363–366.

Knotters, M.: 2001, *Regionalised time series models for water table depths*, PhD thesis, Wageningen University.

Ko, C., Lee, J. and Queyranne, M.: 1995, An exact algorithm for maximum entropy sampling, *Operations Research* **43**(4), 684–691.

Kraan, P.: 1995, Het beoordelen van zoekalgoritmen voor optimale herhalingsvrije proefopzetten in logistische regressie, *Technical report*, Wageningen University.

Krige, D.: 1951, A statistical approach to some basic mine valuation problems on the Witwatersrand, *Journal of the Chemical, Metallurgical and Mining Society South Africa* **52**, 119–139.

Laslett, G. and McBratney, A.: 1990, Further comparison of spatial methods for predicting soil ph, *Soil Science Society America Journal* **54**, 1553–1558.

Lawler, E., Lenstra, J. and Rinnooy Kan, A.: 1985, *The travelling salesman problem: a guided tour of combinatorial optimization*, Wiley, New York.

Littell, R., Milliken, G., Stroup, W. and Wolfinger, R.: 1996, *SAS System for Mixed Models*.

Matheron, G.: 1963, Principles of geostatistics, *Economic Geology* **58**, 1246–1266.

McBratney, A. and Webster, R.: 1981, The design of optimal sampling schemes for local estimation and mapping of regionalized variables: 2. Programs and examples, *Computers and Geosciences* **7**, 335–365.

Müller, R.: 1994, Konkrete lokal ψ-optimale, wiederholungsfreie Versuchsplänne für speziell isotone nichtlineare Regressionfunktionen, *Technical report*, University Dortmund.

Müller, W.: 1995, Optimal design for local fitting, *Journal of Statistical Planning and Inference* **55**, 389–397.

Müller, W.: 1998, *Collecting spatial Data - Optimum design of experiments for random fields*, Physica-Verlag, Heidelberg.

Müller, W.: 1999, Least-squares fitting from the variogram cloud, *Statistics & Probability Letters* **43**, 93–98.

Müller, W. and Pázman, A.: 1998, Design measures and approximate information matrices for experiments without replications, *Journal of Statistical Planning and Inference* **71**, 349–362.

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

Odeh, I., McBratney, A. and Chittleborough, D.: 1995, Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging, *Geoderma* **67**, 215–226.

Pardo-Igúzquiza, E.: 1998a, Comparison of geostatistical methods for estimating the areal average climatological rainfall mean using data on precipitation and topography, *International Journal of Climatology* **18**, 1031–1047.

Pardo-Igúzquiza, E.: 1998b, Optimal selection of number and location of rainfall gauges for areal rainfall estimation using geostatistics and simulated annealing, *Journal of Hydrology* **210**, 206–220.

Pebesma, E. and Wesseling, C.: 1998, Gstat: a program for geostatistical modelling, prediction and simulation, *Computer & Geosciences* **24**(1), 17–31.

Pfeifer, P. and Deutsch, A.: 1980, A three-stage iterative procedure for space-time modelling, *Technometrics* **22**, 35–47.

Prakash, M. and Singh, V.: 2000, Network design for groundwater monitoring - a case study, *Environmental Geology* **39**(6), 628–632.

Pukelsheim, F.: 1993, *Optimal design of experiments*, John Wiley & Sons, New York.

Pukelsheim, F. and Rieder, S.: 1992, Efficient rounding of approximate designs, *Biometrika* **79**(4), 763–770.

Rasch, D.: 1990, Optimum experimental design in nonlinear regression, *Commun. Statist.-Theory Meth.* **19**(12), 4789–4806.

Rasch, D., Hendrix, E. and Boer, E.: 1997, Replication-free optimal design in regression analysis, *Computational Statistics* **12**, 19–52.

Reingold, E., Nievergelt, J. and Deo, N.: 1977, *Combinatorial algorithms: theory and practice*, Prentice Hall, New Yersey.

Rosenthal, W., Hammer, G. and Butler, D.: 1998, Predicting regional grain sorghum production in Australia using spatial data and crop simulation modelling, *Agricultural and Forest Meteorology* **91**, 263–274.

Roth, C.: 1998, Is lognormal kriging suitable for local estimation ?, *Mathematical Geology* **30**(8), 999–1009.

Rouhani, S. and Hall, T.: 1989, Space-time kriging of groundwater data, *in* M. Armstrong (ed.), *Geostatistics*, Vol. 2, Kluwer Academic Publishers, Dordrecht, pp. 639–651.

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

SAS: 1989, *SAS/STAT Users's Guide*, 4 edn, North Carolina.

Silvey, S.: 1980, *Optimal design*, Chapman and Hall, London.

Sølna, K. and Switzer, P.: 1996, Time trend estimation for a geographic region, *Journal of the American Statistical Association* **91**(434), 577–589.

Stein, A.: 1999, Geostatistical procedures for analysing spatial variability and optimizing collection of monitoring data, *in* H. van Lanen (ed.), *Monitoring for groundwater management in (semi-)arid regions*, UNESCO Studies and Reports in Hydrology, Paris, pp. 91–106.

Stein, A., Van Groenigen, J., Jeger, M. and Hoosbeek, M.: 1998, Space-time statistics for environmental and agricultural related phenomena, *Environmental and Ecological Statistics* **5**, 155–172.

USGS: 1997, *GTOPO-30: a digital elevation model at 30-arc second scale*, South Dakota, USA.

Van Groenigen, J.: 1999, *Constrained optimisation of spatial sampling, a geostatistical approach*, PhD thesis, Wageningen University.

Van Groenigen, J. and Stein, A.: 1998, Constrained optimisation of spatial sampling using continuous simulated annealing, *Journal of Environmental Quality* **27**, 1078–1086.

Van Groenigen, J., Stein, A. and Zuurbier, R.: 1997, Optimisation of environmental sampling using interactive gis, *Soil Technology* **10**, 83–97.

Vila, J.: 1991, Local optimality of replications from a minimal D-optimal design in regression: A sufficient and quasi-necessary condition, *Journal of Statistical Planning and Inference* **29**, 261–277.

Wackernagel, H.: 1995, *Multivariate geostatistics*, Springer-Verlag, Berlin.

Wahba, G.: 1990, *Spline models for observational data*, SIAM, Philadelphia.

Warrick, A. and Myers, D.: 1987, Optimization of sampling locations for variogram calculations, *Water Resources Research* **23**(3), 496–500.

Welch, W.: 1982, Branch-and-bound search for experimental designs based on D-optimality and other criteria, *Technometrics* **24**(1), 41–48.

White, L.: 1973, An extension to the general equivalence theorem for nonlinear models, *Biometrika* **60**, 345–348.

Wikle, C., Berliner, M. and Cressie, N.: 1998, Hierarchical bayesian space-time models, *Environmental and Ecological Statistics* **5**, 117–154.

Wikle, C. and Royle, J.: 1999, Space-time dynamic design of environmental monitoring networks, *Journal of Agricultural, Biological, and Environmental Statistics* **4**(4), 489–507.

Zhigljavsky, A.: 1991, *Theory of Global Random Search*, Kluwer Academic Publishers, Dordrecht.

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

Samenvatting

Het milieu staat sinds enige decennia nadrukkelijk in de belangstelling. Dat is niet verwonderlijk omdat onze leefomgeving één van de belangrijkste determinanten is voor de kwaliteit van het leven. Om processen in het milieu alsmede de invloed van de mensheid daarop te bestuderen, moeten gegevens verzameld worden. Statistiek is niet alleen een belangrijk hulpmiddel om deze gegevens te analyseren, maar is ook belangrijk om te bepalen op welke manier deze gegevens verzameld moeten worden. Gegevens worden meestal gedurende een bepaalde periode verzameld door middel van een netwerk van meetstations, kortweg aangeduid met de term *meetnet*. Deze gegevens kunnen worden gebruikt om processen in het milieu met een wiskundig model te beschrijven.

Het voornaamste doel van dit proefschrift is om onderzoek te doen naar het optimaliseren van dergelijke meetnetten, met behulp van wiskundige modellen. Aan de hand van problemen uit de praktijk wordt getoond hoe meetnetten kunnen worden uitgedund of uitgebreid. Deze beslissingen zijn gebaseerd op het minimaliseren van kwantitatieve criteria. In dit proefschrift worden drie verschillende onderwerpen uit de statistiek besproken en toegepast: ruimtelijke(-temporele) interpolatie (hoofdstukken 2,5 en 6), optimale proefopzetten (hoofdstukken 3 en 4) en het optimaliseren van meetnetten (hoofdstukken 4,5 en 6).

In het tweede hoofdstuk van dit proefschrift wordt een zevental interpolatietechnieken besproken: drie vormen van flexibele regressie (*splines*) en vier geostatistische benaderingen. Deze technieken worden toegepast voor het interpoleren van de maandelijkse maximale temperatuur en de maandelijkse gemiddelde neerslag in de provincie Jalisco in Mexico. De resultaten laten zien dat het de moeite waard is om gebruik te maken van de correlatie tussen meteorologische variabelen en de hoogte van het gebied. Interpolatie-resultaten kunnen aanzienlijk verbeteren als de correlatie tussen de hoogte en de meteorologische variabele groot is, zoals die tussen temperatuur en hoogte. Deze extra informatie kan op verschillende manieren verwerkt worden in de bestaande interpolatietechnieken. Zo kan (gewone) kriging worden uitgebreid tot cokriging of regressiekriging. Uitbreiding van de flexibele regressie met een derde dimensie, namelijk de hoogte van het gebied, blijkt een verbetering van de interpolatie-resultaten te geven. Analoog hieraan is een vorm van drie-dimensionale kriging getest. Voor maandelijkse maximale temperatuur wordt geconcludeerd dat deze drie-dimensionale vormen van de interpolatie-technieken de voorkeur verdienen. In de geostatistische literatuur wordt veel aandacht besteed aan het gebruik van extra informatie voor het verbeteren van de predictie-resultaten. Er is echter relatief weinig aandacht voor de behandelde drie-dimensionale vorm van kriging. Hoofdstuk 2 laat zien dat deze vorm meer aandacht verdient. Conclusies voor maandelijkse gemiddelde neerslag zijn moeilijker te trekken, omdat er een grotere variabiliteit in de data is.

Hoofdstuk 3 geeft een overzicht van verschillende methoden voor het bepalen van optimale proefopzetten voor regressiemodellen. Afhankelijk van het probleem resulteert de vraag naar een optimale proefopzet in het oplossen van een (soms uitdagend) globaal optimaliseringsprobleem. Er worden vier verschillende oplossingsmethoden besproken. Ten eerste de analytische werkwijze, waarbij de optimale proefopzet op exacte wijze afgeleid kan worden. Vervolgens, een specifiek algoritme dat (onder bepaalde condities) convergeert naar de optimale proefopzet. In de derde plaats wordt de situatie bestudeerd waarin de keuze van punten in de proefopzet beperkt is tot een verzameling van mogelijke punten. Dit probleem wordt opgelost met een algoritme uit de theorie van de combinatoriek, een zogenaamd *branch-and-bound algoritme* (zie de Appendix van dit proefschrift). Tenslotte wordt aangetoond dat algemene optimalisatie-algoritmen vaak niet voldoen vanwege het feit dat meer dan één optima aanwezig is. Echter, een combinatie van algemene optimalisatie-methoden kan tot goede resultaten leiden.

Het onderwerp in hoofdstuk 4 is het optimaliseren van een meetnet voor het schatten van de semivariantiefunctie. Het criterium is gebaseerd op het minimaliseren van de schattings-variantie van de parameters (D-optimaliteit) van de semivariantiefunctie, analoog aan optimale proefopzetten voor niet-lineaire regressie (zie hoofdstuk 3). Aan de hand van een meetnet voor onderzoek naar zure regen in de Verenigde Staten worden drie aspecten aan de bestaande literatuur toegevoegd.
(*a*) Met behulp van een aantal figuren wordt het probleem gevisualiseerd. Dit richt zich in eerste instantie tot het berekenen van gebieden waar het (in de zin van minimale-uitkomst-criterium) gunstig is om een extra meetstation te plaatsen. Daarnaast wordt getoond wat de invloed is van het modelleren van de correlatie tussen de waarnemingen voor het schatten van de semivariantiefunctie.
(*b*) Een combinatorisch optimalisatie-algoritme (*branch-and-bound*) is toegepast voor het vinden van het optimale meetnet bij een beperkt aantal mogelijke meetstations. Dit algoritme werkt met volledige aftelling, dat wil zeggen dat de optimale combinatie van meetstations met zekerheid gevonden wordt.
(*c*) De gevonden oplossing is alleen optimaal gegeven de *a priori* parameterwaarden van de semivariantiefunctie. Dergelijke *a priori* parameterwaarden komen vaak uit een eerdere studie. Met vervolgonderzoek wordt geprobeerd deze parameterwaarden beter te schatten. Het is dus niet denkbeeldig dat de schatting van de parameters afwijken van de *a priori* parameterwaarden. Door middel van

een robuustheid-studie wordt de vraag beantwoord of de optimale combinatie van meetstations veel verschilt bij afwijkende parameterwaarden. In deze studie blijken de verschillen, ook bij aanzienlijke afwijkingen, zeer beperkt te zijn.

Terwijl verschillende interpolatie-technieken reeds in hoofdstuk 2 aan de orde komen, wordt in hoofdstuk 5 een andere mogelijke interpolatie-techniek toegepast: lokale gewogen regressie. Het idee achter deze laatste techniek is om het trendoppervlak door een lokale benadering te bepalen. Op deze manier is het niet nodig globale modellen te gebruiken voor de trend en de ruimtelijke variabiliteit. Deze interpolatie-techniek wordt toegepast op het Nederlandse meetnet voor SO_2 van het Landelijke Meetnet Luchtkwaliteit van het RIVM (Rijksinstituut voor Volksgezondheid en Milieu). Aan de hand van criteria gebaseerd op lokale gewogen regressie wordt een suggestie gedaan hoe het meetnet uitgedund kan worden. In deze criteria kunnen politieke wensen worden meegewogen, zoals bijvoorbeeld de wens tot meer meetstations bij een hogere bevolkingsdichtheid. Dit hoofdstuk laat zien dat een statistisch criterium eenvoudig kan worden gecombineerd met wensen van beleidsmakers. Een zoekalgoritme wordt gepresenteerd waarmee een (sub)-optimale oplossing kan worden gegenereerd. Oplossingen van drie verschillende criteria worden met elkaar vergeleken.

In de hoofdstukken 4 en 5 zijn meetnetten geoptimaliseerd op basis van een criterium zonder dat daarbij een temporele component in beschouwing wordt genomen. In hoofdstuk 6 wordt een meetnet bekeken waarin zowel de ruimtelijke als de temporele component een rol spelen. Het betreft een meetnet voor grondwaterstijghoogten op de Veluwe, zoals dat door een zeker nutsbedrijf wordt gebruikt. Men verricht daartoe op vaste lokaties (peilbuizen) tweewekelijkse metingen. Het nutsbedrijf stelde de vraag of het aantal metingen gereduceerd zou kunnen worden zonder dat veel informatie over grondwaterstijghoogten verloren zou gaan. Dit wordt onderzocht aan de hand van een stochastische ruimte-tijd model met drie componenten: een gemiddelde stijghoogte per peilbuis, een seizoenseffect per peilbuis en een stochastisch ruimte-tijd gecorreleerde fout. De verschillende tijdreeksen kunnen per lokatie aanzienlijk verschillen. Om de heterogeniteit van de waarnemingen te beperken worden de waarnemingen opgedeeld in clusters. Empirische ruimte-tijd semivariantie geeft inzicht in de ruimtelijke en temporele variabiliteit van de metingen binnen elk van de clusters. Voor de reductie van het meetnet wordt in eerste instantie gekeken naar de frequentie van meten binnen een peilbuis. Een *Gaussische* simulatie is uitgevoerd om verschillende frequenties te kunnen vergelijken ten aanzien van de maximale predictie-fout en de gemiddelde predictie-fout. Daarnaast worden verschillende meetstrategieën met elkaar vergeleken door berekening van de predictie-variantie op basis van een ruimte-tijd semivariantiefunctie.

Curriculum vitae

Eric Boer werd op 16 april 1974 geboren te Oostvoorne. In 1992 behaalde hij zijn diploma Voortgezet Wetenschappelijk Onderwijs aan de christelijke scholengemeenschap "Blaise Pascal" te Spijkenisse. Hij ging daarna Agrosysteemkunde studeren aan de Landbouwuniversiteit Wageningen. Deze opleiding bood hem de mogelijkheid zich te verdiepen in biometrie. Zijn aandacht ging daarbij vooral uit naar optimale proefopzetten in niet-lineaire regressie en statistische kwaliteits- en proces-controle. Hij studeerde af (met lof) in maart 1997.

Vanaf juni 1997 was hij assistent in opleiding bij de leerstoelgroep Wiskundige en Statistische Methoden, die nu onderdeel is van Biometris van Wageningen Universiteit en Research centrum. Het in dat kader uitgevoerde onderzoek heeft geleid tot het voor u liggende proefschrift.

Per 15 september 2001 is de auteur van dit proefschrift werkzaam als statisticus bij het Instituut voor Agrotechnologisch Onderzoek (ATO B.V.) te Wageningen.