

Constrained Optimisation of Spatial Sampling

a Geostatistical Approach

Jan-Willem van Groenigen



GENTRALE LANDBOUWCATALOGUS

0000 0872 8871

Promotor: Dr. ir. J. Bouma
Hoogleraar in de Bodeminventarisatie en Landevaluatie
Landbouwwuniversiteit Wageningen

Co-promotor: Dr. ir. A. Stein
Hoogleraar in de Ruimtelijke Statistiek
ITC (International Institute for Aerospace Survey and Earth
Sciences)
..
Enschede

BIBLIOTHEEK
LANDBOUWUNIVERSITEIT
WAGENINGEN

NN08201, 2589

Jan-Willem van Groenigen

Constrained Optimisation of Spatial Sampling

a Geostatistical Approach

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van de Landbouwniversiteit Wageningen,
dr. C.M. Karssen,
in het openbaar te verdedigen
op 17 maart 1999
des namiddags te vier uur in de Aula

064643

ITC Publication Series

No. 65

The research presented in this thesis was partly conducted at the International
Institute for Aerospace Survey and Earth Sciences (ITC)
P.O. Box 6, 7500 AA Enschede, the Netherlands



CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

© Jan-Willem van Groenigen

ISBN 90 6164 156 X

Constrained Optimisation of Spatial Sampling - a Geostatistical Approach
Thesis Wageningen Agricultural University and ITC,
with ref. Summary in Dutch

Stellingen

behorende bij het proefschrift van Jan-Willem van Groenigen, *Constrained Optimisation of Spatial Sampling - a geostatistical approach*, Landbouwwuniversiteit Wageningen, 17 maart 1999.

1.

In termen van kriging variantie is een regelmatig grid vrijwel nooit het optimale bemonsteringsschema.

Dit proefschrift

2.

Het optimaal gebruiken van beschikbare voorinformatie zal in ruimtelijke bodem- en milieu-inventarisaties steeds belangrijker worden. Een flexibel algoritme voor optimalisatie van bemonstering zoals Spatial Simulated Annealing is hierbij onontbeerlijk.

Dit proefschrift

3.

Bij een gegeven aantal observaties kan het nemen van monsters buiten de grenzen van het onderzoeksgebied leiden tot een belangrijke verbetering van de interpolatie, in termen van zowel kriging variantie als gekwadrateerde kriging fout.

Dit proefschrift

4.

Door de sterke focus op het ontwikkelen van geavanceerde interpolatie- en simulatie-technieken is theorievorming over bemonstering in de geostatistiek jarenlang ten onrechte verwaarloosd.

Dit proefschrift

5.

Aangezien Indicator Kriging geen verdelingsfunctie, maar een *voorspelde* verdelingsfunctie oplevert, kan de kriging variantie in dit algoritme niet worden genegeerd.

Dit proefschrift

6.

Case studies in bodemkundige proefschriften zijn uitsluitend relevant als ze dienen voor het testen of illustreren van het ontwikkelde gedachtegoed. Zij mogen nooit een doel op zich worden.

7.

Iedere wetenschap krijgt de wetenschappers die zij verdient.

vgl. Paul Feyerabend, 'Against method'

8.

Zelfkennis is het begin van alle wijsheid; zelfoverschatting het begin van alle (verlangen naar) macht.

9.

Het is vreemd dat veel ecologische koffie- en thee-merken het Max Havelaar keurmerk niet voeren, aangezien dit suggereert dat de mens de enige productiefactor is die niet duurzaam aangewend hoeft te worden.

10.

Cursussen time-management zijn zinloos, aangezien mensen die er tijd voor vrij kunnen maken, het niet nodig hebben, en mensen die het nodig hebben, er geen tijd voor vrij kunnen maken.

11.

Na zonneschijn komt regen.

12.

Niets is zo ongeloofwaardig als een statisticus met vliegangst.

Voor Opa van Groningen

*"For nitrates are not the land, nor phosphates;
and the length of fiber in the cotton is not the land.
Carbon is not a man, nor salt nor water nor calcium.
He is all these, but he is much more, much more,
and the land is so much more than its analysis"*

John Steinbeck
'The Grapes of Wrath'

Acknowledgements

During the years I worked on this thesis, as in the rest of my life, the care and friendship of many people was crucial to me. Trying to thank them in words would result in acknowledgements both tedious and inadequate. Therefore, I decided to thank people only as far as they gave a direct intellectual or technical contribution to the contents of this thesis.

First, and most importantly, I want to thank my supervisor and co-promotor, Alfred Stein. Alfred, I am very much aware that for an AIO it is far from normal to have a supervisor who gives full support in both scientific and organisational terms. I am very grateful I had one.

Furthermore, I would like to thank my promotor Johan Bouma for his indispensable guidance, and for the pleasant co-operation on several of this thesis' chapters.

Most of the studies presented in this thesis were or will be published together with co-authors. I would like to thank those that I didn't mention above: Ronald Zuurbier, Wouter Siderius, Mahamoud Gandah, Felix Mainam and Gerton Pieters. Without your work and expert knowledge, this thesis would not have existed.

I am grateful to Jaap de Gruijter, Gerard Heuvelink, Edzer Pebesma and Harry-Jan Hendricks-Franssen for their contributions to geostatistical discussions related to this thesis. In terms of geostatistics, the Netherlands is an exceptionally fertile and inspiring country, and I took full advantage of that. Furthermore, I would like to thank numerous anonymous reviewers for their often very useful suggestions and remarks on my papers.

I would like to thank my colleagues from the Soil Sciences division of ITC for many discussions on the application of geostatistics in soil survey, and for tolerating my insatiable appetite for unguarded computers. I especially want to thank David Rossiter for checking my English in the first chapters. Furthermore, I would like to thank Benno Masselink and Job Duim for their technical assistance on some of the graphics (the nice ones), and Daniëla Semeraro, Ceciel Wolters, Mireille Meester and Annette van 't Hooge for their secretarial support.

I am grateful to Wageningen Agricultural University and ITC for giving me the opportunity to do my Ph.D. research.

I would like to thank Ingrid Maters for designing the cover, and René Voorburg for giving a crash-course webpage design. Finally, I want to thank Erna for her very thorough editing of the text (and for many other, far more important things).

Contents

Acknowledgements	vii
Contents	ix
List of Symbols	xi

INTRODUCTION

1 General Introduction.....	1
-----------------------------	---

OUTLINE OF MAJOR TOOLS

2 Using Probability Maps for Phased Sampling.....	7
3 Spatial Simulated Annealing.....	23

OPTIMISING SAMPLING FOR SPATIAL INTERPOLATION

4 Minimising Kriging Variance: Methodology.....	41
5 Minimising Kriging Variance: Influence of Variogram Parameters.....	61
6 Optimising Multivariate Interpolation.....	75

OPTIMISING SAMPLING FOR MODEL ESTIMATION

7 Establishing Soil/Yield Relations in Precision Agriculture Studies.....	93
8 Sampling strategies for Effective Variogram Estimation.....	105

CONCLUSIONS AND REFERENCES

9	Conclusions and Further Research.....	125
	References.....	133
	Bibliography.....	139
	Summary	141
	Samenvatting	145
	Curriculum Vitae	149

List of Symbols

The following list contains symbols that are used in more than one chapter of this thesis.

A ,	:	Research area.
A_R	:	
A_S	:	Part of research area that can be sampled, $A_S \subset A_R$.
h	:	Length of separation vector, $\ \bar{x}_i - \bar{x}_j\ $.
\bar{h}	:	Separation vector, $\bar{x}_i - \bar{x}_j$.
L_i	:	i^{th} Soil layer.
n	:	Number of sampling points, $\bar{x}_1, \dots, \bar{x}_n$.
n_e	:	Number of evaluation points on a fine raster grid, $\bar{x}_1^e, \dots, \bar{x}_{n_e}^e$.
nc	:	Number of lag classes for experimental variogram.
$P_c(S_i \rightarrow S_{i+l})$:	Transition probability of S_i to S_{i+l} during optimisation with Spatial Simulated Annealing.
ReV	:	Regionalized variable, usually denoted by $Z(\cdot)$.
S	:	Sampling scheme, consisting of sampling points $\bar{x}_1, \dots, \bar{x}_n$.
S_i	:	Intermediate sampling scheme in the i^{th} step of optimisation using Spatial Simulated Annealing.
\bar{x}_0	:	Location vector of predicted point using (indicator) kriging.
\bar{x}_i	:	Location vector of the i^{th} observation point.
$\bar{x}_{e,j}$:	Location vector of the j^{th} evaluation point.
$V_S(\bar{x})$:	Location vector of the nearest observation point \bar{x}_i to \bar{x} .
$z(\bar{x})$:	Observation of the ReV $Z(\cdot)$ at location \bar{x} .
$Z(\cdot)$:	Regionalized Variable (ReV).
α	:	Level of significance.
$\phi(S)$:	Fitness function to be minimised, expresses the performance of sampling scheme S .

- $\gamma(\mathbf{h})$: (Semi-) variance as function of separation vector \mathbf{h} .
- $\hat{\gamma}(\mathbf{h})$: Estimated (semi-) variance as function of \mathbf{h} .
- λ_i : Weight of the i^{th} observation point in the kriging predictor of \bar{x}_0 .
- σ_i : Standard deviation of separation vectors \mathbf{h} in lag class i .
- σ_{OK}^2 : Ordinary kriging variance
- ζ_i, ζ_i^S : Realised number of point pairs in lag class i .
- ζ_i^* : Ideal number of point pairs in lag class i , as defined by the user.

Chapter 1

General Introduction

"Forty-two!" yelled Loonquawl. 'Is that all you've got to show for seven and a half million years' work?' 'I checked it very thoroughly', said the computer, 'and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is.'"

Douglas Adams
'the Hitch Hikers' Guide to the Galaxy'

1.1. Spatial sampling or soil sampling?

This thesis is written by a soil scientist who specialised in geostatistics. Therefore, all case studies and most examples are drawn from soil science. Yet, the title of this thesis is 'Constrained Optimisation of *Spatial* Sampling' instead of '*...Soil* Sampling'. There are two reasons for this:

- i)* The thesis is only concerned with soil sampling for characterisation of spatial distribution. Issues like optimising sampling of excavated soil for waste disposal, the optimal number of samples for bulking, calculation of the mean phosphate saturation, *etc.* fall outside its scope. Therefore, the spatial character of the study was made explicit in the title.
- ii)* Although the case studies in this thesis are exclusively drawn from soil science, the developed techniques should be easy to modify for application in other scientific fields. The central concept in geostatistics is the theory of Regionalized Variables (ReV's), and any scientific field dealing with such ReV's could potentially benefit

from these techniques. These fields include, among many others, remote sensing (e.g. Csillag *et al.*, 1996), hydrology (e.g. Hendricks-Franssen and Gómez-Hernández, 1997), meteorology (e.g. Seo *et al.*, 1990) and even marine biology (e.g. Petitgas, 1993).

1.2. Motives for the study

The importance of sampling strategies in soil-related surveys stems from the fundamental fact that our knowledge on soils in their natural state is at best incomplete, at worst erratic. The data we have on soils is usually either indirect (such as aerial photography) or based upon destructive techniques (such as laboratory analysis). Maps on soil properties in their natural state therefore have to be inferred using conceptual models such as soil-landscape relationships, quantitative models such as provided by geostatistics, or (preferably) a combination of both. They are never fully known. This "*frustrating feature of reality*" (Isaaks and Srivastava, 1989: p.107) forces the geostatistician to collect samples as a basis for spatial characterisation.

Yet, collecting samples has been relatively neglected in the geostatistical literature. While exceedingly complex algorithms for interpolation and (more recently) stochastic simulation have been developed, sampling strategies have not drawn much attention. Most textbooks on geostatistics start with data analysis, taking the data collection for granted, or at most dedicate a few lines to it. Deutsch and Journel (1998), while discussing 13 different types of kriging and 8 types of stochastic simulation, fail to give any recommendations on the collection of data to feed these algorithms. Although Goovaerts (1997a) makes some remarks on the (nested) sampling strategy used to collect his main data set, he states that "*In this book, one considers the situation where data have already been collected, possibly with no statistical treatment in mind*" (Goovaerts, 1997a: p.75). Isaaks and Srivastava (1989) focus on how to correct for inadequate sampling strategies, rather than how to avoid them. Webster and Oliver (1990) include a much more detailed discussion on sampling strategies. Their main focus, however, is on sampling strategies derived from classical sampling theory (sampling *designs* such as systematic sampling, random sampling, *etc.*). In classical sampling theory (as opposed to geostatistics), several well-established sampling designs are routinely applied (e.g. Cochran, 1977; Thompson, 1992). More recently, optimisation strategies for such sampling designs in a spatial context have been developed (De Gruijter and Ter Braak, 1990; Domburg *et al.*, 1994).

The fact that few sampling strategies (apart from the well-known regular grids) belong to the established tools of the geostatistician, does not mean that no significant research on the subject has been done. Landmark papers were published on the optimal grid spacing (McBratney and Webster, 1981), the type of regular grid (Yfantis *et al.*,

1987), optimal estimation of the variogram (Webster and Oliver, 1992) and on geostatistical vs. classical sampling theory (De Gruijter and Ter Braak, 1990). However, these results have not yet evolved into sampling strategies that can be routinely applied in practice, as have many interpolation and simulation algorithms.

One of the reasons for this is the wide variety of optimisation criteria met in different soil-related surveys. Cressie remarks on sampling strategies that "*at the very basis of optimal statistical design is ... the choice of what is to be estimated or predicted, second the choice of the estimator or predictor...*" (Cressie, 1991: p.314). A few examples from the literature can show how far optimisation criteria, either implicitly or explicitly stated, can diverge in soil-related spatial studies:

- i) Yfantis *et al.* (1987) used the mean kriging variance as a discrimination criterion between different types of regular grids.
- ii) Warrick and Myers (1987) used the point pair distribution of the sampling scheme for estimation of the variogram as an optimisation criterion for a Monte-Carlo optimisation.
- iii) Stein *et al.* (1988b) calculated the mean (squared) prediction error to assess the use of water table height as a covariable in cokriging of moisture deficit.
- iv) Webster and Oliver (1992) used the fluctuation of the experimental variogram values as a means of selecting between different sampling schemes.
- v) Brus (1994) aimed at minimal sampling variance of the mean Phosphate saturation over the whole study area.
- vi) Watson and Barnes (1995) defined several optimisation criteria, among them optimisation of the chance of detecting the maximum value in the area of interest.

Before any optimisation of the sampling scheme can be tried, the optimisation criterion should be explicitly stated. In fact, the formulation of an optimisation criterion may contribute to the understanding of the problem at hand. For example, in studies aiming at optimal variogram characterisation, formulation of the optimisation criterion may well be the most difficult part of the whole survey, as will be shown in this thesis.

A second reason for the lack of practical sampling strategies in geostatistics is that in soil survey many other, non-pedological criteria and issues may play an important role. Cressie observes on the problem of optimal sampling that "*The statistical problem is part of a much bigger picture...*" (Cressie, 1991: p.268). A good example in this context is the practice of soil remediation in urban areas, where the soil surveyor has to deal with such diverse scientific fields as ecology, toxicology, chemical technology and psychology (Okx *et al.*, 1996). As a further complication, sampling constraints such as buildings, roads *etc.* can easily include 90% of the area. Apart from that, the surveyor should take into account financial constraints and is always tied to environmental legislation, which may not necessarily require the most reasonable course of action.

Therefore, an optimisation method for spatial sampling should also be both flexible and robust, and should be able to handle all types of errors, deviations and non-scientific considerations that are usually met in practice, while still leaving room for the decision-making processes that are related to many types of spatial soil studies. This thesis presents such an optimisation method.

1.3. Purposes of this study

The main purpose of this thesis is the development of an all-purpose, flexible and robust optimisation algorithm for sampling in geostatistical studies. This overall purpose leads to several aims:

- i) Formulation of a range of optimisation criteria that honour a wide variety of aims in soil-related surveys.
- ii) Development of an optimisation algorithm for spatial sampling that is able to handle these different optimisation criteria.
- iii) Incorporation of ancillary data such as co-related imagery, historic knowledge and expert knowledge in the sampling strategy.
- iv) Comparison of the performances of the developed optimisation algorithms with established sampling strategies.
- v) Application of developed optimisation techniques in practical soil sampling studies.

1.4. Definitions and scope

This thesis deals exclusively with optimisation in a geostatistical context. The variables that are considered are Regionalized Variables (ReV's), and the case studies presented therefore focus on issues like characterisation of the auto-correlation structure and optimal interpolation using different types of kriging. I use the term *sampling scheme* for sampling strategies in a geostatistical context (*i.e.* a list of optimal sampling locations for characterisation of the ReV). The term *sampling design* is used to refer to sampling strategies based upon classical sampling theory, and indicates a method of drawing sampling locations (*e.g.* simple random *vs.* stratified sampling) rather than actual sampling locations.

As the purpose of this study was the development of an all-purpose optimisation algorithm, examples and case studies were drawn from as wide a variety of applications as possible. The studies in this thesis range from plot scale to geomorphological unit scale, from precision agriculture to soil contamination, from tropical to temperate

climates, and from univariate to multivariate analyses. Such a wide variety of case studies prohibits extensive discussion and interpretation of all results. Therefore, the case studies should be seen as an illustration of the developed methodology rather than as a purpose in itself. Most of the case studies in this thesis are part of a larger research- or mapping effort, conducted by experienced soil scientists, and will therefore be included in more extensive reports elsewhere.

1.5. Outline of the thesis

This thesis is essentially a collection of papers and should be regarded as such. Chapters 2 to 8 have been or will be published in international peer-reviewed journals. Apart from the standardised layout and some minor editing for reasons of consistency (mainly notation), a combined references list and a list of symbols has been composed. Although this thesis represents a coherent line of research, some inevitable gaps and overlaps are therefore to be expected, especially in the introductions and conclusions of the papers/chapters. However, in my opinion these drawbacks were considerably outweighed by the advantage of having critical feedback from other scientists during the peer-reviewing process.

Chapters 2 and 3 start with the outline of the main tools that I developed. These are applied, adapted and extended in chapters 4 to 8.

In chapter 2 it is shown how probability maps produced using Indicator Kriging can be used in a multi-stage sampling approach to focus sampling on areas with higher risk of contamination. The method is applied in an environmental case study, and is tested using stochastic simulation. Compared with conventional sampling schemes, this method results in more efficient remediation maps with similar health risks.

Chapter 3 introduces Spatial Simulated Annealing (SSA) as a general optimisation algorithm for spatial sampling schemes. It is shown how the Simulated Annealing algorithm is adapted for spatial purposes, and how ancillary information can be incorporated in the sampling strategy. The functionality of the algorithm is demonstrated using two optimisation criteria from the literature.

Chapters 4, 5 and 6 deal with optimisation of sampling schemes for spatial interpolation. In Chapter 4, minimisation of the mean kriging variance is added to the optimisation criteria of SSA. It is shown how both kriging neighbourhood and anisotropy influence optimal sampling schemes. The optimisation criterion is illustrated in a case study on a river terrace in Thailand.

Chapter 5 further explores the possibilities of minimising the kriging variance using SSA. Minimisation of the maximum kriging variance is added as an optimisation criterion, and the influence of variogram parameters on optimised sampling schemes is investigated. It is shown that all variogram parameters influence the optimised

sampling scheme, and that there is a considerable difference in this respect between minimising the mean kriging variance and minimising the maximum kriging variance.

In Chapter 6, a new optimisation criterion is introduced that can assist in optimising sampling schemes for multivariate contamination studies. Using a spatial weight function, priorities in sampling can be set using historic information, expert judgement or preliminary observations. This technique is applied in a highly complex contamination study in the Rotterdam harbour.

Chapter 7 deals with optimising sampling using the type of co-related imagery that is often met in precision agriculture (PA) studies. Using a simple scoring technique, yield maps are predicted. These maps assist in optimising sampling for finding soil-yield relations.

Chapter 8 deals with the question of optimising sampling for variogram estimation. A hybrid sampling scheme is introduced, combining the advantages of optimal coverage of the area and short range observations. Finally, it is shown that accuracy of the experimental variogram is usually of little value without considering the effects on kriging accuracy.

Finally, chapter 9 summarises the main conclusions of the thesis, and gives some recommendations for further research.

1.6. Software

Spatial Simulated Annealing (SSA) is presented in this thesis as an optimisation algorithm for spatial sampling. It is described in chapter 3, and further extended in chapters 4 to 7. During the research I programmed the SSA optimisation algorithm as the SANOS (Simulated ANnealing for Optimising Sampling) program. This program (written in C++) can be downloaded in a preliminary version from <http://www.itc.nl/~groenig>. It includes all optimisation criteria that are presented in this thesis. The site also includes some example files and a brief user manual. In the future, updates on both software and user manual will be made available at this site.

Chapter 2

Using Probability Maps for Phased Sampling¹

Abstract

A phased sampling procedure is proposed to optimise environmental risk assessment. Subsequent sampling stages were used as quantitative pre-information. With this pre-information probability maps were made using indicator kriging to direct subsequent sampling. In this way, better use of the remaining sampling stages was ensured. Phased sampling was applied to a lead-pollution study in the Dutch city of Schoonhoven. Environmental risks were quantified by the probability of exceeding the intervention level. Using six conditional simulations of stochastic fields, phased sampling schemes were compared to conventional sampling schemes in terms of type-I and type-II errors. The phased schemes had much lower type-I errors than the conventional schemes, and comparable type-II errors. Moreover, the phased sampling schemes left a smaller fraction of the not-remediated area polluted than the conventional ones did. They predicted almost 70% of the area correctly, as compared to 55% by conventional schemes.

¹ *Published as:* Van Groenigen, J.W., Stein, A. and Zuurbier, R. (1997). *Optimisation of environmental sampling using interactive GIS.* Soil Technology 10: 83-97.

2.1. Introduction

In many environmental studies, Geographical Information Systems (GIS's) are routinely applied. Questions about the reliability of GIS-generated data, error propagation (Heuvelink *et al.*, 1989), the division of tasks between user and computer (Okx *et al.*, 1990), and their interaction (Stein *et al.*, 1995) have been studied in the past. Systems are being proposed or developed which perform increasingly complex tasks. Intelligent GIS (Burrough, 1992), knowledge based systems (Domburg, 1994), expert systems (Burrough, 1986), decision support systems (Armstrong and Densham, 1990), and fuzzy Soil Information Systems (Kollias and Voliotis, 1991) are distinguished.

Still, the reliability of the output of a GIS, such as maps, basically depends upon the quality of the data. When considering soil data that has to be collected in the field, data quality is determined in part by the sampling scheme: a poorly designed sampling scheme yields unreliable results (Corsten and Stein, 1994). In the past, attention was focused on sampling schemes which minimise the uncertainty of maps (McBratney *et al.*, 1981). Also, qualitative and quantitative pre-information has been used to determine an optimal sampling scheme for soil pollution (Van Tooren, 1993) and variograms from previous comparable surveys have been used to optimise sampling schemes (Domburg *et al.*, 1994; McBratney *et al.*, 1981).

Until now, phased sampling was barely investigated for geostatistical studies, although adaptive sampling has been applied to estimate parameters of distributions in non-geostatistical studies (*e.g.* Thompson, 1992). In this study, a data set is set up using different sampling stages. This is done by using already analysed samples as pre-information for subsequent sampling. Phased sampling is applied to an area polluted with lead in the Dutch city of Schoonhoven. The aim was to provide a map showing the probability that a critical intervention threshold was exceeded. Moreover, the quality of the sampling scheme was compared with that of other schemes using conditional simulations (Cressie, 1991; Deutsch and Journel, 1992) to identify type-I and type-II errors.

2.2. Materials and methods

2.2.1. Study area

The study area is located in the Dutch city of Schoonhoven. In this city, a possibly severe lead-pollution was detected. Supposedly, most of this pollution was caused by a single factory that has been under operation, producing lead-containing chemicals, for over 200 years. The spatial distribution of the pollution was caused by a combination

Table 2.1. *Statistical parameters of the preliminary research and the three sampling stages for the two layers L₁ (0-0.2 m) and L₂ (0-0.5 m).*

Statistics	Pre-stage	Stage 1		Stage 2		Stage 3		
		L ₁	L ₂	L ₁	L ₂	L ₁	L ₂	
No. of samples	-	28	76	100	162	143	201	177
Mean	mg kg ⁻¹	274	497	451	323	333	349	309
Stand. Dev.	mg kg ⁻¹	312	692	780	414	483	481	320
Coeff. of var.	-	1.14	1.39	1.73	1.28	1.45	1.37	1.03
Max.	mg kg ⁻¹	1300	4050	5000	2200	4250	4600	1950
Min.	mg kg ⁻¹	17	21	11	10	11	10	11
Perc. above z _i	%	17.1	19.7	19.0	13.6	15.0	13.8	14.7
Median	mg kg ⁻¹	177.5	165	195	305	185	190	195

of atmospheric deposition from periodical cleaning of the factory, man-made deposition from contaminated sewer sludge, horse dung used during the production process and permanent background values. In the Netherlands, environmental standards are set to classify areas according to their degree of pollution. In particular, if concentrations on a single contaminant exceed the so-called intervention level, environmental measures have to be taken. For lead, the intervention level is equal to 600 mg kg⁻¹ dry matter.

To identify the extension of the pollution above the intervention level a survey was carried out. The aim was to delineate the extent of the pollution as precisely as possible, because remediation is expensive and health risks are at stake. Hence the samples have to be located as efficiently as possible. In addition to an estimate of the pollution at each location, it is important to know the accuracy of that estimation, since decision making on remediation of the soil should be based on probabilities of exceeding the intervention level. Typical probabilities to consider are 0.01 and 0.05. Maps showing these probabilities are obtained with indicator kriging (Deutsch and Journel, 1992).

Preliminary research in Schoonhoven was done in 1988. Several transects were sampled in the surroundings of the factory. Analysis of these data indicated a lead pollution with a peak near the factory and decreasing concentrations with increasing distance from the factory (Table 2.1). As the total number of samples was only 28, little information was available on the spatial distribution of the pollution. But since the highest measurement equalled 1300 mg kg⁻¹, and 17% of the measurements were above the intervention level, a thorough survey was carried out in 1992. During this survey, several depths were sampled at each sampling location, from which only two depth classes had sufficient data to apply a geostatistical analysis:

- i) layer 1 (L₁), ranging to 0.2 m below the soil surface, including samples taken at 0-0.2 m and at 0.1-0.2 m below the soil surface.

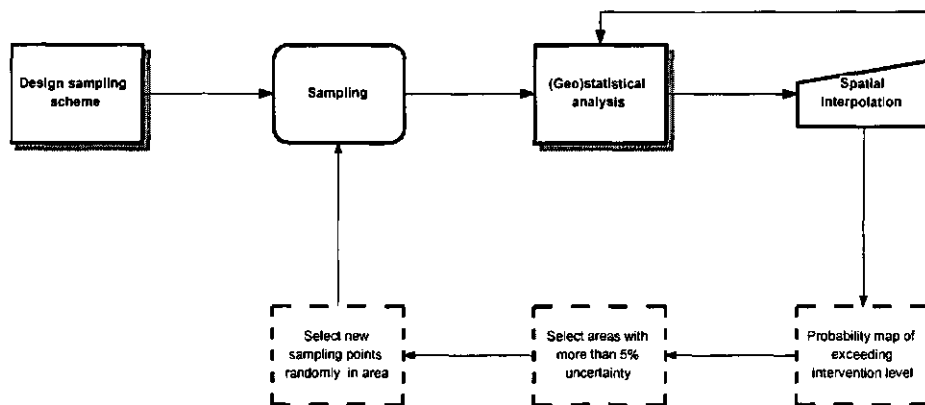


Figure 2.1. Schematic overview of a normal soil-survey using geostatistics (non-dotted figures) and the extensions when using phased sampling (dotted figures)

- ii) layer 2 (L_2), ranging to 0.5 m below the soil surface, including samples taken at 0-0.5 m, 0.1-0.5 m and 0.3-0.5 m.

Mixed samples were taken over different depth intervals. Data were stored in a Geographical Information System (GIS).

2.2.2. Phased sampling

The aim of environmental sampling can be to collect data such that the area with a low, fixed probability of being polluted with a contaminant above a threshold level is determined as precisely as possible. In this way, risk-qualified remediation can be executed. Commonly, sampling is conducted as follows (Figure 2.1): first, a sampling scheme is designed, using knowledge derived from earlier surveys on the soil parameter, geostatistics, historical information and organoleptic judgement. Next, data is collected following this scheme. In the field, deviations from the sampling scheme are likely to occur because of sampling constraints. For example, sampling below a house may be prohibitive. After sampling, spatial analyses are carried out, e.g. using geostatistics. At this stage, interactive data exploration as used in Haslett *et al.*, (1990) may be applied. With trial and error and a good data presentation, insight is gained in the spatial properties of the data. If necessary, measurements can be re-analysed, removed or added after additional sampling. Also, the best method of spatial interpolation is chosen interactively, using expert-judgement.

In this study, a phased sampling procedure is proposed. Data collected during

sampling is used to direct further sampling. Let the total number of observations to be collected (n) be fixed. Then the following sequence takes place:

- i) The sampling scheme chosen at first contains n_1 observations, with $n_1 \leq n$. For example, n_1 may be chosen equal to $n_1 = \frac{1}{2}n$. The sampling scheme can be of a conventional type, e.g. random or grid, possibly stratified.
- ii) Collected data is stored and used to produce maps and probability maps. The data is used *as if* a full data set was collected.
- iii) Next, n_2 points (for example $n_2 = \frac{1}{2}n_1$) are selected at locations that require additional sampling, and a second sampling stage is carried out.
- iv) The steps ii) and iii) are repeated until sufficient precision is obtained or the maximum number of data is reached.

In this procedure, it is assumed that the spatial distribution of the pollution does not change significantly during the sampling period. The procedure is shown in Figure 2.1 by the dotted figures and lines. If we consider A our 1-, 2- or 3-dimensional sampling area, the aim of the study is to determine the sub-area $F_1(\alpha) \subset A$ that has a probability higher than α of having concentrations of variable Z above the critical intervention level, z_1 . To predict the locations with high pollution the areas with a probability $\alpha > 0.05$ of exceeding z_1 are selected. These areas are excluded from further sampling. In this way geostatistical knowledge and field-knowledge which are useful to design sampling schemes increased during sampling, whereas by means of one stage procedures collected data are evaluated only afterwards.

2.2.3. Geostatistics

The aim of geostatistics is to analyse regionalized variables (ReV's). Suppose that the contaminant $Z(\bar{x})$ can be considered an ReV, where \bar{x} denotes the location in A and Z denotes the concentration of the contaminant. Usually, observations are collected, denoted with $z(\bar{x}_1), \dots, z(\bar{x}_n)$. The spatial dependence is commonly expressed in a variogram, defined as half the variance of pair differences of an ReV at two locations, \bar{x} and $\bar{x} + \bar{h}$, as a function of the distance \bar{h} between these two locations (Webster and Oliver, 1990). The variogram may be used for a range of spatial interpolation techniques (e.g. kriging). Kriging provides the best linear unbiased predictor of a spatial variable at unvisited locations. When applied to a large number of closely located grid nodes, it can be used to make maps which show the predicted spatial distribution of the variable. In this study, several variables were interpolated. Lead content itself, but also an indicator variable, equal to 0 if the intervention threshold (600 mg kg^{-1}) is exceeded and 1 when it is not.

An interpolated indicator variable can be used to predict the conditional probability that the intervention value is exceeded, given the observations. Therefore, it can be used to create probability maps, which show the probabilities of exceeding a critical value (Journal, 1983; Deutsch and Journal, 1992).

When setting up a spatial sampling scheme in a geostatistical study, the questions to be decided upon are the number of observations, n , and the sampling locations $\bar{x}_1, \dots, \bar{x}_n$. In principle, global criteria may be defined, like the distribution of the observations in such a way that a prescribed precision of an interpolated map is obtained. This results normally in the optimal spacing of some regular grid. In this chapter we propose a phased procedure. To compare the interactive schemes with conventional ones, use was made of conditional simulations. Special attention was given to extreme spatial variability, and skewed distributions, commonly occurring in environmental pollution. Therefore, sequential indicator simulation (SIS) was applied (Deutsch and Journal, 1992; Bierkens and Burrough, 1993a/1993b). The SIS algorithm uses as input a data set on $Z(\bar{x})$, which is transformed using an indicator-function for several values along the distribution curve, with for each value an indicator variogram. The following sequence of events takes place:

- i) Start with an existing data set and define a grid for mapping.
- ii) A previously unvisited node in the grid \bar{x}_i is drawn randomly.
- iii) A conditional distribution function (cdf) is estimated at \bar{x}_i with indicator kriging.
- iv) From the cdf a realisation $z^{(1)}(\bar{x}_i)$ is obtained by drawing a random number $\pi^{(1)}$ between 0 and 1, and finding the corresponding quantile of the conditional distribution function.
- v) $z^{(1)}(\bar{x}_i)$ is added to the data set, and the procedure starts at a second unvisited node, randomly drawn from the grid, until all nodes of the grid have been drawn.

As each node is added in a random order to the conditioning data, extensive simulation will reproduce the imposed variogram (for proof, see Journal, 1989, pp. 34-35). This results in a spatial variation which is much higher, and closer (but not similar) to the real variation, than spatial variation obtained with kriging. Below, SIS will be used to simulate environmental pollution.

2.3. Results

2.3.1. Actual sampling and probability maps

The survey was conducted in three stages. At the first stage an equilateral triangular grid was applied with edges of 50 m, covering an area of about 600 x 800 m, surrounding

the factory. After this first stage, two additional stages were performed, yielding a three-stage sampling procedure. Table 2.1 summarises the statistical parameters of the subsequent stages.

At stage one, relatively small data-sets were obtained for L_1 and L_2 , consisting of 76 and 100 data, respectively. Compared to the preliminary stage, the means of these data (497 mg kg^{-1} and 451 mg kg^{-1} , respectively) are much higher, probably due to the closer distance to the source of pollution. Also, the maximum values of 4050 and 5000 mg kg^{-1} are much higher than the preliminary measurements. The experimental variogram as well as the indicator variograms, however, fitted poorly (Figure 2.2). This was probably caused by a combination of three factors:

- i) the *quantity* of the data set: 76 and 100 measurements are still small numbers to deal with varying soil parameters.
- ii) the *quality* of the sampling scheme. As the used triangular grid had a spacing of 50 m, almost no information was available on spatial correlation at shorter distances.
- iii) the *variability* of the ReV. The experimental variogram equalled a pure nugget effect.

The data set of stage two largely solved the first problem by adding previously unanalysed samples, yielding now a total of 162 and 143 samples for L_1 and L_2 , respectively. Also, several measurements were re-analysed, and some errors were corrected. This resulted in a much better fit of the indicator variograms.

Because measurements at small distances were still lacking, a large nugget effect remained, which made it difficult to use the probability maps as described above. Sampling at the third stage focused on sampling at short distances. The sampling points were randomly drawn at short distances from existing sampling points, close to the centre of pollution. This yielded data sets, with 201 and 177 samples for L_1 and L_2 , respectively. The variograms showed a much better fit and had smaller nugget-variances (Figure 2.2).

Figure 2.3 shows the predicted pollution, and the probability of exceeding the intervention level for L_2 , both calculated using indicator kriging. A major pollution occurs in the surroundings of the factory, south west of the centre of the map. North east of the factory, a small area with increased pollution is delineated, which could be caused by atmospherical deposition, the predominant wind direction being south west. Also, severe pollution is predicted in the north and the south direction of the area, which, however, seem to be caused by boundary effects of the kriging procedure. The map with the probabilities of exceeding the intervention level is relatively flat. The remainder of the map shows only small changes in probabilities, apart from an area in the east that is almost certainly not polluted.

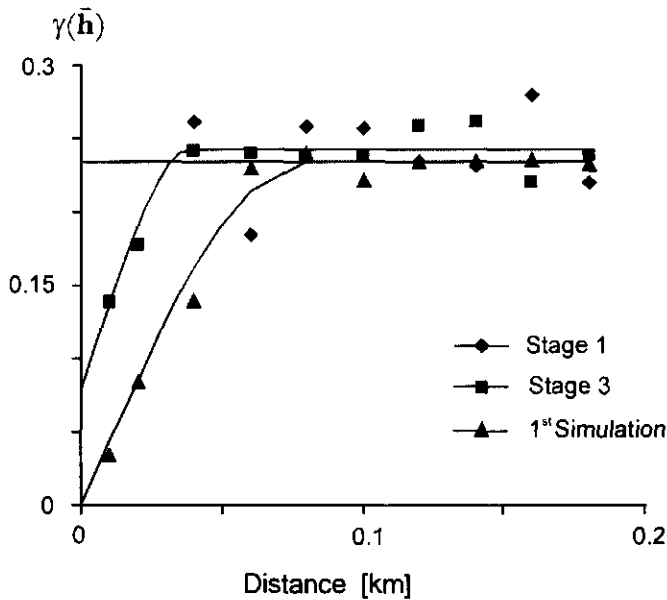


Figure 2.2: Experimental and fitted indicator variogram models for the median of the lead content (220 mg/kg) for the data collected at the first stage, the third stage and for the first simulated lead pollution.

2.3.2. Stochastic simulation

To compare phased sampling with different other sampling schemes in a quantitative way, use was made of stochastic simulations using SIS, conditioned on L_1 . The aim of this study was to determine $F_1(0.05)$, which is defined as the area that has a probability larger than 0.05 of exceeding the intervention level. All sampling schemes were tested by comparing the polluted area obtained by simulation on their efficiency in estimating the (simulated) pollution. Each scheme consisted of $n = 300$ observations, but differed in terms of the number sampling stages and the way in which the locations of the samples were selected (Table 2.2).

- i) Scheme one (S^1) consists of three stages, yielding $n_1 = 158$, $n_2 = 90$ and $n_3 = 52$ observations, respectively. At the first stage a square grid is applied with grid spacing of 60 m, to which a few random points are added to improve estimation of the nugget effect. At two succeeding stages random sampling within the selected area is applied.
- ii) Scheme two (S^2) consists of two stages, with $n_1 = 215$ and $n_2 = 85$ observations, respectively.

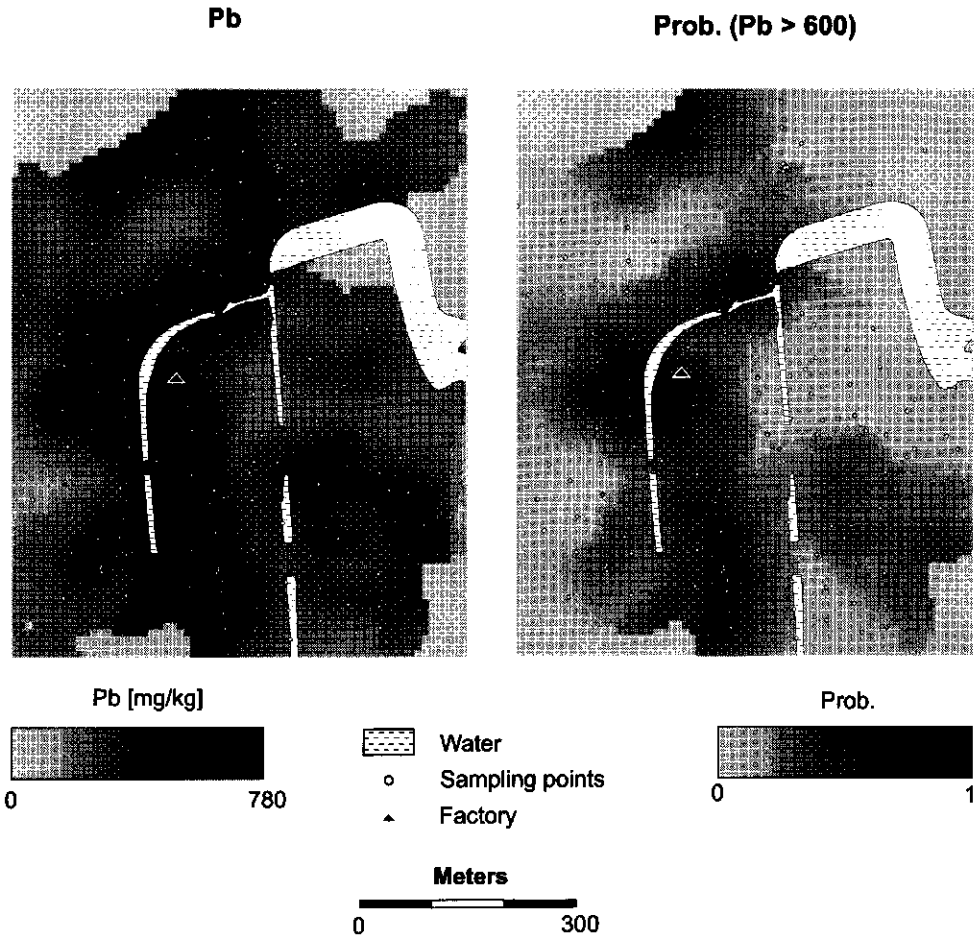


Figure 2.3. Predicted lead contents in Schoonhoven (left) and probability of exceeding intervention level z , for layer 1 (right).

- iii) Scheme three (S^3) is a one-stage sampling, using a 40 x 40 m square grid.
- iv) Scheme four (S^4) also is a one-stage sampling, with 300 points distributed randomly.

These four sampling schemes were applied to six different conditional simulations. Figures 2.4 to 2.7 show four of these simulations, each of these containing one of the four sampling schemes. Similar pictures were obtained for the two other simulations, with every combination of sampling scheme and simulation covered. The indicator variogram for the median is given in Figure 2.2. Maps with simulated pollution show several hot-spots and almost no unpolluted areas. The simulated data has the same

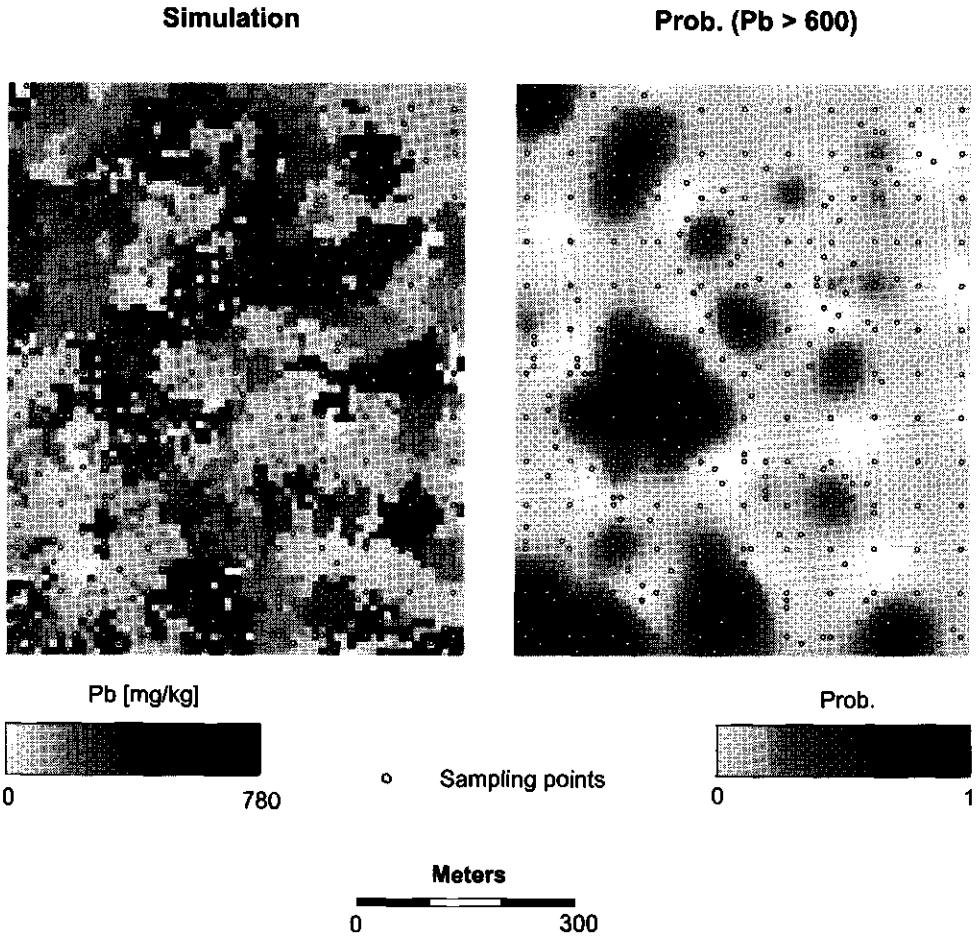


Figure 2.4. Simulated lead contents for simulation 1 (left) and predicted probability of exceeding intervention level z_1 with S^1 (right).

distribution as the Schoonhoven data-set (Table 2.1).

Three-staged phased sampling following S^1 is shown in Figure 2.4. Phased sampling results in more detailed information of the areas with a higher pollution. Therefore, the probability-map (Figure 2.4) predicts most of the areas with concentrations exceeding 600 mg kg^{-1} . However, it also recognises areas which are only moderately polluted. Two-staged sampling following S^2 shows a slightly less pronounced view (Figure 2.5). The predictions are less extreme, more areas with concentrations exceeding 600 mg kg^{-1} were detected and less moderately polluted areas were detected. The spacing of S^1 at the first stage is wider than that of S^2 , hence causing S^1 to overlook polluted

Table 2.2. Overview of the sampling schemes used to survey the simulated pollution. The number of sampling stages varies from 1 to 3, the type of sampling is either random or a square grid.

Scheme	Number of stages	Number of samples		Sampling scheme	
		total	per stage	type	spacing
S ¹	3	300	158	square	60 m
			90	random	-
			52	random	-
S ²	2	300	215	square	50 m
			85	random	-
			300	square	40 m
S ⁴	1	300	300	random	-

areas of a very small size, and tending to over-estimate the pollution of moderately polluted areas. The grid sampling following S³ picks up almost all of the polluted areas, but areas which are not polluted are often included. This results in small areas which can be declared unpolluted (Figure 2.6). The random sampling following S⁴, misses several important polluted areas because random sampling tends to leave large areas unsampled (Figure 2.7).

The aim of the survey was to determine those areas which were polluted with a probability of at least 0.05. To quantify the performances of the four schemes, the sizes of areas that were falsely or rightly classified as polluted or non-polluted were calculated and described as type-I and type-II errors (Table 2.3). Of the non-polluted areas, S₂ classifies the smallest area as having a probability of exceeding z_1 higher than 0.05 (28.3%). S¹ scores second-best (32.3%), while S³ (42.3%) and S⁴ (44.2%) perform much worse. This means that, although the errors are high, the type-I error is smallest for S². If attention is focused on polluted but not remediated areas (type-II error), the largest part of the polluted soil (12.3%) is remediated in the case of S³, which therefore has the smallest type-II error, closely followed by S² (12.2%), S¹ (12.0%) and S⁴ (11.4%).

In the practice of soil remediation, making a decision as to remediate the area is often based upon a 0.95 certainty level of having removed all contaminated soil. This

Table 2.3. Results of the four different sampling schemes used in estimating the simulated pollution, and the health risk, expressed as percentage of the not remediated area that is polluted.

Scheme	Not polluted		Polluted		Health risk (%)
	Not remediated (%)	Remediated (%)	Remediated (%)	Not Remediated (%)	
S ¹	52.9	32.3	12.0	2.8	5.0
S ²	56.9	28.3	12.2	2.6	4.4
S ³	42.9	42.3	12.3	2.5	5.5
S ⁴	41.0	44.2	11.4	3.4	7.7

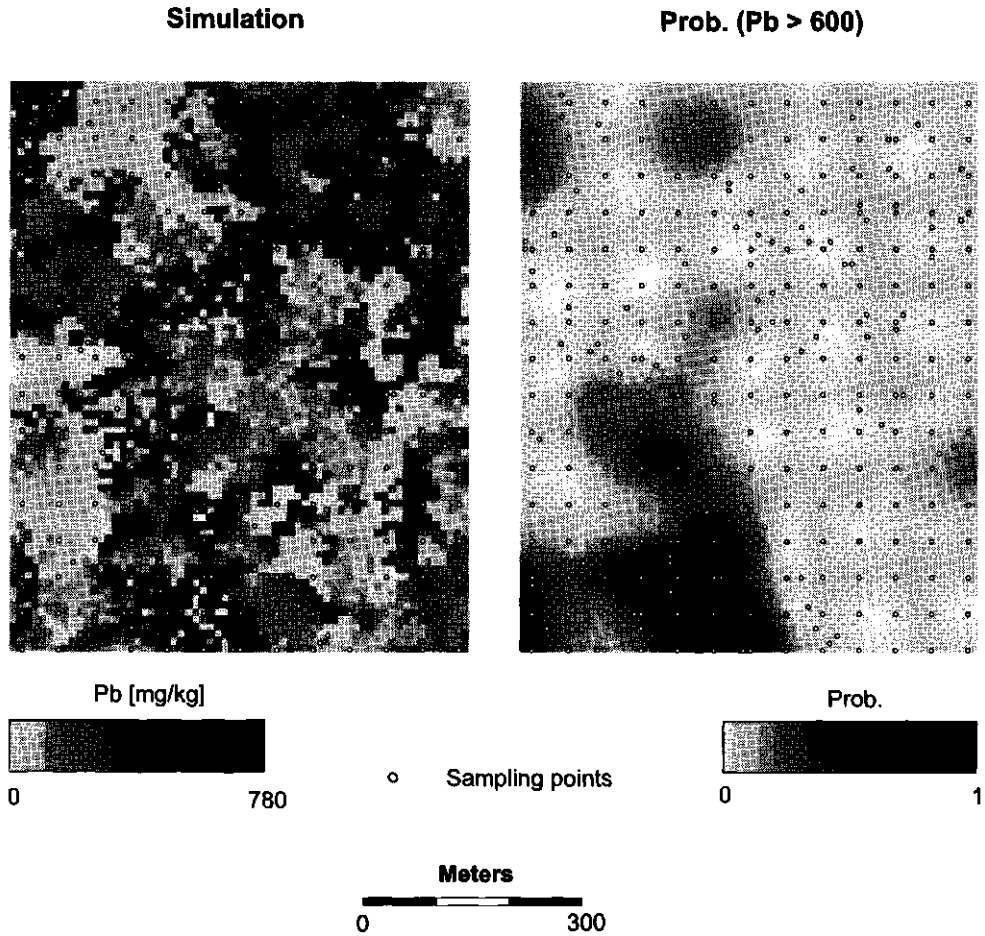


Figure 2.5. Simulated lead contents for simulation 2 (left) and predicted probability of exceeding intervention level z_1 with S^2 (right).

implies that an error of 0.05 is accepted. The last column in Table 2.3 shows how high this error, which can be described as health risk, is for the different sampling schemes. Both S^1 and S^2 are within this certainty level. S^3 does not reach the demand, because it is very inefficient in remediating polluted soil: to remediate 0.1% more polluted soil (as compared to S^2), around 10% of unpolluted soil had to be remediated. In summary, S^2 performed better than S^3 and S^4 , whereas S^1 failed to predict several small polluted areas. The 60 meter grid spacing of the first stage of S^1 is probably too wide for this purpose.

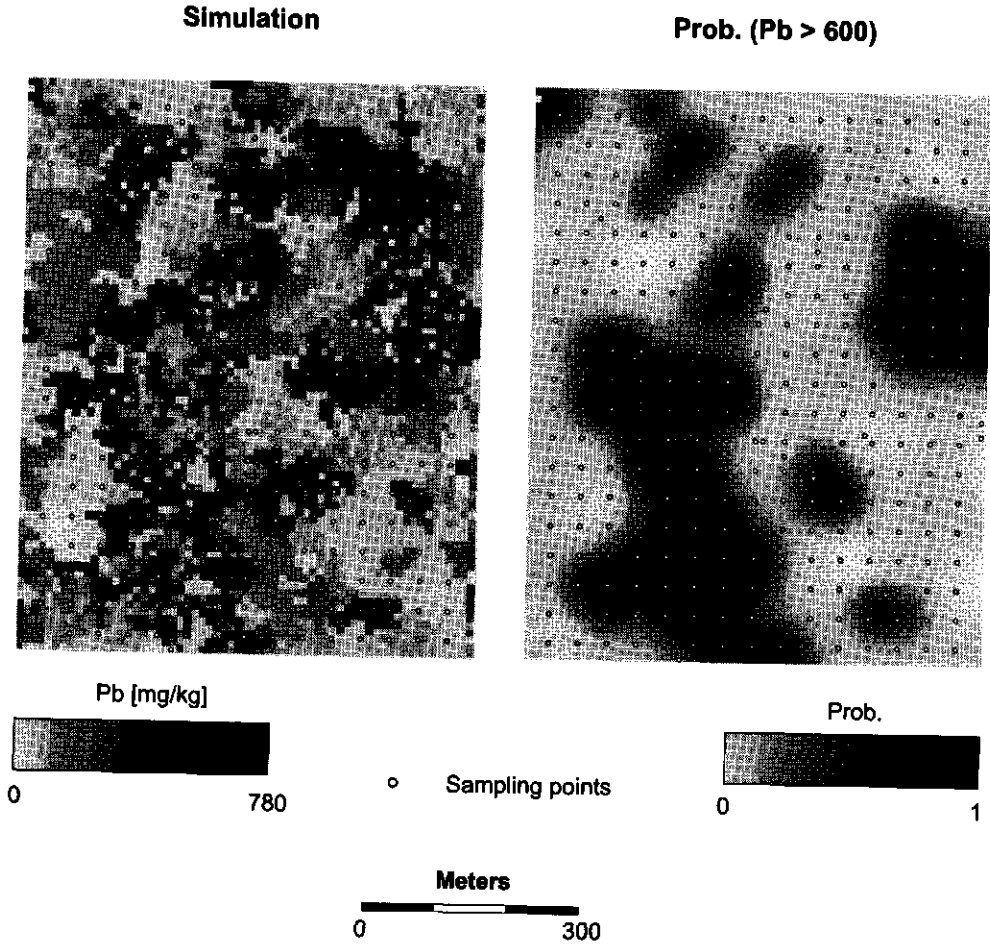


Figure 2.6: Simulated lead contents for simulation 3 (left) and predicted probability of exceeding intervention level z_1 with S^3 (right).

2.4. Discussion

Phased sampling, as applied in the Schoonhoven study, was conducted in three stages, comprising nearly 200 data-points. The pollution showed a large spatial variation, which made it difficult to make good predictions. To assess the benefits of staged sampling for this case, successful use could be made of conditional simulation procedures:

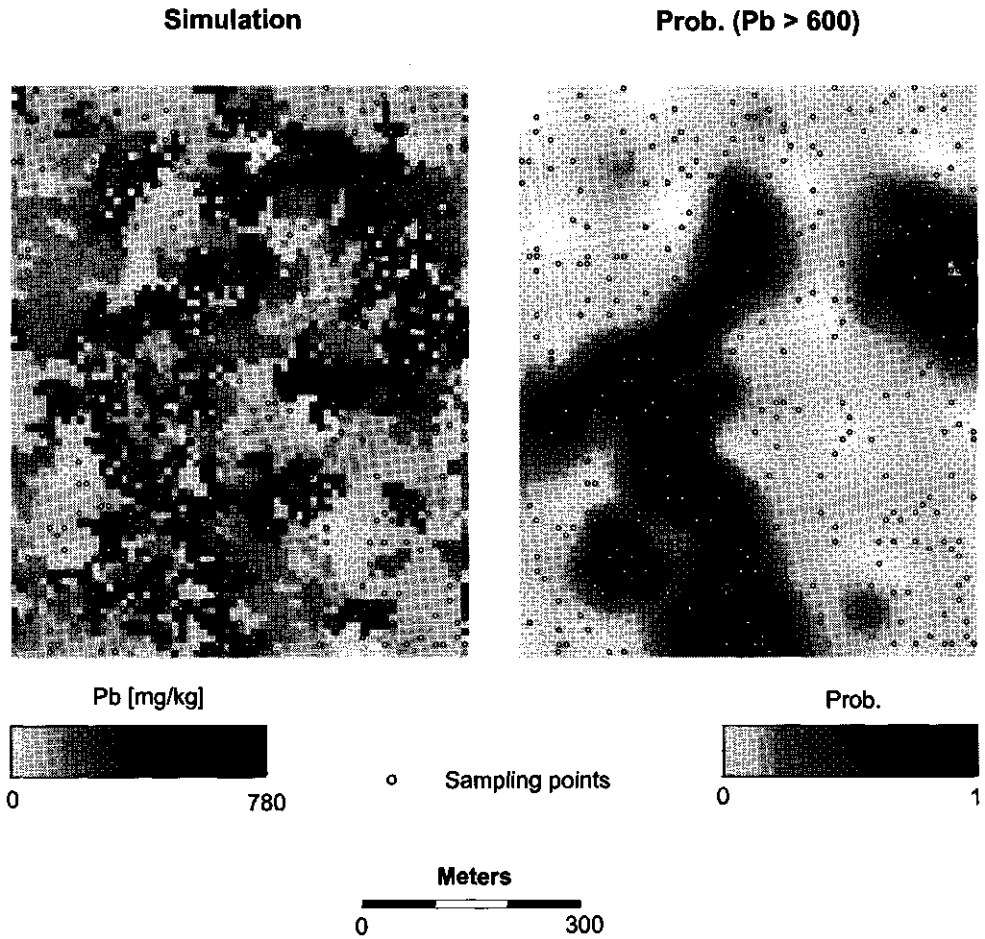


Figure 2.7. Simulated lead contents for simulation 4 (left) and predicted probability of exceeding intervention level z_1 with S^4 (right).

- i) to quantify the quality of the predictions in terms of type-I and type-II errors made by the sampling schemes and by using these simulated fields as references when quantifying the quality of the predictions.
- ii) to show that phased sampling schemes were superior to classical schemes, by comparing different sampling schemes on the same conditional simulations.

Simulated pollution fields showed good results for phased sampling, in particular for S^2 . A much larger unpolluted area is left unremediated, while the health risk defined

as the fraction of polluted soil in the not remediated area does not exceed the 0.05-limit. The two non-phased schemes (S^3 and S^4) delineate a much larger area as possibly polluted.

Use of conditional simulations has some drawbacks. In this study the simulations are made for grid-cells with a 10 m width and hence represent a continuous variable on a discrete grid. However, it is still a close approximation. Also, resampling on simulated fields allows a perfect sampling grid. A grid applied in practice is often disturbed by houses, roads, *etc.* Errors in location during sampling were not simulated. A statistical problem connected with using phased sampling, is that it may lead to a biased data set, as the samples will be more prevalent in areas with a higher pollution. In this study, use was made of a declustering algorithm to correct for this effect. Declustering, however, corrects only for univariate statistical properties, such as the cumulative distribution function, and not for bivariate ones, such as the variogram (Deutsch and Journel, 1992).

An alternative would be to use only the variograms of stage 1. Because the data set derived from the first stage of phased sampling is unbiased, the resulting variograms are also unbiased. But use of variograms derived from a small number of data reduces the accuracy severely.

Further research is needed to determine the selection of additional sampling points. In this study only simple random sampling was applied for that purpose. Not much research is being done on this problem of optimal allocation, although several methods are being proposed (Cressie, 1991). As was shown in this study, phased sampling is attractive and useful to select areas where a quantitative spatial variable exceeds a fixed environmental threshold.

2.5. Acknowledgements

The authors are grateful to the province of Zuid-Holland for making the data available. Also contributions by mr. K. Kappen and mr. K. Asmus are gratefully acknowledged.

Chapter 3

Spatial Simulated Annealing¹

Abstract

In this chapter, Spatial Simulated Annealing (SSA) is presented as a method to optimise spatial sampling schemes. Sampling schemes are optimised at the point-level, taking into account sampling constraints and preliminary observations. The method is illustrated by two optimisation criteria. The first optimises even spreading of the points over a region, whereas the second optimises variogram estimation using a proposed criterion from the literature. For several examples it is shown that SSA is superior to conventional methods of designing sampling schemes. Improvements up to 30% occur for the first criterion, and an almost complete solution is found for the second criterion. SSA is especially useful in studies with many sampling constraints. It is flexible in implementing additional, quantitative criteria.

¹ *Published as:* Van Groenigen, J.W. and Stein, A. (1998). *Constrained optimisation of spatial sampling using continuous simulated annealing*. *Journal of Environmental Quality* 27: 1078-1086.

3.1. Introduction

Characterisation and remediation of polluted soils is a major environmental activity, requiring large amounts of money and man-power. For example, in the Netherlands alone, it was estimated that in urban regions 110,000 sites would have to be surveyed and possibly remediated, at a cost of approximately 50 billion Dutch guilders (about 25 billion US\$). This included almost every former industrial site (RIVM, 1991). Spatial sampling is a crucial activity in such studies. The size and nature of the sampling scheme is strongly influencing the costs of the survey and the reliability of its results (Winkels and Stein, 1997). Therefore, a careful design of the sampling scheme can lead to considerable savings of time and money. Spatial statistics can assist in optimising such a sampling scheme.

This chapter will be focussed on pollutants that can be described as a Regionalized Variable (ReV). More formally, consider a variable $Z(\bar{x})$ depending upon the 1-, 2-, or 3-dimensional vector \bar{x} in region A . This variable may denote the lead pollution on a former factory ground, the Phosphate saturation along a transect, the thickness of a contaminated layer, *etc.* For optimal sampling, we have to decide upon the number of observations n and the locations of these observations x_1, \dots, x_n . In this chapter we will address the problem of defining a sampling scheme $S = \{x_1, \dots, x_n\}$, such that a well-defined, quantitative criterion $\phi(S)$ is optimised.

In all environmental surveys, the size of A is finite, *i.e.* it is delimited by boundaries. Most often, historic information is available on previous use of the region, indicating already useful delineations. In addition, previous observations may be available, say S^1 , as soil remediation is typically a process that involves several stages of surveying and decision making. Hence $S = S^1 \cup S^2$, where only S^2 has to be optimised. Both the boundaries and the previous observations have to be considered for optimising a sampling scheme in practical conditions.

This chapter describes a method that optimises spatial sampling schemes, taking into account physical sampling constraints and delineations, as well as previous measurements.

3.2. Optimising sampling using geostatistics

In the past, various ways of optimising spatial sampling have been proposed. Using classical sampling theory, Thompson and Seber (1996) derived estimators for finite populations that are distributed in space and proposed adaptive sampling strategies to estimate scattered populations. They found that by taking into account scattering of

populations, estimates of population size could be improved. De Gruijter and ter Braak (1990) used classical sampling theory to estimate the spatial means of continuous variables.

Geostatistics offers methods for interpolation and analysis of ReV's $Z(\bar{x})$ (Matheron, 1973). Furthermore, its ability to provide risk-qualified predictors of exceeding threshold values is an indispensable tool for environmental decision-making (Cressie, 1991; Journel, 1983).

In the past, much geostatistical research has been dedicated to optimisation of the sampling scheme for estimation of the variogram $\gamma(\bar{h})$, that characterises the spatial correlation of a ReV. The variogram can be calculated if the mathematical expectation of the ReV exists and does not depend on the location \bar{x} :

$$E\{Z(\bar{x})\} = m, \quad \forall \bar{x} \quad (3.1)$$

(Journel and Huijbrechts, 1978). The variogram is defined as:

$$\gamma(\bar{h}) = \frac{1}{2} E\{Z(\bar{x}) - Z(\bar{x} + \bar{h})\}^2. \quad (3.2)$$

It describes spatial dependence as a function of separation vector \bar{h} . The n collected observations are denoted by $z(\bar{x}_1), \dots, z(\bar{x}_n)$. We distinguish between a stochastic variable that is denoted by capitals, and observations denoted by lower case characters. The variogram can be estimated for any distance or direction class \bar{h} by

$$\hat{\gamma}(\bar{h}) = \frac{1}{2n(\bar{h})} \sum_{i=1}^{n(\bar{h})} \{z(\bar{x}_i) - z(\bar{x}_i + \bar{h})\}^2, \quad (3.3)$$

where $n(\bar{h})$ is the number of pairs of points with a separation vector approximately equal to \bar{h} .

Optimising the sampling scheme for estimation of the variogram is not an easy task. Every individual observation may contribute to pair differences in different distance classes. Russo (1984) and Warrick and Myers (1987) optimised for distribution of the points pairs over the distance and direction classes. An ideal distribution was decided upon *a priori*. Russo and Jury (1988) used simulations of ReV's to evaluate this procedure. They found that the procedure yielded better short range estimates of the variograms than those obtained by systematic sampling. Using a similar evaluation procedure, Corsten and Stein (1994) showed that nested sampling designs produced inaccurate experimental variograms, as compared to random and systematic sampling designs.

Brus and de Gruijter (1994) showed how classical sampling theory could provide the error variance of the local or non-ergodic variogram. Webster and Oliver (1992) found that at least 100 sampling points are needed for estimating the variogram, and 225 to estimate it reliably. Yfantis *et al.* (1987) compared the performance of square, equilateral triangular, and hexagonal grids. They found that the equilateral triangular grid yielded the most reliable estimation of the variogram.

Given the variogram, kriging can be used for spatial interpolation, by means of predicting the values at the nodes of a fine-meshed grid (Deutsch and Journel, 1992). Although kriging provides the best linear unbiased predictor of the ReV at any unvisited location, the sampling scheme is still of crucial importance. Ill-designed sampling schemes can result in much higher costs of sampling (as more samples are necessary to achieve the same precision), higher costs of required action (such as remediation) because of imprecise predictions, and even unnecessary health risks in the case of environmental pollution (Stein *et al.*, 1995).

McBratney *et al.* (1981) presented procedures for optimising the spacing of a sampling grid for minimisation of the kriging variance, given an *a priori* variogram. McBratney and Webster (1983) did the same for the co-kriging variance. Yfantis *et al.* (1987) found that if the variogram had a relatively low nugget, equilateral triangular grids yielded more reliable kriging predictions than either hexagonal or square grids. When the nugget is relatively high and the sampling density is relatively scarce, they found that the hexagonal grid yielded the lowest kriging variances. Christakos and Olea (1992) refined these findings, and presented a case-specific methodology for choosing between the different grid designs. Sacks and Schiller (1988) presented several annealing-based algorithms for optimising a sampling scheme out of a small grid of possible sampling locations. They distinguished between several optimisation criteria, among which the minimisation of the mean kriging variance, and minimisation of the maximum kriging variance. Watson and Barnes (1995) defined several optimisation criteria for infill sampling of an existing network, focussing on dealing with extremes. However, they did not present a suitable optimisation algorithm to apply these criteria, and used the criteria only in a very limited, simulated case study.

Van Groenigen *et al.* (1997) (Chapter 2) proposed an interactive sampling procedure for characterising health risks in urban regions. Using probability maps of environmental threshold values calculated using indicator kriging, additional sampling was focused on areas with imprecise predictions.

In this chapter, a procedure is proposed to optimise sampling schemes for different quantitative optimisation criteria, taking into account physical sampling barriers and earlier measurements. Using this procedure, two optimisation criteria are evaluated, representing different definitions of optimality.

The Minimisation of the Mean of Shortest Distances (MMSD) criterion aims at regular spreading of all sampling points over the sampling region. Regular spreading

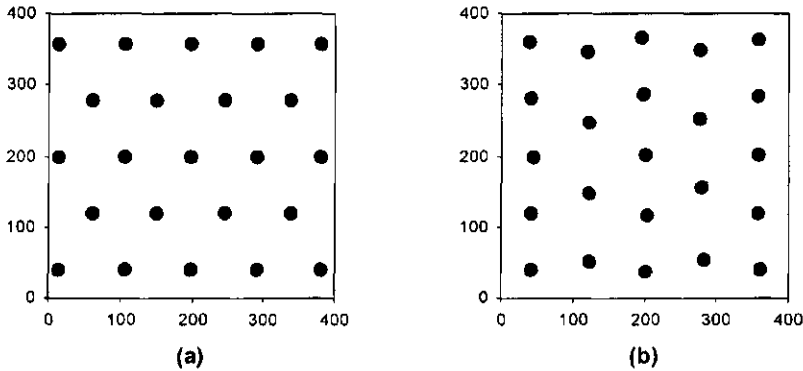


Figure 3.1. An optimised sampling scheme using a triangular grid (a) and using SSA with the MMSE criterion (b)

can be formulated as minimising the expectation of the distance between an arbitrarily chosen point within the region, and its nearest sampling point. For sampling scheme S , minimising this expectation leads to the following minimisation function:

$$\min_S \int_A \|\bar{x} - V_S(\bar{x})\| \quad (3.4)$$

where \bar{x} is a two-dimensional location vector, and $V_S(\bar{x})$ denotes the location vector of the nearest sampling point $\bar{x}_i \in S$. An equilateral triangular grid (Figure 3.1a) optimises this criterion in theory (Flatman and Yfantis, 1984). In the practice of spatial sampling many constraints to this are met. We will give an example from soil sampling in an urban region (Figure 3.3), where several types of area-specific information are encountered during sampling:

- i) The area of the sampling region is finite. Therefore, there are boundary-effects which will make almost any regular grid sub-optimal (Christakos and Olea, 1992).
- ii) The sampling region can be composed of several sub-regions, different in size and shape. Some of these sub-regions may be impossible to sample. Within these regions which can not be sampled, we made a division into a research sub-region and a non-research sub-region. For remediation studies, non-research sub-regions may include ponds and regions that were remediated previously, whereas research sub-regions may include houses, below which the extent of the contamination should be predicted.
- iii) Measurements from a preliminary study are available, and should be included in the optimised sampling scheme.

These features ensure that in practice a regular grid often is not an optimal solution with respect to the MMSD-criterion, or can not be realised because of sampling constraints.

The Warrick and Myers (WM-) criterion optimises the fit of the realised distribution of point pairs for the experimental variogram to an *a priori* defined, ideal distribution (Warrick and Myers, 1987). For a sample size of n , the number of point pairs equals $\frac{1}{2}n(n-1)$. The variogram is estimated for each distance and (possibly) direction class using Equation (3.3). If the number of lag classes equals nc , the WM-criterion evaluates the following expression:

$$\sum_{i=1}^{nc} [a \cdot (\zeta_i^* - \zeta_i)^2 + b \cdot \sigma_i] \quad (3.5)$$

where a and b are user-defined weights, and ζ_i denotes the number of point pairs in the i^{th} lag class, with optimal values of ζ_i specified in ζ_i^* . The second term denotes the standard deviation σ_i of the point pairs from the median of class i . Warrick and Myers minimised Equation (3.5) using a Monte Carlo algorithm that could also include earlier sampling points. However, this algorithm was susceptible to local minima, and could not handle complex sampling barriers (Warrick and Myers, 1987).

3.3. Simulated Annealing

Simulated annealing (SA) is a combinatorial optimisation algorithm, originating from statistical physics. It was developed independently by Kirkpatrick *et al.* (1983) and Cerny (1985). Other names for the same algorithm include Monte Carlo annealing, probabilistic hill climbing, statistical cooling and stochastic relaxation (Aarts and Korst, 1989). In many studies, it has been applied successfully as a universal optimisation method (*e.g.* Goldstein and Waterman, 1987). It is also widely applied in geostatistics for simulation of ReV's (Deutsch and Cockerham, 1994; Deutsch and Journel, 1992; Goovaerts, 1996). Related algorithms have been applied to optimisation of spatial sampling (Sacks and Schiller, 1988) and to the restoration of degraded images (Geman and Geman, 1984). One of its properties is its insensitivity to local extremes.

A central concept in SA is the fitness function $\phi(S)$ that has to be optimised. Suppose that we can define a combinatorial optimisation problem in which $\phi(S)$ has to be minimised. Starting with S_0 , let S_i and S_{i+1} represent two solutions with fitness $\phi(S_i)$ and $\phi(S_{i+1})$, respectively. Typically, S_{i+1} is derived from the neighbourhood of

S_i by a random perturbation of one of the variables of S_i . A probabilistic acceptance criterion decides whether S_{i+1} is accepted or not. This probability $P_c(S_i \rightarrow S_{i+1})$ of S_{i+1} being accepted can be described as:

$$P_c(S_i \rightarrow S_{i+1}) = 1, \quad \text{if } \phi(S_{i+1}) \leq \phi(S_i)$$

$$P_c(S_i \rightarrow S_{i+1}) = \exp\left(\frac{\phi(S_i) - \phi(S_{i+1})}{c}\right), \quad \text{if } \phi(S_{i+1}) > \phi(S_i) \quad (3.6)$$

where c denotes a positive control parameter. The parameter c is lowered according to a cooling schedule as the process evolves, to find the global minimum. A transition takes place if S_{i+1} is accepted. Next, a solution S_{i+2} is derived from S_{i+1} , and the probability $P_c(S_{i+1} \rightarrow S_{i+2})$ is calculated with a similar acceptance criterion as Equation 3.6 (Aarts and Korst, 1989).

A mathematical description of the SA-algorithm is given by the theory of finite Markov chains (Seneta, 1981). At each value of c , several transitions have to be made before the annealing can proceed, and c can take its next value (Aarts and Korst, 1989).

3.4. Spatial Simulated Annealing

Sacks and Schiller (1988) proposed several SA-related algorithms for optimising spatial sampling schemes using geostatistical criteria. Although this research was related to the proposed method, some crucial differences will be covered in the description of the algorithm below.

In order to modify simulated annealing for optimisation of spatial sampling, the fitness function, a generation mechanism and the cooling scheme have to be decided upon (Aarts en Korst, 1989). They are discussed below.

3.4.1. Fitness functions

Let the total research region be denoted by A_R and the sub-region that can be sampled by $A_S \subset A_R$, thus excluding roads, houses *etc.* Next, the MMSD criterion is estimated by the fitness function $\phi_{\text{MMSD}}(S)$, which is an estimator of the function formulated in Equation (3.4):

$$\phi_{\text{MMSD}}(S) = \frac{1}{n_c} \sum_{j=1}^{n_c} \|\bar{x}_c^j - V_S(\bar{x}_c^j)\|, \quad (3.7)$$

with $\bar{x}_z^j \in A_R$ denoting the j^{th} evaluation point. The n_c evaluation points are located on a finely meshed grid over the whole area. In order to yield a reliable value of $\phi_{\text{MMSD}}(\mathbf{S})$, the number of evaluation points should be much higher than the number of sampling points.

By choosing the evaluation points on a finely meshed grid over the whole region A_R , while locating the sampling points strictly in A_S , the algorithm spreads the sampling points optimally over the whole region A_R , while taking physical sampling constraints into account. This is an important difference with the method of Sacks and Schiller, which could not handle such sampling constraints.

The WM-criterion solely depends upon the distances between sampling points. Therefore, the fitness function can be directly calculated using Equation 3.5:

$$\phi_{\text{WM}} = \sum_{i=1}^{nc} [\mathbf{a} \cdot (\zeta_i^* - \zeta_i)^2 + \mathbf{b} \cdot \sigma_i] \quad (3.8)$$

where \mathbf{a} and \mathbf{b} are user-defined weights that can be used to define the relative importance of the two parts of the function (Warrick and Myers, 1987).

3.4.2. Generation mechanism

The aim of a generation mechanism is to generate a new solution S_{i+2} out of the solution S_{i+1} , by means of a random perturbation in one of the variables of S_{i+1} (Davis, 1990). In SSA, this is done by moving one randomly chosen sampling point \bar{x}_i over a vector $\bar{\mathbf{h}}$, with the direction of $\bar{\mathbf{h}}$ drawn randomly, and $\|\bar{\mathbf{h}}\|$ taking a random value between 0 and $\bar{\mathbf{h}}_{\text{max}}$. One of the modifications of SSA as compared to ordinary SA and the method of Sacks and Schiller, is that $\bar{\mathbf{h}}_{\text{max}}$ initially is equal to half the length of the sampling region, and decreases with time. This increases the efficiency of the demanding recalculations after each modification in the sampling scheme, because it can be expected that with optimisation of sampling schemes, successful modifications consist of increasingly smaller values of $\|\bar{\mathbf{h}}\|$ as the SSA process advances. This is because the process deals with many similar variables (*i.e.* the co-ordinates of the sampling points). Therefore, moving sampling points randomly over large distances will not contribute much to finding the minimum towards the end of the optimisation process. Furthermore, contrary to earlier optimisation methods, co-ordinates of the sampling points are treated as continuous variables, rather than chosen from a discrete grid. This is in line with earlier studies, where SA was applied to continuous problems (*e.g.* Bohachevsky *et al.*, 1986; Vanderbilt and Louie, 1984). At the final value of the

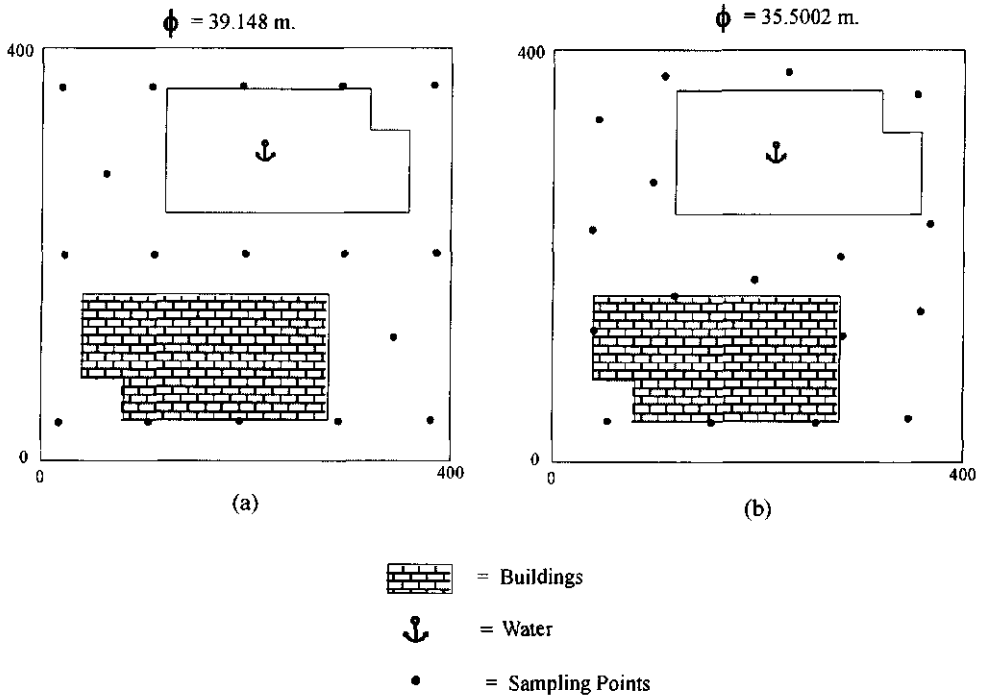


Figure 3.2. A sampling scheme with two enclosures of a different nature, using a triangular grid (a) and using SSA with the MMSD criterion (b)

control parameter, \bar{h}_{\max} will be almost equal to zero.

3.4.3. Cooling schedule

For the cooling schedule, which expresses c as a function of the progress of the optimisation, we used a basic set of empirical rules which have been proposed in many studies (e.g. Kirkpatrick *et al.*, 1983; Aarts and Korst, 1989). We start with an initial value c_0 which has an acceptance ratio of 0.95 or higher for alternative solutions. The decrement of c is given by

$$c_{k+1} = \alpha \cdot c_k, \quad k = 1, 2, \dots, \tag{3.9}$$

with $0 < \alpha < 1$. The maximum period of time for one Markov chain k to remain at any value of c is fixed, and the final value of c is explicitly given to the SSA algorithm.

From these data, α can be calculated. The acceptance criterion is similar to the one given in Equation 3.6, substituting ϕ_{MMSD} for ϕ . Use of a variable c which ensures that inferior solutions are accepted with decreasing probability as the process evolves, is the most important difference with the proposed algorithms by Sacks and Schiller (1988).

The spatial character of the problem of designing sampling schemes made it logical to use a geographical information system (GIS) (Burrough, 1993). We used the GRID module of ARC/INFO to describe the sampling barriers. Calculations on Spatial Simulated Annealing were done with the specially developed software package SANOS (Simulated ANnealing for Optimisation of Sampling), which will be made available in the future in a decision support system for soil sampling (Domburg *et al.*, 1997).

3.5. Examples

The examples which we consider here were encountered during a recent study on lead pollution in an urban region (Van Groenigen *et al.*, 1997) (Chapter 2). The optimised sampling schemes using the MMSD-criterion will be compared to a traditional equilateral triangular grid. The sampling schemes produced using the WM-criterion will be compared to the optimised sampling schemes as presented by Warrick and Myers (1987). Additionally, the WM-criterion was used to design a D-optimal sampling scheme for fitting the experimental variogram (Rasch, 1990).

3.5.1. Examples of the MMSD-Criterion

Figure 3.1a shows an equilateral triangular grid of 23 points which in theory is optimal with respect to the MMSD-criterion. The region is a square of 400 x 400 m and the evaluation points were chosen on a 4 m grid, yielding 10000 points. The ϕ_{MMSD} of this scheme is 32.668 m. Figure 3.1b shows the solution using SSA, starting with a random sampling scheme also consisting of 23 sampling points. This optimised scheme closely resembles the triangular equilateral grid, but with small deviations caused by boundary constraints. The scheme is point symmetric. The ϕ_{MMSD} for this scheme is 31.870 m, which shortens the mean distance to sampling points by 2.4%, as compared to the triangular grid.

The second example illustrates the effects of research- and a non-research sub-regions that can not be sampled. Figures 3.2a and 3.2b show a research region A_R for soil remediation. The upper sub-region (water) does not belong to A_R nor A_S , while the lower sub-region (a building) is part of A_R but not of A_S . Figure 3.2a shows a sampling scheme based on an equilateral triangular grid, from which the points which

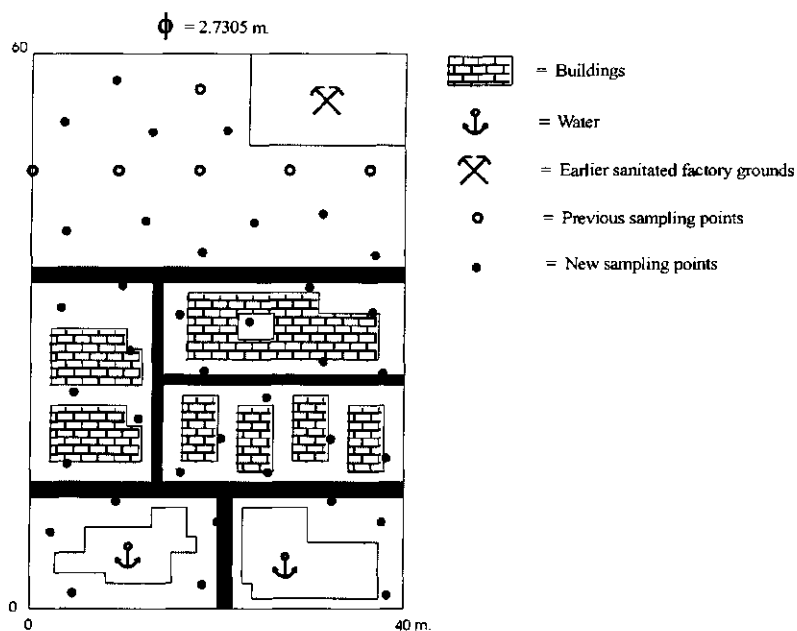


Figure 3.3. An urban sampling area with typical sampling constraints and preliminary observations. The sampling scheme is optimised using SSA with the MMSD criterion.

could not be sampled have been deleted, resulting in a symmetrical sampling scheme of 17 points, and a ϕ_{MMSD} of 39.148 m. Figure 3.2b shows the solution using SSA, also using 17 points. This scheme scores 35.500 m, which is an improvement of 8%. From the figures it appears that the main advantage of SSA is that it recognises the differences between the upper and lower sub-regions. In the lower sub-region (the building) the points are located at the boundaries to cover it optimally given the sampling constraints. In the upper part (the pond) the points are located at some distance from the sub-region. This is because of the differentiation between the whole research region A_R , and the part of the research area that can be sampled A_S , as explained above. Here, the use of prior information has an effect both on the sampling scheme and on the obtained optimum.

Figure 3.3 displays an urban region, where a complex set of prior information is available, including buildings, already remediated factory grounds, ponds and previously sampled points. For optimisation using SSA, the previously sampled points are included as a fixed subset of the sampling scheme S . Figure 3.3 shows the sampling scheme calculated with SSA, for 37 additional sampling points. This solution has a ϕ_{MMSD} of 2.7305 m. A sampling scheme based upon an equilateral grid, also consisting of 37

Table 3.1. Evaluation of the fitness function for the sampling design obtained with the SSA algorithm, as compared to a triangular grid.

Case	n	n _e	ϕ_{MMSD}		Calculation time SSA (s)
			triangular grid	SSA	
1	23	20000	0.8167	0.8009	1260
2	17	20000	0.9787	0.8875	2040
3	37	20000	3.9018	2.7305	18000

additional points has a ϕ_{MMSD} of 3.9018 m. This amounts to an improvement of approximately 30% for SSA.

Use of SSA takes into account the previously sampled points. The sampling points are often located close to the boundaries of regions belonging to A_R but not to A_S , such as buildings and roads, whereas they maintain a larger distance from the regions which belong to neither A_R nor A_S (water and already remediated former factory grounds). Figure 3.4 shows the ϕ_{MMSD} as a function of k (Equation 3.9), during optimisation of Figure 3.1b. There is a tendency towards a global minimum, with local minima being avoided. Table 3.1 summarises the performance of the SSA algorithm for the MMSD-criterion, including calculation times with a Pentium 120 MHz PC. It shows reasonable calculation times, although they increase with the complexity of the problem.

3.5.2. Examples of the WM-Criterion

To test SSA for the WM-criterion, four cases from Warrick and Myers (1987) were recalculated. The emphasis in that paper was on the formulation of the optimisation criterion, not on the Monte Carlo algorithm. The use of SSA for these problems is therefore a logical extension of this work. Additionally, a sampling scheme is optimised for modelling of the experimental variogram, using the WM-criterion. All cases assume a 400 x 400 m field.

Case 4 considers 16 sampling points as pre-information, distributed according to a rectangular grid. A set of 14 additional points has to be placed optimally according to the first term in Equation 3.8 (with $\mathbf{a} = 1$ and $\mathbf{b} = 0$). Ten distance classes were distinguished of 20 m width each. A uniform distribution among these classes was defined as optimal:

$$\zeta_i^* = \frac{n_s \cdot (n_s - 1)}{2 \cdot n_c} = 43.5, \quad \forall i \in 1, \dots, n_c. \quad (3.10)$$

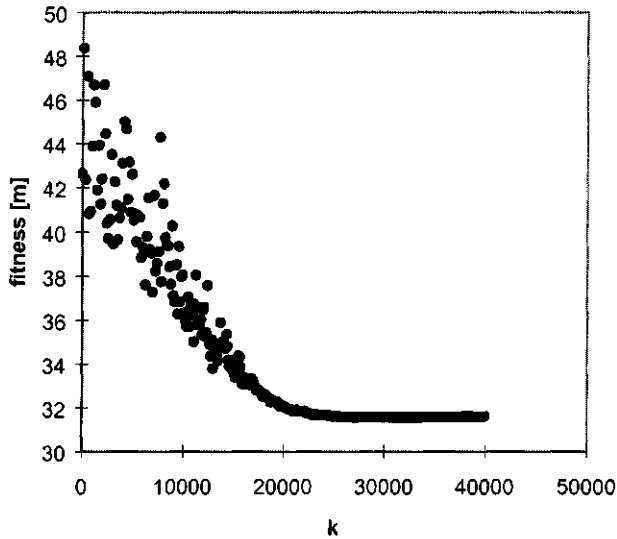


Figure 3.4. The optimisation process, ϕ_{MMSD} as a function of the number of Markov chains k .

The results of the optimisation are given in Figure 3.5a and in Table 3.2. Although a completely uniform solution is impossible because of the predefined grid, we observe a considerable improvement compared to the Monte Carlo algorithm, as the minimum number of point pairs in a class rises from 17 to 29 (Table 3.2).

The fifth case involves 50 sampling locations of which the distribution was optimised according to the first term in Equation 3.8 (with $a = 1$ and $b = 0$). Again, a uniform distribution was defined as optimal:

$$\zeta_i^* = \frac{n_s \cdot (n_s - 1)}{2 \cdot nc} = 40.83, \quad \forall i \in 1, \dots, nc, \quad (3.11)$$

with results given in Figure 3.5b and Table 3.3. This is an almost optimal solution, as all classes consist of either 40 or 41 point pairs, and only one pair of points falls outside the lag range.

The sixth case, consisting of 30 sampling locations, considers optimisation according to the second term in Equation 3.8 (with $a = 0$ and $b = 1$). This term denotes the mean deviation of the point pairs from the class median. The classes are the same as in the fourth case. An additional constraint was added to ensure that each class would contain at least 25 point pairs (Figure 3.5c and Table 3.2). The mean deviation

Table 3.2. *Distribution of point pairs over the lag classes (case 4) and mean deviation from the median of the distance class (case 6).*

Distance class (m)	case 4		case 6	
	Monte Carlo	SSA	Monte Carlo	SSA
0 - 20	17	30	25	37
20 - 40	18	30	29	28
40 - 60	17	30	30	54
60 - 80	19	29	46	55
80 - 100	32	32	51	26
100 - 120	51	29	25	52
120 - 140	31	29	28	50
140 - 160	46	35	28	68
160 - 180	23	37	25	40
180 - 200	26	42	25	25
> 200	155	112	123	0
m_{avg}	-	-	3.9	2.3

from the median of the distance classes reduced to 2.35 m, which is a considerable improvement compared to the 3.9 m obtained originally (Warrick and Myers, 1987). It is interesting to notice that the optimised scheme is not satisfactory in this case; although the mean deviation is very low, several sampling points are placed at the same location. This means that the class 0 - 20 m has a very high standard deviation (due to the many distances of 0), in order to optimise the overall performance of the sampling scheme. This second part of the WM-criterion is therefore in its present form not very useful for optimising spatial sampling schemes.

In the seventh case, 30 sampling locations were optimally located according to the same distance classes as in the fourth case, as well as to two direction classes equal to $0^\circ \pm 45^\circ$, and $90^\circ \pm 45^\circ$, respectively. A uniform distribution was defined as optimal:

$$\zeta_i^* = \frac{n_s \cdot (n_s - 1)}{2 \cdot nc} = 21.75, \quad \forall i \in 1, \dots, nc \quad (3.12)$$

Figure 3.5d and Table 3.4 show that the solution is optimal because no point pairs fall outside the class-range, and all combinations of distance and direction classes contain either 21 or 22 point pairs. This is a considerable improvement compared to

Table 3.3. *Distribution of point pairs over the distance classes generated by the SSA algorithm, compared to the Monte Carlo method.*

Case 5					
Class (m)	Monte Carlo	SSA	Class (m)	Monte Carlo	SSA
0 - 15	33	41	240 - 255	35	40
15 - 30	36	41	255 - 270	36	41
30 - 45	37	41	270 - 285	41	40
45 - 60	37	41	285 - 300	42	40
60 - 75	40	41	300 - 315	37	40
75 - 90	43	41	315 - 330	39	41
90 - 105	43	41	330 - 345	44	41
105 - 120	45	41	345 - 360	45	41
120 - 135	41	41	360 - 375	44	41
135 - 150	41	41	375 - 390	42	41
150 - 165	37	41	390 - 405	36	41
165 - 180	37	41	405 - 420	41	41
180 - 195	35	41	420 - 435	41	41
195 - 210	39	41	435 - 450	40	41
210 - 225	46	41			
225 - 240	47	40	> 450	25	0

the earlier solutions with a minimum and maximum of 8 and 28, respectively.

Finally, the WM-criterion was used to optimise a sampling scheme for modelling of the experimental variogram. This was done using the D-optimality criterion for fitting of non-linear functions (Rasch, 1990). The use of D-optimality for designing sampling networks has been proposed in the past (Zimmerman and Homer, 1991). To optimise fitting of an exponential variogram model

$$\gamma(\mathbf{h}) = c_0 + c(1 - \exp(-\mathbf{h}/r)), \quad (3.13)$$

where c_0 denotes the nugget, $c_0 + c$ the sill, and $3r$ the effective range, estimators of the variogram at three distance lags $\hat{\gamma}(\bar{\mathbf{h}}_1)$, $\hat{\gamma}(\bar{\mathbf{h}}_2)$ and $\hat{\gamma}(\bar{\mathbf{h}}_3)$ are needed, with

$$\bar{\mathbf{h}}_1 \approx 0$$

$$\bar{\mathbf{h}}_3 = 3r$$

and

$$h_2 = r + \frac{h_1 \cdot \exp(-h_1/r) - h_3 \cdot \exp(-h_3/r)}{\exp(-h_1/r) - \exp(-h_3/r)} \quad (3.14)$$

The point pairs should be equally distributed over the 3 lag classes:

$$\zeta_1^* = \zeta_2^* = \zeta_3^* = \frac{n \cdot (n-1)}{6 \cdot n} \quad (3.15)$$

If we assume that for a particular ReV an exponential variogram model holds with $3r = 200$ m, then Equation 3.14 requires that \bar{h}_2 should be equal 84 m. This requirement leads to a scheme with three sharply delineated clusters, almost exactly reproducing the required point pair distribution. Table 3.5 shows the results for a sample region with $n = 50$. The scheme is extremely clustered, raising considerable doubts on its practical use.

3.6. Discussion and conclusions

In this chapter, we presented the Spatial Simulated Annealing (SSA) procedure to optimise spatial sampling schemes. We found that SSA provides a robust algorithm, being insensitive to local minima. It can easily be adapted to sampling constraints met in practice. SSA translates various optimisation criteria into actual optimal sampling schemes. Therefore its application yields better sampling schemes and more insight into the implications of the choice of any quantitative optimisation criterion. In the future, other optimisation criteria will be added to SSA. In particular minimisation of the kriging variance, as presented in Sacks and Schiller (1988), may prove a valuable tool in environmental studies, where accuracy of predictions is a crucial issue. So far, the number of sampling points was kept fixed, and no cost models of the sampling schemes were considered. This may not always be a realistic assumption. In the future, we aim at integration with existing cost models as in Domburg *et al.* (1997).

We showed that SSA is superior to current methods of designing spatial sampling schemes for two different criteria. For the MMSD-criterion, which aims at even spreading of the observations over the region, SSA is most beneficial when prior spatial information is available. This may apply to a wide range of soil pollution studies in urban regions, where typically many sampling constraints are met. In such cases, SSA can enhance efficiency up to 30%. In the more theoretical examples (Figures 3.1

Table 3.4. *Distribution of point pairs over the distance and direction classes generated by the SSA Algorithm, compared to the Monte Carlo Method*

Case 7					
<i>class (m)</i>	<i>Monte Carlo</i>	<i>SSA</i>	<i>class (m)</i>	<i>Monte Carlo</i>	<i>SSA</i>
<i>H, 0 - 20</i>	12	22	<i>V, 100 - 120</i>	22	22
<i>V, 0 - 20</i>	8	22	<i>H, 120 - 140</i>	20	22
<i>H, 20 - 40</i>	22	21	<i>V, 120 - 140</i>	20	21
<i>V, 20 - 40</i>	20	22	<i>H, 140 - 160</i>	26	22
<i>H, 40 - 60</i>	18	22	<i>V, 140 - 160</i>	28	22
<i>V, 40 - 60</i>	24	21	<i>H, 160 - 180</i>	27	22
<i>H, 60 - 80</i>	23	22	<i>V, 160 - 180</i>	27	21
<i>V, 60 - 80</i>	27	22	<i>H, 180 - 200</i>	17	22
<i>H, 80 - 100</i>	22	22	<i>V, 180 - 200</i>	20	22
<i>V, 80 - 100</i>	22	21			
<i>H, 100 - 120</i>	25	22	<i>> 200</i>	5	0

and 3.2) the relative improvement due to SSA is probably not enough to justify use, as site location is generally more difficult than when a regular grid is used.

For the WM-criterion, which aims at optimal estimation of the variogram, SSA was always superior to currently available Monte-Carlo optimisation, and solves one well-defined case almost completely. Use of SSA for designing sampling schemes can therefore improve variogram estimation. Further, we found that the second term of the WM-criterion is not satisfactory, as it yields degenerate sampling schemes with several sampling points at the same locations.

It was shown how the WM-criterion could be used for designing a D-optimal sampling scheme, yielding however a strongly clustered scheme. The question remains whether such a scheme is desirable, as it only gives information on a very small part of the research region and prior information on the variogram must be available. Also, the high number of point pairs may be misleading. If used for kriging purposes, a declustering algorithm should be applied (Deutsch and Journel, 1992). The WM criterion should only be used in cases where some additional sampling points can be added to an existing sampling scheme in order to improve variogram estimation. An example of such a case is provided in Figure 3.5a.

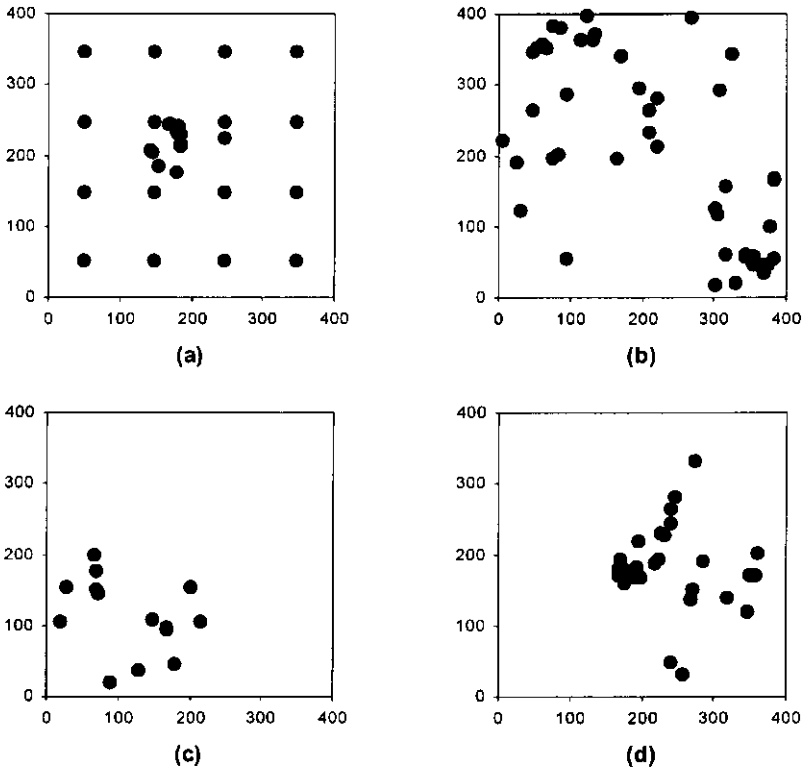


Figure 3.5. Optimal sampling schemes in a 400x400 m sampling region, according to the Warrick/Myers criterion, (a) 16 fixed and 14 variable points, optimality defined as a uniform distribution, (b) 50 variable points, optimality defined as a uniform distribution, (c) 30 variable points, optimality defined as minimisation of the standard deviation from the median of the distance classes, and (d) 30 variable points with optimality defined as uniform distribution over both distance and direction classes.

Chapter 4

Minimising Kriging Variance: Methodology¹

Abstract

This chapter introduces minimisation of the mean kriging variance as a new optimisation criterion in Spatial Simulated Annealing (SSA). Sampling schemes are optimised at the point level. Boundaries and previous observations can be taken into account. We applied it to texture and Phosphate content on a river terrace in Thailand. First, sampling was conducted for estimation of the variogram using ordinary SSA. The variograms thus obtained were used in extended SSA, yielding a reduction of the mean kriging variance of the sand percentage from 28.2 to 23.7 (%)². The variograms were used subsequently in a geomorphologically similar area. Optimised sampling schemes for anisotropic variables differed considerably from those for isotropic ones. The schemes were especially efficient in reducing high values of the kriging variance near boundaries of the area.

¹ Published as: Van Groenigen, J.W., Siderius, W. and Stein, A. (1999). *Constrained optimisation of soil sampling for minimisation of the kriging variance*. *Geoderma* 87: 239-259.

4.1. Introduction

In the practice of spatial (soil) sampling, the sampling scheme is a major factor influencing the efficiency and costs of a survey. In optimising sampling schemes, (geo)statistical theory plays an important role. A sampling scheme might be optimised for interpolation of a Regionalised Variable (ReV) using various criteria. One of the most important geostatistical interpolators is the ordinary kriging algorithm. In this chapter a method will be presented for designing sampling schemes with minimal kriging variance. A regular grid is usually recommended to achieve this aim (e.g. Christakos and Olea, 1992; Yfantis *et al.*, 1987). However, such a grid does not account for all prior information on the ReV and the research area which might be used for optimising the sampling procedure. This prior information can be classified as follows:

- i) *Qualitative maps.* These can be used as a basis for stratification. Stein *et al.* (1988b), Heuvelink and Bierkens (1992) and Boucneau *et al.* (1998) used information provided by soil maps to improve the kriging predictor. Brus (1994) used soil maps for estimation of the mean phosphate characteristics by classical sampling theory.
- ii) *Quantitative maps.* These can be either maps of the ReV in question, or a co-related ReV. Stein *et al.* (1988a) used cokriging for estimation of the moisture deficit, with the mean highest groundwater level as covariable. More recently, Csillag *et al.* (1996) used a co-related remotesensing image as a basis for stratification of the sampling scheme. Van Groenigen *et al.* (1997) (Chapter 2) used probability maps of the ReV exceeding a threshold-value to direct subsequent sampling to uncertain areas.
- iii) *Earlier point observations.* Domburg *et al.* (1997) proposed a method for optimising sampling using classical sampling theory. They found that a variogram calculated from earlier point observations could be used as information for optimising the sampling scheme. McBratney *et al.* (1981) used variograms that can be derived from previous data to calculate optimal sampling density, given a certain required accuracy. More recently, Odeh *et al.* (1990) combined this technique with a fuzzy k-means analysis. They used digital gradient segmentation of the fuzzy membership values to vary sampling density within the research area. Van Groenigen and Stein (1998) (Chapter 3) proposed an optimisation method that could handle earlier data points in the optimisation process.
- iv) *Sampling barriers.* In almost all soil survey studies, sampling barriers are met. In soil surveys on a regional level, these can be inaccessible parts of areas with a poor infrastructure. On a different scale, in soil pollution studies these can be buildings which cannot be sampled. Van Groenigen and Stein (1998) (Chapter 3) proposed the SSA method for constraining an optimised sampling scheme to these barriers.

In this chapter, SSA is extended to deal with minimisation of the kriging variance

of a ReV. This extension includes use of previous samples to direct additional sampling for minimal kriging variance.

The proposed method is applied in a case study on a river terrace in Thailand. The aim of the case study was firstly to investigate efficiently the spatial variability in the region. Secondly, experimental variograms were used as *a priori* variograms in an area which was geomorphologically similar to the first area.

4.2. Materials and methods

4.2.1. Spatial Simulated Annealing

Spatial Simulated Annealing (SSA) is a continuous version of the discrete Simulated Annealing (SA) optimisation method (e.g. Aarts and Korst, 1989). The insensitivity of the algorithm to local extremes makes it very suitable for constrained optimisation of spatial sampling schemes in the presence of complex pre-information. Below, a concise description of SSA is given. For a more extensive introduction and discussion of the method, see Van Groenigen and Stein (1998) (Chapter 3).

Consider the collection of possible sampling schemes consisting of n observations, S^n , with a so called fitness function $\phi(\cdot): S^n \rightarrow \mathcal{R}^+$ to be minimised. Optimisation starts with a random scheme $S_0 \in S^n$. It then involves a sequence of random perturbations S_{i+1} of S_0 that have a probability $P_c(S_i \rightarrow S_{i+1})$ of being accepted. This transition probability is defined in the Metropolis criterion:

$$\begin{aligned} P_c(S_i \rightarrow S_{i+1}) &= 1, & \text{if } \phi(S_{i+1}) \leq \phi(S_i) \\ P_c(S_i \rightarrow S_{i+1}) &= \exp\left(\frac{\phi(S_i) - \phi(S_{i+1})}{c}\right), & \text{if } \phi(S_{i+1}) > \phi(S_i) \end{aligned} \quad (4.1)$$

where c denotes the positive control parameter, which decreases as optimisation progresses. If S_{i+1} is accepted, it serves as a starting point for a next scheme S_{i+2} and the process continues in a similar way (Aarts and Korst, 1989). In SSA, random perturbations of the sampling scheme S_i consist of transformations of randomly drawn observations over a vector with random length and a random direction. We use a vector $\bar{\delta}_i^n$ of n elements. At each step an element of this vector is drawn at random, and is assigned a random value. All other values are equal to 0. This vector has the property that $|\bar{\delta}_i^n| \rightarrow 0$ when $i \rightarrow \infty$. SSA can include earlier observations into the optimisation by treating them as an integral but fixed part of the sampling scheme, *i.e.* with corresponding $\bar{\delta}_i^n$ values set equal to 0, for all i . Boundaries of the region and inaccessible subregions can be taken from a GIS file.

So far, two optimisation criteria have been translated into a fitness function and applied in SSA (Van Groenigen and Stein, 1998) (Chapter 3):

- i) *MMSD-criterion*: this criterion aims at even spreading of the observations over the entire research area by minimising the distance between an arbitrarily chosen point and its nearest observation (van Groenigen and Stein, 1998) (Chapter 3).
- ii) *WM-criterion*: this criterion, which is taken from the literature, optimises the fit of the realised distribution of point pairs for the experimental variogram to a pre-defined, ideal distribution (Warrick and Myers, 1987; Russo, 1984; Russo and Jury, 1988). The desired distribution can be based upon expert judgement, allowing the user to give special attention to certain aspects of the variogram (e.g. short distance observations for estimation of the nugget). The minimisation function is a simple sum of squares of the deviation between the desired distribution ζ^* and the realised distribution ζ^S :

$$\phi_{WM}(S) = \sum_{i=1}^{nc} (\zeta_i^* - \zeta_i^S)^2 \quad (4.2)$$

where nc denotes the number of lag classes, and ζ_i^S is the number of realised point pairs in lag class i (Warrick and Myers, 1987).

4.2.2. Fitness function for minimising kriging variance

Ordinary kriging is a widely used method for spatial interpolation. For this study it is relevant that the ordinary kriging predictor for an arbitrary point \bar{x}_0 is a linear combination of the measured ReV $z(\bar{x}_i)$ in a neighbourhood of n locations, with weights λ_i :

$$\hat{z}(\bar{x}_0) = \sum_{i=1}^n \lambda_i \cdot z(\bar{x}_i). \quad (4.3)$$

These weights are calculated using the variogram assuming a constant expectation, *i.e.* $E\{Z(\bar{x})\} = m$. The variance of the ordinary kriging predictor, $\sigma_{OK}^2(\bar{x}_0 | S_i)$, depends only on the sampling scheme S_i , consisting of observation points \bar{x}_i , the variogram and the kriging neighbourhood (Cressie, 1991). In the past, this has been used to calculate the optimal grid spacing, given a prescribed accuracy (McBratney *et al.*, 1981; McBratney and Webster, 1981). Sacks and Schiller (1988) used SA-related algorithms to optimise sampling schemes on a small grid of possible sampling locations. However, these optimisation procedures could not deal with previous measurements and sampling barriers. The SSA method presented here allows for designing an optimal

sampling scheme with minimal kriging variance as a criterion. We defined minimisation of the integral of the kriging-variance over the study area as the aim:

$$\min_S \int_A \sigma_{\text{MEAN_OK}}^2(\bar{x}_0 | S_i) \cdot d\bar{x}_0. \quad (4.4)$$

In most studies the kriging predictions are calculated on the nodes of a raster. Hence, SSA discretises the area for evaluation of Equation 4.4. The fitness function can then be defined as the mean of kriging variances calculated at the nodes of a fine raster:

$$\phi_{\text{MEAN_OK}}(S_i) = \frac{1}{n_e} \sum_{j=1}^{n_e} \sigma_{\text{OK}}^2(\bar{x}_{e,j} | S_i) \quad (4.5)$$

with $\bar{x}_{e,j}$ denoting the j^{th} raster node, and n_e the number of evaluated raster nodes. This optimisation criterion has also been used elsewhere (Sacks and Schiller, 1988; Christakos and Olea, 1992).

4.2.3. The optimisation procedure

The subsequent steps of the optimisation algorithm can be described as follows:

- i) a variogram $\gamma(\mathbf{h})$ for the ReV is estimated using previous observations or inference from previous studies.
- ii) a sampling scheme $S_0 \in S^n$ is defined, consisting of a subscheme $S_0^f \in S^{n_f}$ with a set of n_f earlier, fixed observations, and a randomly drawn subscheme $S_0^v \in S^{n_v}$ that consists of the n_v observations to be optimised.
- iii) the area is discretised, and the raster is defined with n_e raster nodes.
- iv) kriging variances for all raster nodes are calculated, and the mean kriging variance $\phi_{\text{MEAN_OK}}(S_0)$ is calculated using Equation 4.5.
- v) S_0 is transformed along a vector δ_i^n , yielding sampling scheme S_1 and the new mean kriging variance $\phi_{\text{MEAN_OK}}(S_1)$ is calculated. Notice that all elements of $\bar{\delta}_i^n$ are 0 for $\mathbf{n} = 1, \dots, n_f$, and only one element of $\bar{\delta}_i^n$ for $\mathbf{n} = n_f + 1, \dots, n_f + n_v$ has a value other than 0. The direction of this transformation vector is drawn randomly, and the length is drawn randomly within the interval $[0, \mathbf{h}_{\text{max}}]$. If transformation over $\bar{\delta}_i^n$ results in one of the sampling points falling outside the area, an alternative vector $\bar{\delta}_i^n$ is drawn, until an acceptable vector is obtained.
- vi) S_1 is accepted as a basis for further optimisation depending on the Metropolis criterion (Equation 4.1).
- vii) the process proceeds at point v), with S_1 replacing S_0 if it was accepted. The process ends if during a prespecified time no new sampling schemes have been

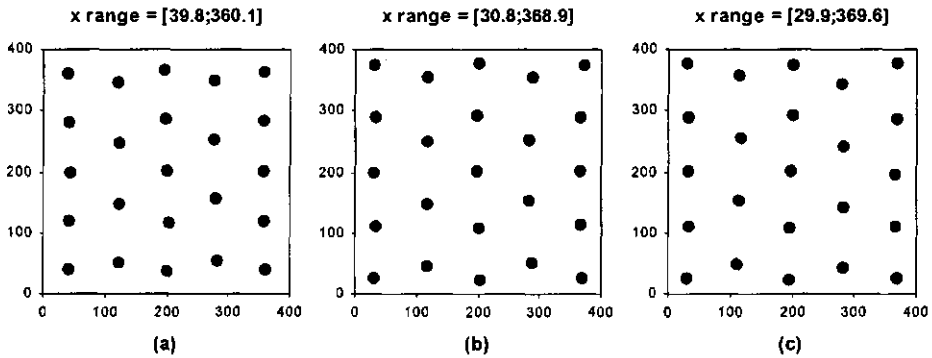


Figure 4.1. Optimised sampling schemes and the x-range of sampling locations for a kriging neighbourhood of 1 (a), 6 (b) and 12 (c), with a linear variogram without nugget.

accepted.

During optimisation, a record of kriging neighbourhoods of all the raster points is kept. The kriging variance at a raster point is recalculated only if the translated sampling point was or will be in the kriging neighbourhood. This saves valuable calculation time.

Optimisation of the total sampling scheme S , consisting of a fixed part S_i^f and a variable part S_i^v , with

$$\bar{\delta}_i^n = \begin{pmatrix} \bar{\delta}_i^{n_i} \\ \bar{\delta}_i^{n_v} \end{pmatrix} = \begin{pmatrix} 0 \\ \bar{\delta}_i^{n_v} \end{pmatrix} \quad (4.6)$$

ensures incorporation of the earlier observations in the optimised sampling scheme. The c value of the Metropolis criterion (Equation 4.1) is lowered as optimisation continues. This is done using an Equation suggested by Aarts en Korst (1989):

$$c_{k+1} = \alpha \cdot c_k, \quad k = 1, 2, \dots, \quad (4.7)$$

where α is a user-chosen parameter with $0 < \alpha < 1$, typically chosen to be close to 1 (e.g. 0.999), and k the number of optimisation steps that have been performed. The period selected for each optimisation step k can be specified by the user.

The length of the random transformation vector $\bar{\delta}_i^n$ is drawn randomly between 0 and \bar{h}_{\max} . Initially, \bar{h}_{\max} is chosen half of the area size, and is lowered using a similar equation as the c variable (Equation 4.7) (Van Groenigen and Stein, 1998) (Chapter 3).

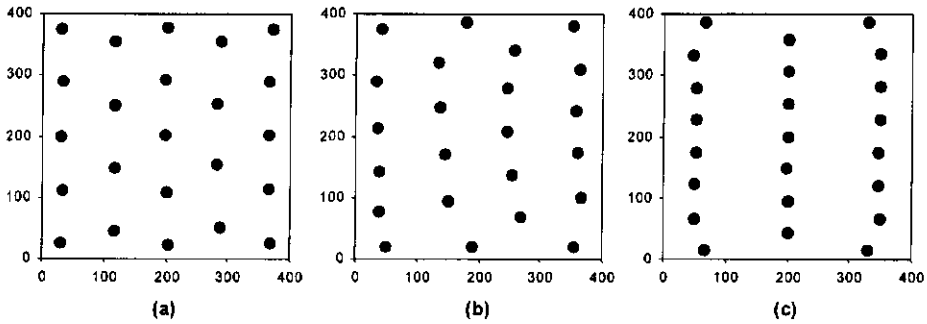


Figure 4.2. *Optimised sampling schemes for anisotropy ratio 1.0 (a), 0.5 (b) and 0.25 (c), with a linear variogram without nugget.*

4.2.4. Examples

Figure 4.1 presents examples of optimised sampling schemes, showing the effect of the kriging neighbourhood. A simple square area was chosen, with a small sampling scheme of 23 sampling points. Figure 4.1a shows the optimised sampling scheme for a kriging neighbourhood of 1 and a linear variogram without a nugget (which is similar to the nearest neighbour method). Figures 4.1b and 4.1c show the optimised sampling schemes for a kriging neighbourhood of 6 and 12, respectively. Although the point configurations of Figures 4.1a, b, and c are the same, the points are located closer to the boundaries when the kriging neighbourhood increases. A kriging neighbourhood of 1 (Figure 4.1a) results in minimum and maximum x-co-ordinates of 39.8 and 360.1, respectively, while a kriging neighbourhood of 12 has a range of 29.9 to 369.6. The small differences between the kriging neighbourhood of 6 and 12 indicate that the kriging neighbourhood in larger optimisations can be kept quite low, if calculation times are expected to be problems.

Figure 4.2 shows the effect of anisotropy on optimised sampling schemes. Figure 4.2a shows the optimised sampling scheme for an isotropic linear variogram with no nugget. Figures 4.2b and c show the optimised sampling schemes for anisotropy ratios of 0.5 and 0.25, respectively. It is clear from these figures that anisotropy has a considerable influence on the optimised sampling schemes. Table 4.1 compares the performance of the optimised sampling schemes of Figure 4.2 with a traditional, triangular equilateral grid. Although the optimised sampling scheme performs slightly better with the isotropic variogram (Figure 4.2a), it is clear that most benefit is gained if the anisotropy is more pronounced. The optimised sampling scheme with anisotropy ratio of 0.25 (Figure 4.2c) has a $\phi_{\text{MEAN_OK}}$ of 81.83 [unit]², while the equilateral triangular grid has a $\phi_{\text{MEAN_OK}}$ of 117.45 [unit]².

Table 4.1. Mean kriging variances for optimised sampling schemes of figure 2 compared to an equilateral triangular grid for different anisotropy ratios. n is 23 observations.

Figure	Anis. ratio	$\phi_{\text{MEAN_OK}}$	$\phi_{\text{MEAN_OK}}$
		optimised scheme	triangular grid
4.2a	1.0	39.99	40.62
4.2b	0.5	56.92	63.16
4.2c	0.25	81.83	117.45

4.3. Case study

4.3.1. Description of the study area

The study area is located in the alluvial plain of the Ping river in northern Thailand, about 60 km north of the city of Chiang Mai. The climate is subhumid, with a rainy season in the period June - October. A high river terrace and some surroundings were chosen for intensive sampling. Figure 4.3 shows a photo interpretation of the terrace and its immediate surroundings. The study area is clearly delineated to the west by the levee, and to the east by the irrigation channel. The size of the area is approximately 1 km x 3 km.

Table 4.2 relates the photo interpretation to drainage class. The river sediments are coarsest and best drained on the levee in the western part of the area. They become finer in the overflow mantle (coarse-fine loamy), via the overflow basin I (fine loamy) and overflow basin II (fine loamy - clayey) to the decantation basin (clayey). To the east, the decantation basin is delineated by the footslope of the neighbouring hills, which consists of coarser material. Furthermore, in the northern part of the area, an old stream channel is visible.

The area shows a clear anisotropy; the variability in landforms is highest in the direction perpendicular to the river. This is in line with what could be expected from geomorphological theory.

The main soil type in the area is classified as Ustic Epiaquert (Soil Survey Staff, 1996) or Eutric Vertisol (FAO, 1994). Most areas on the footslope lack the diagnostic features for Vertisol classification. These soils are classified as Vertic Epiaqualfs (Soil Survey Staff, 1996). Table 4.3 shows some features of a typical Ustic Epiaquert mini-profile in the area. Notable are the heavy texture and the slickensides, which are diagnostic features for Vertisols.

Ping River

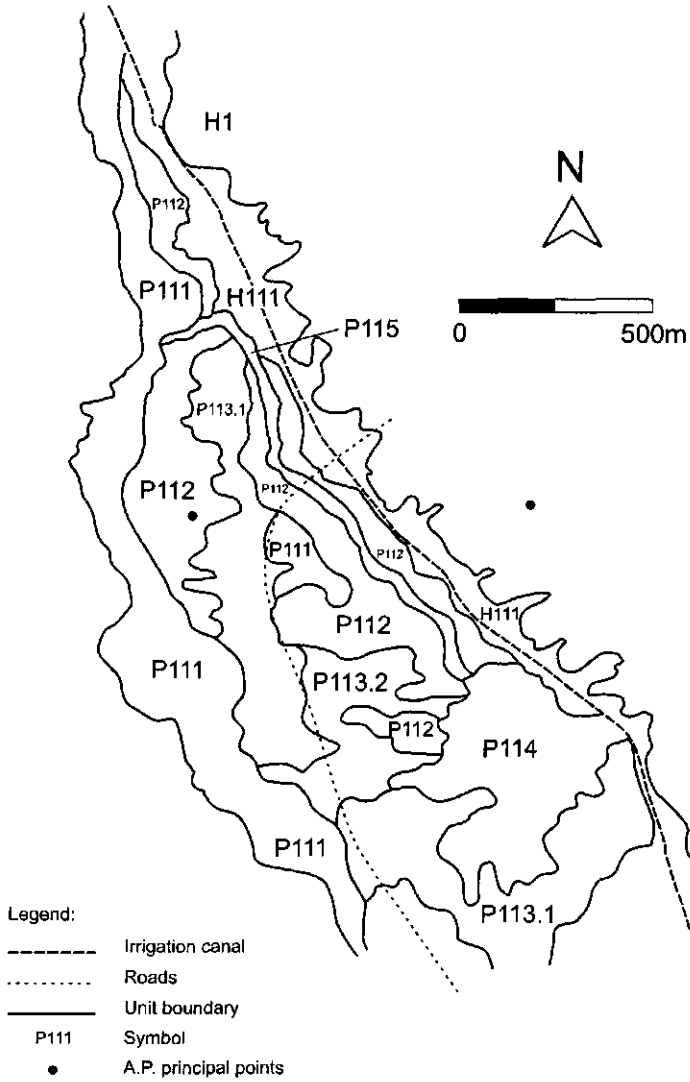


Figure 4.3. Photo interpretation of the landforms within the study area.

Table 4.2. Symbols of the photo interpretation of the landforms within the research area.

<i>Landscape</i>	<i>Relief</i>	<i>Parent Material</i>	<i>Landform</i>	<i>Drainage</i>
<i>Hilland (H)</i>	<i>Foothills (H1)</i>	<i>Metamorphic (H11)</i>	<i>Footslope (H111)</i>	<i>well drained</i>
<i>Alluvial Plain (P)</i>	<i>High Terrace (P1)</i>	<i>Fine loamy to clayey sediments (P11)</i>	<i>Levee (P111)</i>	<i>well drained</i>
			<i>Overflow Mantle (P112)</i>	<i>mod. well drained</i>
			<i>Overflow Basin (P113.1)</i>	<i>imperfectly drained</i>
			<i>Overflow Basin (P113.2)</i>	<i>poorly drained</i>
			<i>Decantation Basin (P114)</i>	<i>very poorly drained</i>
			<i>Old stream channel (P115)</i>	<i>poorly drained</i>

4.3.2. Sampling

Due to financial constraints, the total number of samples was restricted to 60. Each sample was analysed for texture, P-Olsen, N-Kjel., cations, pH and CEC (Hesse, 1971). Samples were taken only from the upper 0.2 m, because of partial flooding of the area for paddy rice.

Because of the limited number of samples, a careful trade-off had to be made between two, partly conflicting, demands:

- i) *Coverage of the area* is necessary, in order to capture the main features of the spatial distribution.
- ii) *Precision of the experimental variogram* partly depends on the number of point pairs per lag class. Typically, a sampling scheme that covers the area evenly has very few to zero point pairs at short distances, thus yielding a poor experimental variogram for those distances.

Because of these conflicting interests, a combination of two criteria was used with SSA. The scheme was designed in two steps:

Table 4.3. Some diagnostic features of a typical Ustic Epiaquet profile in the study area.

<i>Hor. Symb.</i>	<i>Depth (cm)</i>	<i>Diagn. Hor.</i>	<i>Org. M. (%)</i>	<i>pH (H₂O)</i>	<i>Sand (%)</i>	<i>Silt (%)</i>	<i>Clay (%)</i>	<i>Remarks</i>
<i>A_p</i>	0-15	<i>ochric</i>	1.72	6.0	0.9	40.2	59.0	
<i>AB</i>	15-21	-	-	-	-	-	-	
<i>B_{tg}</i>	21-33	<i>(argillic)</i>	0.91	7.9	6.0	38.0	56.0	<i>slickensides</i>
<i>C_{g1}</i>	33-53	-	0.79	7.9	3.8	40.3	56.0	<i>slickensides</i>
<i>C_{g2}</i>	53-	-	-	-	-	-	-	<i>slickensides</i>



Figure 4.4. *The optimised sampling scheme for estimation of spatial variability.*

- i) The first 30 observations were selected according to the MMSD-criterion (Van Groenigen and Stein, 1998) (Chapter 3). In this way, a good coverage of the area with these observations was ensured.
- ii) Subsequently, the additional 30 observations were selected using the WM-criterion of Equation 4.2. The scheme was designed for the estimation of the experimental variogram using the *total number of 60 points*. Special attention was paid to short distances.

Figure 4.4 shows the resulting sampling scheme. While the first 30 points have been distributed evenly over the area, the last 30 observations are clustered within the centre of the area. This is because of the high number of point pairs at short distances that was demanded. Table 4.4 shows the ideal distribution of point pairs ζ^* together with the realised point pair distribution ζ^s . The algorithm manages to fill the gaps in the point pair distribution that were left by the first 30 observations, and create a reasonably equal distribution of point pairs.

Table 4.4. Point pair distribution of the first 30 sampling points, and of the final 60 sampling points, together with the ideal distribution ζ^* .

Distance class (m)	ζ^*	ζ^S	
		first 30 points	60 points
0 - 25	147.5	0	101
25 - 50	147.5	0	95
50 - 100	147.5	0	101
100 - 200	147.5	0	104
200 - 300	147.5	18	98
300 - 400	147.5	40	111
400 - 500	147.5	9	118
500 - 600	147.5	37	131
600 - 700	147.5	36	116
700 - 800	147.5	19	80
800 - 900	147.5	34	121
900 - 1000	147.5	27	139
> 1000	0.0	215	455
Total	1770.0	435	1770

4.3.3. Spatial variability

Table 4.5 shows the descriptive statistics of the analysed variables. To avoid bias by clustering, only the (non clustered) first 30 observations are used. The variability of texture is shown in the texture-triangle of Figure 4.5. There is considerable variability in texture, ranging from very heavy clay with 1% sand, up to loam with almost 50% sand. The loam occurs mainly on the footslope. The most heavy clays are found in the decantation basin.

Figure 4.6a shows the all-directional experimental variogram and the modelled variogram of Nitrogen content. The experimental variogram can be modelled by a linear model with a nugget of $0.0001 (\%)^2$. There is no range within the maximum calculated lag distance of 1000 m. The variogram of Phosphate content (Figure 4.6b) shows a very different type of spatial correlation. The large number of point-pairs at short distances allows an accurate estimation of the nugget, which seems to be extremely low. This variogram can be modelled by an isotropic exponential variogram with a nugget of $18.0 (\text{mg kg}^{-1})^2$, a sill of $567.0 (\text{mg kg}^{-1})^2$ and an effective range of 247 m. The spatial structure of the sand content shows a clear axis of anisotropy, with the longest range in the NW/SE direction, which coincides with the direction of the Ping river, the source of the sediments. This is shown in the variogram surface of the sand percentage (Figure 4.7).

Table 4.5. Descriptive statistics of the non clustered first 30 sampling points

Variable	units	mean	st. dev.	min.	max.
Sand	%	9.8	13.8	0.7	47.5
Silt	%	41.1	8.3	20.3	59.0
Clay	%	49.0	14.6	17.2	77.1
P-Olsen	mg kg ⁻¹	15.9	30.1	1.7	159.1
Org. C	%	1.4	0.4	0.5	2.3
Kjel-N	%	0.1	0.0	0.1	0.2
pH (H ₂ O)	-	6.3	0.6	4.9	7.2
pH (KCl)	-	5.2	0.6	4.1	6.1
CEC	cmol (+) kg ⁻¹	18.9	4.5	9.7	25.5
Ca	cmol (+) kg ⁻¹	14.0	3.7	6.7	20.4
Mg	cmol (+) kg ⁻¹	3.1	0.9	1.3	4.5
Na	cmol (+) kg ⁻¹	0.1	0.1	0.0	0.5
K	cmol (+) kg ⁻¹	0.4	0.3	0.1	1.5
Base-sat.	%	91.4	8.5	74.0	100.0

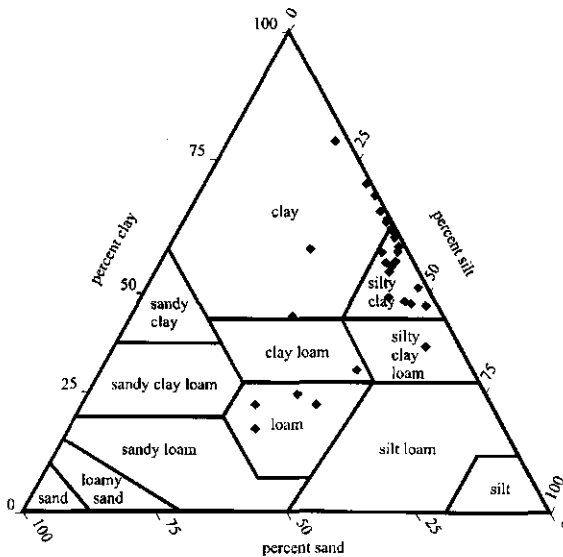


Figure 4.5. Variability of texture, using the non-clustered data.

4.3.4. Additional schemes for minimising the kriging variance

From Figures 4.6 and 4.7, it appears that spatial variability changes considerably from variable to variable. When the number of observations is judged inadequate for spatial interpolation of a variable, additional observations can be selected for minimal

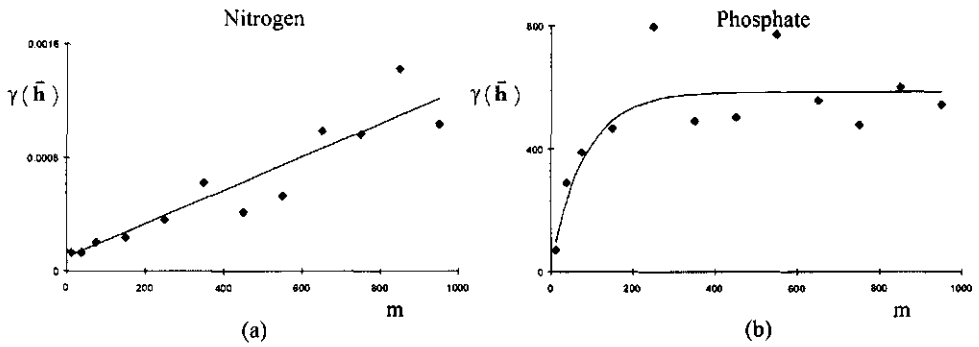


Figure 4.6. Variogram of N-Kjel. (a) and P-Olsen (b).

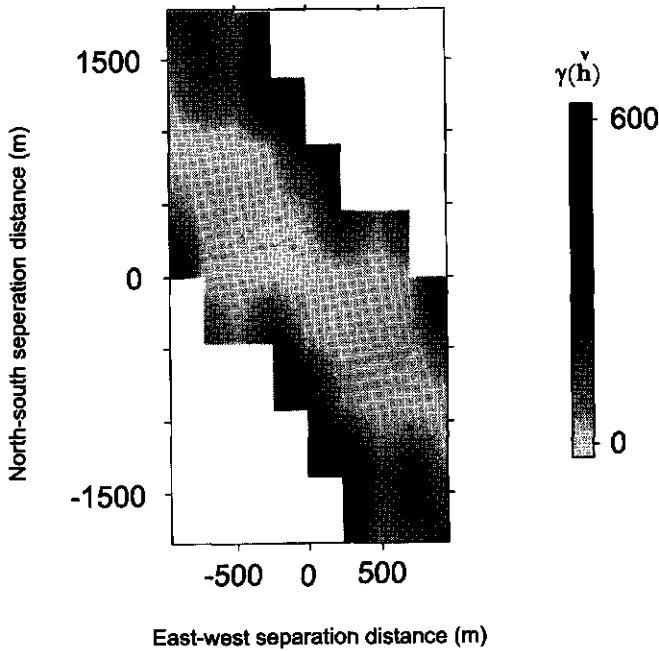


Figure 4.7. Variogram surface of sand percentage.

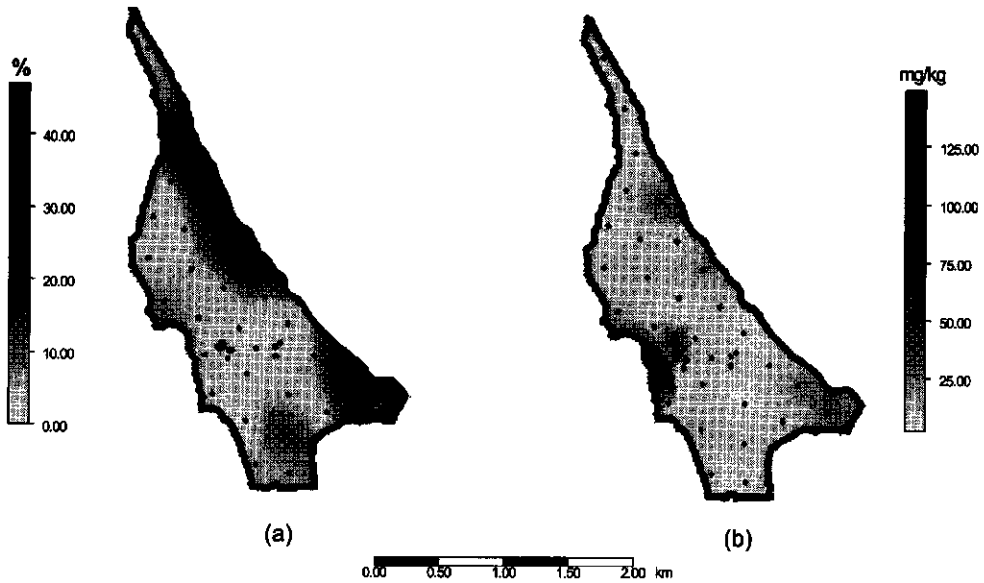


Figure 4.8. Kriged map of sand percentage (a) and P-Olsen (b).

kriging variance, using SSA. These additional sampling schemes have been calculated for two different soil variables.

Figures 4.8a and 4.9a show the kriged map of the sand percentage and its kriging variance, respectively. These maps were calculated with the OKB2D-routine of GSLIB (Deutsch and Journel, 1992). An anisotropic variogram model was used. The kriging predictions are very coarse, as there were only 60 observations, with 30 of them strongly clustered. The kriging variance map, however, gives good insight in the limitations of the MMSD-criterion (or a grid) for kriging purposes.

Figure 4.9b shows the optimised sampling scheme for minimal kriging variance, if 30 additional observations are added. Several of these observations are very close to the boundaries of the area. Moreover, the effect of anisotropy is clear from the smaller spacing of the observations in the direction of the shortest range. Note the different scales in Figures 4.9a and 4.9b. The maximum kriging variance of the sand percentage decreases from 146 to 69 ($\%^2$).

Figure 4.10 shows the mean kriging variance of the sand percentage during the optimisation process, as a function of the number of perturbations. From this figure, it is clear that the optimisation method is insensitive to local minima, in search of the global minimum. Starting with 30 randomly drawn additional observations, the mean kriging variance changes from 28.2 to 23.7 ($\%^2$).

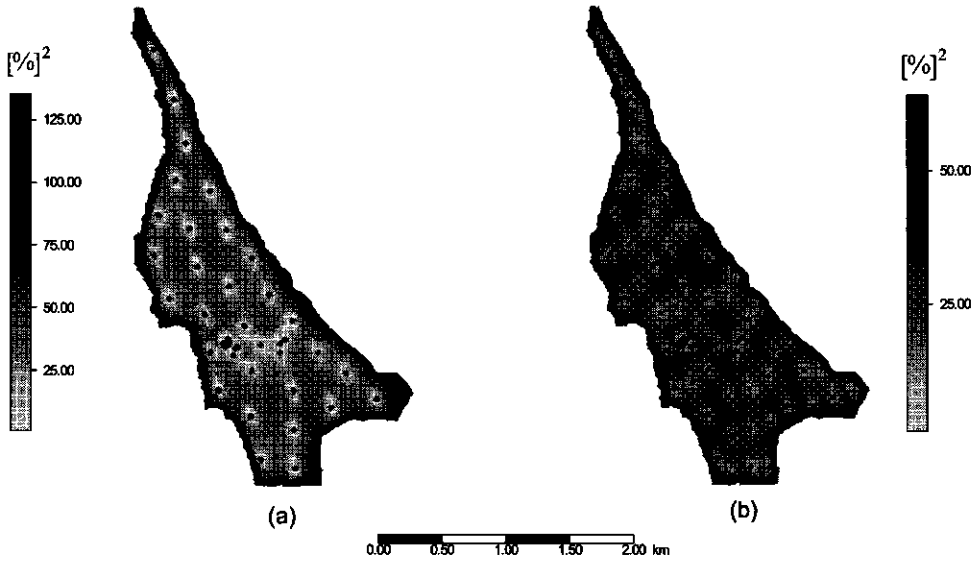


Figure 4.9. Kriging variance of sand percentage with original data (a) and with the additional sampling scheme (b).

Figures 4.8b and 4.11a show the kriged map of the Phosphate content and the kriging variances, respectively. There seems to be a weak correlation between the

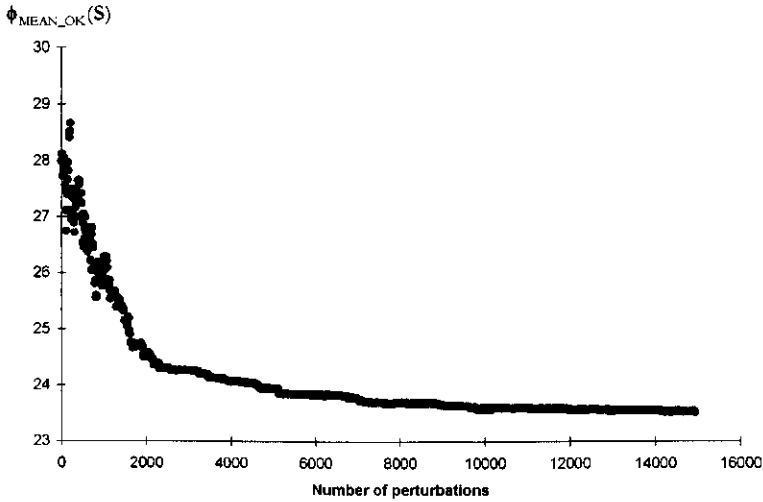


Figure 4.10. The optimisation process for the additional sampling scheme for sand percentage.

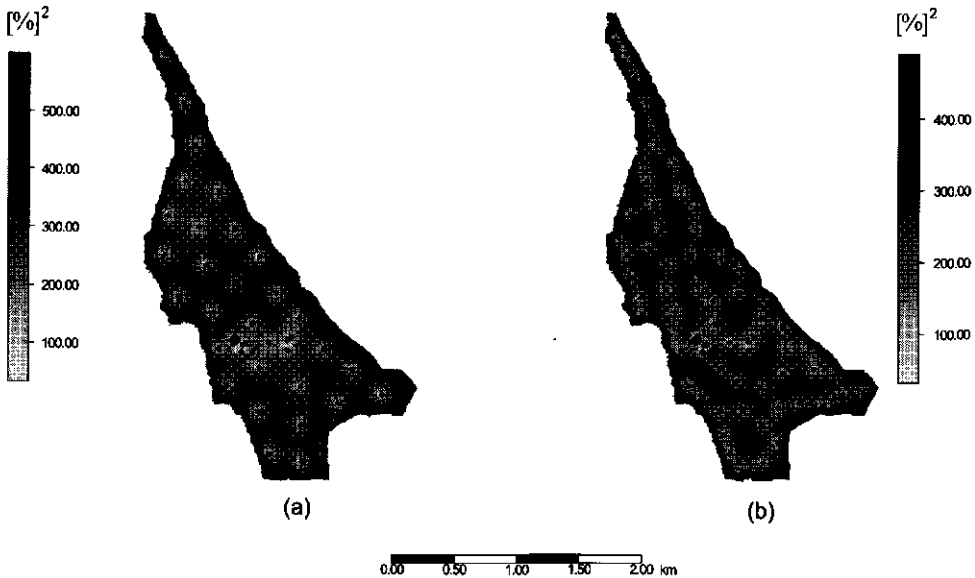


Figure 4.11. Kriging variance of *P*-Olsen with original data (a) and with the additional sampling scheme (b).

sand percentage and the phosphate content on the colluvial soils on the eastern part of the sampling area. This is probably partly due to different management practices (fertilisation) on these lighter soils, where the cash crops (vegetables) are grown. However, the variogram of the Phosphate content is isotropic. This results in a different configuration of additional data points for Phosphate (Figure 4.11b). Because of isotropy of the fitted variogram, the additional data points have no extra emphasis on a particular direction.

If we compare the detailed photo interpretation of the drainage class with the interpolated result of the sand percentage, it is clear that there is a considerable similarity to the texture distribution. In the kriged map of the sand content, the general trends of the texture distribution are similar (decantation basin, footslope). However, as the number of data points was very limited and partly clustered, only a very general geostatistical interpolation was possible.

4.3.5. Minimisation of kriging variance with an *a priori* variogram

If a reliable *a priori* estimation of the variogram of a variable can be made, the sampling scheme can be optimised directly for minimal kriging variance. In our case, it can be inferred from geomorphological classification. A second study terrace, which is

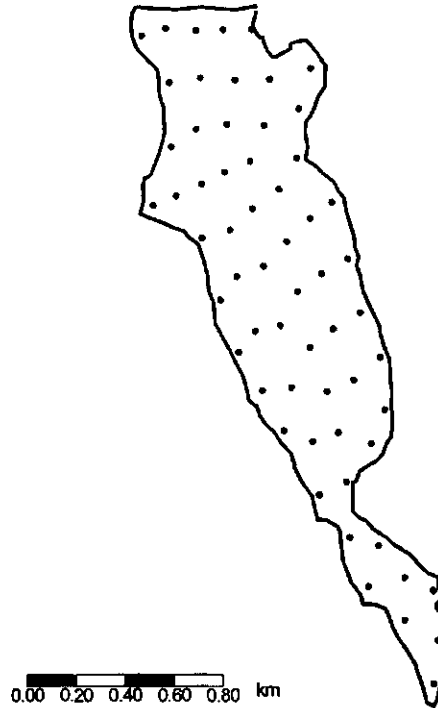


Figure 4.12. *An a priori optimised sampling scheme for anisotropic sand percentage.*

located at the opposite side of the river, has roughly the same soil characteristics as the first. Therefore, the sand percentage variograms of the first terrace were used to optimise a sampling scheme for minimal kriging variance at the second terrace. The axes of anisotropy are rotated clockwise by 10° , as the direction of the second terrace is slightly different from the first terrace.

Figure 4.12 shows this sampling scheme, which consists of 60 points. Again, there is a directional orientation of the sampling scheme, with shorter distances in the direction perpendicular to the river.

4.4. Discussion

There has been repeated criticism in the past, as to the usefulness of the kriging variance as a measure of prediction accuracy (Deutsch and Journel, 1992). Strictly speaking, kriging variance can only be interpreted as prediction accuracy if the intrinsic hypothesis holds. In practice, this assumption is often questionable. Minimising the

kriging variance then reduces to the comparison of different data configurations, without taking into account trends or heteroscedasticity (Journel and Huijbrechts, 1978). In the future, methods that are less sensitive to problems related to stationarity (such as indicator techniques) might be introduced in the SSA procedure as well. However, ordinary kriging is still often used in practice. Deutsch and Journel (1992) state that the ordinary kriging will remain the anchor algorithm in geostatistics. Therefore, we feel that the minimisation of its prediction accuracy is an important issue.

The most widely used nonlinear interpolation technique, indicator kriging (IK), does not yield a similar measure of the accuracy of the interpolation. The aim of indicator kriging is estimation of the conditional distribution function of a ReV. IK uses ordinary kriging of indicator-transformed variables at several cut-off values for this (Deutsch and Journel, 1992). Although IK can yield conditional probability maps of exceeding any threshold value at any location, the accuracy of the interpolation is not accounted for. The use of disjunctive kriging (DK) leads to similar problems when it is used for calculating conditional probabilities (Webster and Oliver, 1989). As this study was specifically aimed at optimising the accuracy of the interpolation, we focussed on ordinary kriging.

The WM-criterion leads in this case study to a strongly clustered sampling scheme. It is well known that for univariate statistics the data of such sampling schemes should be declustered to avoid bias (Isaaks and Srivastava, 1989). The effects of clustering on the experimental variogram are less obvious. Corsten and Stein (1994) found that nested sampling schemes yielded relatively inaccurate estimators of variogram parameters as compared to more conventional sampling schemes. Russo and Jury (1988) found that an optimisation criterion closely related to the WM-criterion yielded more reliable experimental variogram values, especially at short ranges. We feel that strong clustering in one place can only be justified by a very limited number of observations, as in the presented case study. If the budget allows for more observations, more spreading of the points over the area should be aimed at. Strong clustering may lead to a false sense of security because of the high number of point pairs at short distances.

In all types of kriging, the prediction accuracy can only be relative to the chosen variogram model. This means that errors from variogram fitting, choice of variogram model or choice of anisotropy parameters are not accounted for. These issues still need expert judgement, although fuzzy techniques can probably help (Bardossy *et al.*, 1990a and 1990b).

The proposed optimisation method can only be applied if a variogram can be defined. If this is not possible, either by estimation from samples of inference from earlier surveys, other optimisation criteria should be used. The MMSD criterion could be very helpful in such cases.

In anticipating the nature of the anisotropy of certain soil variables (*e.g.* texture), the detailed photo interpretation proved to be a valuable tool during the optimisation

process. Geomorphological classification was used as a tool for choosing an *a priori* variogram model. In our opinion, it is always preferable to model the variogram directly from field data. However, if this is not possible, we feel that geomorphological inference provides at least a sound scientific basis for estimation of the experimental variogram.

4.5. Conclusions

In this chapter, it has been shown how sampling schemes can be optimised by minimising the average kriging variance using SSA. It was shown that kriging neighbourhood has a small but distinct effect on the configuration of the optimised sampling scheme, whereas anisotropy can influence it considerably. Existing observations or available variogram models can be used. Minimisation of the kriging variance is especially useful as both the MMSD-criterion and traditional grids tend to neglect boundary effects which may be important for kriging purposes. The SSA method was applied in a case study in Thailand, where it was shown that the algorithm could handle realistic data sets.

4.6. Acknowledgements

The authors would like to thank Benno Masselink for his assistance with the graphics, and the field staff of the soil science division of ITC for their soil surveys in previous years, which was gratefully used as starting point for the case study. Furthermore, the authors would like to thank Harry-Jan Hendricks-Franssen for his modification of the OKB2D routine from GSLIB, which was used to check the results of the SSA optimisation. The authors are very grateful to Johan Bouma for his careful checking of the manuscript. Finally, we would like to thank the referees D.E. Myers, A. Papritz, J.J. de Gruijter and A.B. McBratney for their useful comments and suggestions.

Chapter 5

Minimising Kriging Variance: Influence of Variogram Parameters¹

Abstract

Using Spatial Simulated Annealing (SSA), spatial sampling schemes can be optimised for minimal kriging variance. In addition to optimising for minimal mean kriging variance, a new criterion for minimising the maximum kriging variance is presented in this chapter. In a simple case with 23 observations, performance of a sampling scheme obtained with SSA were compared with a triangular grid. SSA reduced the mean kriging variance from 40.64 [unit]² to 39.99 [unit]². The maximum kriging variance was reduced from 86.83 [unit]² to 53.36 [unit]². An additional sampling scheme of 10 observations was optimised for an irregularly scattered data set of 100 observations. This reduced the mean kriging variance from 21.62 [unit]² to 15.83 [unit]². The maximum kriging variance was reduced from 70.22 [unit]² to 34.60 [unit]². As the kriging variance depends on variogram parameters, we investigated their influence on the optimised sampling schemes. A Gaussian variogram produced a different sampling scheme as compared to an exponential variogram with the same nugget, sill and (effective) range. Exponential, spherical and linear variograms without nugget resulted in similar sampling schemes. A very short range resulted in random sampling schemes, with observations separated by distances larger than twice the range. For a spherical variogram, magnitude of the relative nugget effect did not affect the sampling schemes, except for very high values (0.75).

¹Based on: Van Groenigen, J.W. (submitted). *The influence of variogram parameters on optimal sampling schemes for kriging*. Geoderma.

5.1. Introduction

Optimisation of spatial sampling is one of the most challenging issues in soil geostatistics. In the past, many research efforts have been aiming at how to process a data set (using interpolation and stochastic simulation), rather than how to collect it. Despite important contributions to the discussion on spatial sampling (e.g. McBratney *et al.*, 1981; Burgess and Webster, 1984; Yfantis *et al.*, 1987; Brus and De Gruijter, 1997), sampling schemes for geostatistical interpolation are still mainly based on some sort of regular grid. Taking into account geographical information, preliminary observations, and information on spatial correlation could potentially improve sampling strategies considerably (Van Groenigen and Stein, 1998) (Chapter 3). For interpolation purposes, such sampling schemes might be designed for minimising the accuracy of the ordinary kriging predictor.

Ordinary kriging (OK) is still the most widely used interpolation algorithm in geostatistics (Deutsch and Journel, 1992). Although more powerful algorithms like indicator kriging (IK) are now available, the relative transparency and straightforwardness of the OK algorithm ensures its continuing popularity. In fact, even IK uses the OK algorithm to interpolate conditional probabilities.

The accuracy of the kriging predictor is usually expressed in the kriging variance. This kriging variance relies solely upon the spatial correlation, the locations and number of observations in the kriging system and the location of the predicted point. The spatial correlation is usually estimated in the variogram, which presents the variance as a function of the separation vector between two points. Therefore, the kriging variance can be calculated before sampling takes place, provided that a sampling scheme and the variogram are available. This makes it possible to optimise a sampling scheme for minimal kriging variance, provided a variogram can be defined.

In the past, work has been done on optimising sampling schemes for minimal kriging variance. Methods were developed for calculating the optimal sampling density of a grid, given a prescribed maximum kriging variance (McBratney *et al.*, 1981; McBratney and Webster, 1981). Furthermore, the performances of different types of regular grids for kriging purposes were investigated (e.g. Yfantis *et al.*, 1987; Christakos and Olea, 1992).

Sacks and Schiller (1988) used a stochastic optimisation algorithm for optimising the sampling scheme for minimal prediction error. They used a very limited, discrete solution space, and a small sample size. More recently, Van Groenigen *et al.* (1999) (Chapter 4) showed how the Spatial Simulated Annealing (SSA) algorithm could be used for constructing sampling schemes with minimal kriging variance, using a continuous solution space and with a more realistic number of observations. They found that anisotropy of the variogram has a considerable influence upon the optimised

sampling scheme, with highest sampling density in the direction of highest variability. Moreover, they showed that the size of the kriging neighbourhood has a very limited, but distinct effect on the sampling schemes, drawing closer to the boundaries as the kriging neighbourhood increases. They illustrated with a case study how SSA could be used for real field data, making use of GIS-stored data.

The aim of this chapter is to investigate the influence of variogram parameters on the optimised sampling scheme. Van Groenigen *et al.* (1999) (Chapter 4) showed that anisotropy and kriging neighbourhood influence the optimised sampling scheme. In this chapter it will be shown that other variogram parameters also influence the optimised sampling scheme. Furthermore, a new optimisation criterion is introduced, *i.e.* minimising the maximum occurring kriging variance rather than the mean kriging variance. Finally, it will be shown how SSA is able to complete existing, inadequate sampling schemes for minimisation of the kriging variance. In this way, considerable improvements in prediction accuracy can be achieved with only modest additional sampling efforts.

5.2. Materials and methods

5.2.1. Variogram and kriging

When the intrinsic hypothesis holds, the isotropic variogram is defined as

$$\gamma(\mathbf{h}) = \frac{1}{2} \mathbf{E} \left[\{Z(\bar{\mathbf{x}}) - Z(\bar{\mathbf{x}} + \mathbf{h})\}^2 \right] \quad (5.1)$$

where $Z(\bar{\mathbf{x}})$ denotes Regionalized Variable (ReV) Z at location $\bar{\mathbf{x}}$. \mathbf{h} Denotes the separation distance between two locations. In most geostatistical studies, the variogram is estimated by a series of observations $z(\bar{\mathbf{x}})$:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2n(\mathbf{h})} \sum_{i=1}^{n(\mathbf{h})} \{z(\bar{\mathbf{x}}_i) - z(\bar{\mathbf{x}}_i + \mathbf{h})\}^2 \quad (5.2)$$

where $n(\mathbf{h})$ denotes the number of point pairs for distance class \mathbf{h} . This estimated variogram is fitted to a suitable model $\gamma(\mathbf{h})$, often using nonlinear regression. Subsequently, $\gamma(\mathbf{h})$ is used to fill the matrices needed for the ordinary kriging system. The ordinary kriging predictor can be written as a weighted average of \mathbf{n} observations:

$$\hat{z}(\bar{x}_0) = \sum_{i=1}^n \lambda_i \cdot z(\bar{x}_i) \quad (5.3)$$

where λ_i denotes the weight of the i^{th} observations. The kriging variance can be written as

$$\sigma_{OK}^2(\bar{x}_0) = \sum_{i=1}^n \lambda_i \gamma(\bar{x}_i - \bar{x}_0) + \Psi \quad (5.4)$$

where Ψ denotes a Lagrange multiplier (Webster and Oliver, 1990). The only factor influencing the kriging variance are the variogram $\gamma(\mathbf{h})$, the number of observations n , the sampling locations \bar{x}_i and the location \bar{x}_0 .

5.2.2. Spatial Simulated Annealing

Spatial Simulated Annealing (SSA) was developed as an optimisation algorithm for spatial sampling schemes. As a modification of simulated annealing (SA), SSA is able to optimise sampling schemes at the point level. Starting with a random sampling scheme, a sequence of random alterations in the locations of observations is drawn. As the process evolves, the maximum length of the random vectors over which the observations are transformed decreases. Simultaneously, the probability that an inferior alteration is accepted decreases. In this way, the sampling scheme is optimised according to the chosen optimisation criterion. SSA is very suitable for completing inadequate, existing data sets, making full use of the existing observations. Although no proof can be given that SSA always results in a global optimal solution, Van Groenigen and Stein (1998) (Chapter 3) showed that the found solutions are superior to traditional sampling strategies. For an extensive presentation of the SSA algorithm, see Van Groenigen and Stein (1998) (Chapter 3). SSA can handle a variety of quantitative optimisation criteria. In this chapter, we will focus on minimisation of the kriging variance, as presented in Van Groenigen *et al.* (1999) (Chapter 4). Additionally, we will propose an alteration of this criterion below.

5.2.3. Criterion 1: minimisation of the mean kriging variance

Following Van Groenigen *et al.* (1999) (Chapter 4), the aim of this optimisation criterion is minimisation of the integral of the ordinary kriging variance over the area of interest. This leads to the following minimisation criterion:

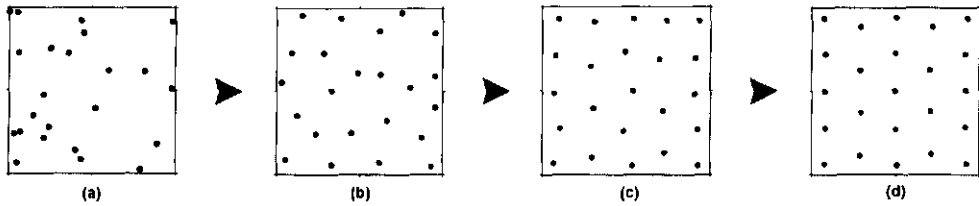


Figure 5.1. The optimisation process. Intermediate solutions for an isotropic linear variogram at start (a), 33% (b), 67% (c) and 100% (d) of the process.

$$\min_S \int_A \sigma_{OK}^2(\bar{x}_0 | S) \cdot d\bar{x}_0 \tag{5.5}$$

where S denotes the sampling scheme, \bar{x}_0 denotes the location vector in the area A , and σ_{OK}^2 is the ordinary kriging variance. As this integral can generally not be solved analytically, a discretisation of the problem is necessary. This leads to a minimisation function defined as the mean kriging variance over a finely meshed grid of evaluation points \bar{x}_e :

$$\phi_{MEAN_OK}(S) = \frac{1}{n_e} \sum_{j=1}^{n_e} \sigma_{OK}^2(\bar{x}_{e,j} | S) \tag{5.6}$$

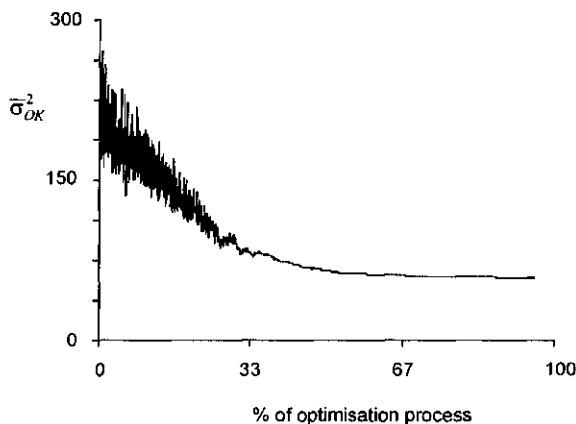


Figure 5.2. The optimisation process. Mean kriging variance as a function of the elapsed time.

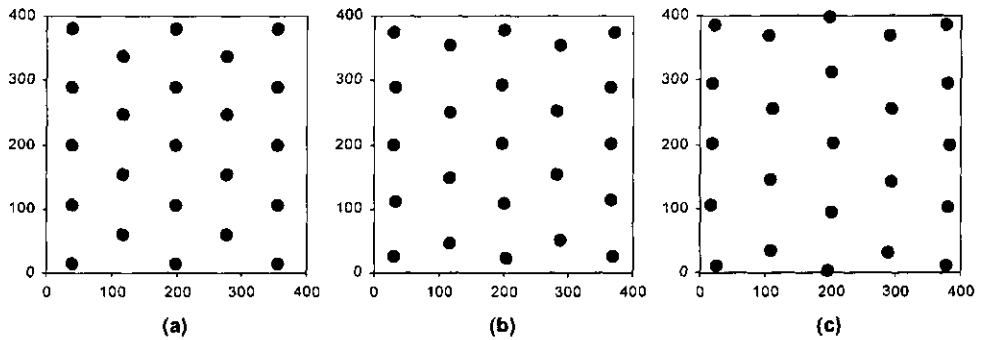


Figure 5.3. A triangular grid (a) and optimised sampling schemes for minimal mean kriging variance (b) and minimal maximum kriging variance (c).

where n_e denotes the number of raster nodes. This will be referred to as the MEAN_OK-criterion.

5.2.4. Criterion 2: minimisation of the maximum kriging variance

As an alternative criterion, we propose a minimax version of the MEAN_OK-criterion. Instead of minimising the mean kriging variance, the maximum value is minimised:

$$\phi_{\text{MAX_OK}}(\mathbf{S}) = \max(\sigma_{\text{OK}}^2(\bar{x}_{e,j} | \mathbf{S}), \quad \forall \bar{x}_{e,1}, \dots, \bar{x}_{e,n_e}) \quad (5.7)$$

This will be referred to as the MAX_OK-criterion. The use of a minimax criterion might be especially useful in studies where the quality of a survey relies upon the weakest chain (*i.e.* the prediction with the lowest accuracy). This might be the case, for example, in environmental studies dealing with health risks. The MAX_OK criterion is similar

Table 5.1. Comparison of an equilateral triangular grid with two optimised sampling schemes using SSA with the MEAN_OK and MAX_OK criterion. n is 23 observations

Fig.	Optimisation	Mean kriging variance ($\phi_{\text{MEAN_OK}}$)	Maximum kriging variance ($\phi_{\text{MAX_OK}}$)
5.3a	triangular grid	40.62	86.83
5.3b	MEAN_OK	39.99	74.61
5.3c	MAX_OK	41.41	53.36

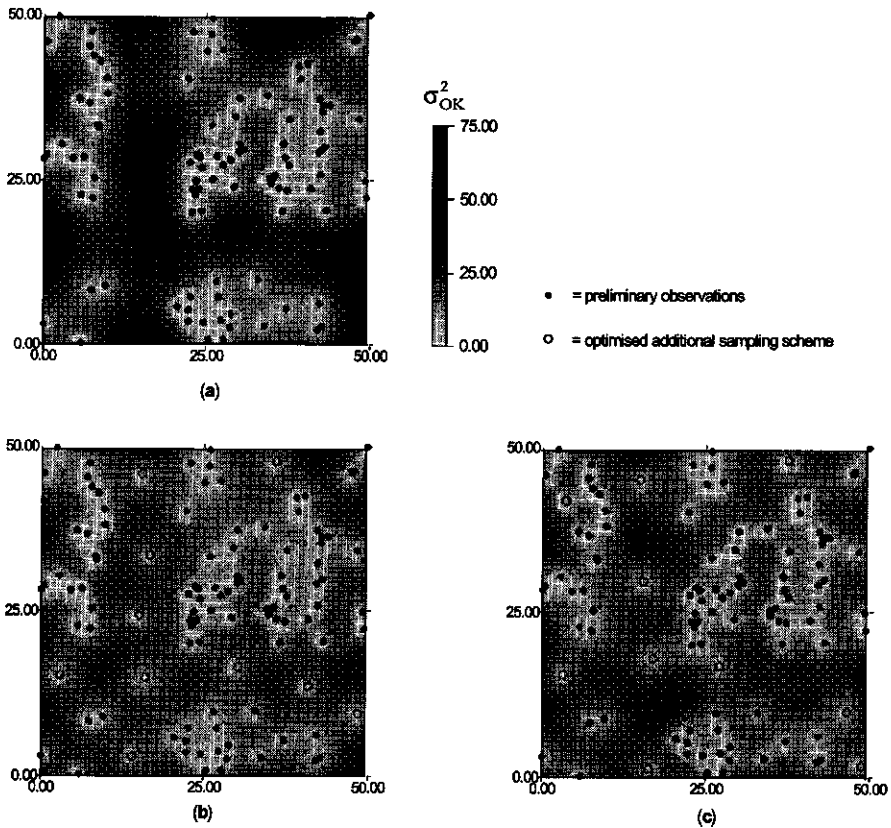


Figure 5.4. A preliminary sampling scheme with kriging variance (a), and the additional optimised sampling scheme of 10 observations using minimal mean kriging variance (b) and minimal maximum kriging variance (c).

to the criterion that was used to calculate optimal grid spacing for a certain minimal accuracy of the kriging predictor, as presented in McBratney *et al.* (1981).

5.3. Examples

Below, examples will be given of the way optimised sampling schemes are influenced by the chosen optimisation criterion and variogram model. Most examples are for a square area with 23 observations. This low number of observations was chosen to

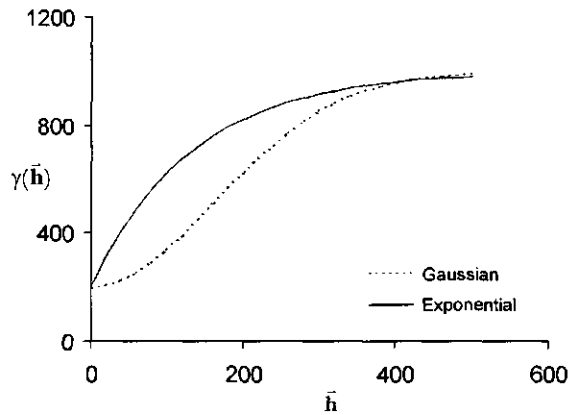


Figure 5.5. A Gaussian and exponential variogram, with nugget 200, sill 1000, and effective range 400.

show the effect on the optimised schemes as clearly as possible.

5.3.1. Influence of the optimisation criterion

Figure 5.1 shows the optimisation process for a linear variogram without a nugget. The process starts with a random sampling scheme (Figure 5.1a). As the process continues, the sampling scheme moves more away from a random configuration, towards even spreading over the area (Figures 5.1b and 5.1c). Finally, optimisation yields a point symmetric configuration (Figure 5.1d). Figure 5.2 shows the mean kriging variance during optimisation. At the start, the kriging variance fluctuates more or less randomly. During the optimisation, the mean kriging variance steadily decreases while avoiding local minima. Finally, the mean kriging variance does not decrease any more.

Figure 5.3a shows an equilateral triangular grid consisting of 23 observations. Figures 5.3b and 5.3c show optimised sampling schemes with a linear variogram without nugget for the MEAN_OK and MAX_OK criteria, respectively. Both these last two figures show the same configuration of points, but with the MAX_OK criterion the

Table 5.2. Kriging variance for a preliminary sampling scheme of 100 observations, and with 10 optimised additional observations for the MEAN_OK and MAX_OK criterion.

Fig.	Optimisation	No. of observations	Mean kriging variance ($\phi_{\text{MEAN_OK}}$)	Maximum kriging variance ($\phi_{\text{MAX_OK}}$)
5.4a	preliminary	100	21.62	70.33
5.4b	MEAN_OK	110	15.83	58.54
5.4c	MAX_OK	110	16.46	34.60

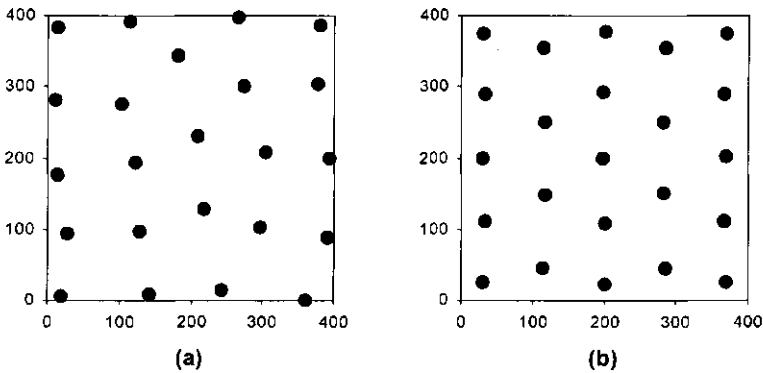


Figure 5.6. Influence of variogram models on the optimised sampling scheme for a Gaussian model (a) and an exponential model (b) with similar nugget, sill and (effective) range.

points are drawn much closer to the boundaries. This can be explained by the usually high kriging variance at the edges of a kriged surface. Table 5.1 shows the scoring of these three schemes, according to both criteria. It can be concluded that the triangular grid is inferior for both criteria. The optimisation effect for the MEAN_OK criterion is only limited (a mean kriging variance of 39.99 [unit]² as compared to 40.62 [unit]² using the grid) in this theoretical example. The optimisation effect of applying the MAX_OK criterion is considerable, reducing the maximum kriging variance from 86.83 [unit]² with the triangular grid to 53.36 [unit]².

It should be observed that when in this chapter, and throughout this thesis, two sampling schemes are called similar, it is meant that they can be rotated and/or reflected to fit each other, depending on the (random) path that the SSA algorithm followed.

Differences between the two criteria are more clearly shown in another example. Figure 5.4a shows an area with 100 observations that are irregularly scattered, and the kriging variance is calculated with a spherical variogram with nugget 0, sill 100 and range 30. SSA was used to add 10 observations to enhance the quality of the data set for kriging purposes. Figures 5.4b and 5.4c show the optimised sampling schemes for the MEAN_OK and MAX_OK criterion, respectively. Table 5.2 shows the performances of the sampling schemes thus obtained, as compared to the preliminary sampling scheme. Using only 10 additional observations, the mean kriging variance could be reduced from 21.62 [unit]² to 15.83 [unit]² using the MEAN_OK criterion. Use of the MAX_OK criterion resulted in a reduction of the maximum kriging variance from 70.33 [unit]² to 34.60 [unit]². The differences between the two criteria are nicely illustrated in the lower right corner of Figures 5.4b and 5.4c. The MEAN_OK criterion leaves this area with a relatively high kriging variance in order to lower the mean kriging variance. The MAX_OK criterion places one observation in this corner, in order to

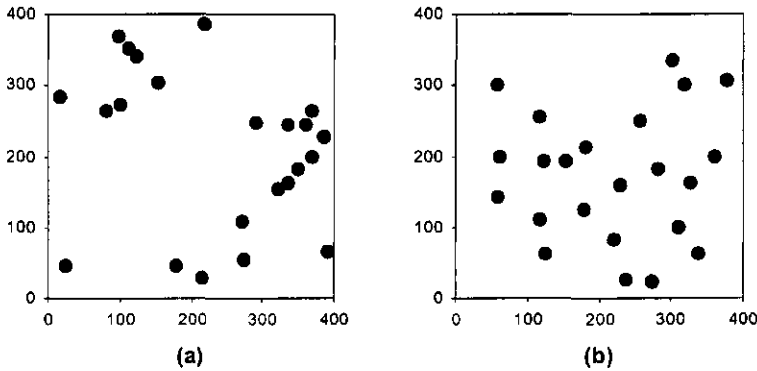


Figure 5.7. Influence of range on the optimised sampling scheme, for a spherical model with a range of 0 (a) and 20 (b).

lower this maximum kriging variance (Figure 5.4c). A surveyor aiming at lowering the kriging variance should carefully consider whether the MEAN_OK or MAX_OK criterion should be used, since they can result in quite different schemes.

5.3.2. Influence of variogram model

Figure 5.5 shows a Gaussian and an exponential variogram with the same parameters. The Gaussian model is:

$$\gamma(\mathbf{h}) = 200 + 800 \cdot \left[1 - \exp\left(-\frac{\mathbf{h}^2}{(230.94)^2}\right) \right] \quad (5.8)$$

This reflects a nugget of 200, a sill of 1000 and an effective range of 400 ($230.94 \cdot \sqrt{3}$). These parameters are also used in the exponential model:

$$\gamma(\mathbf{h}) = 200 + 800 \cdot \left[1 - \exp\left(-\frac{\mathbf{h}}{133.33}\right) \right] \quad (5.9)$$

The effective range ($3 \cdot 133.33$) of this model is similar to that of the Gaussian model. Figure 5.6 shows the optimised sampling schemes for these two variograms, using the MEAN_OK criterion. While the exponential variogram resulted in a point symmetric

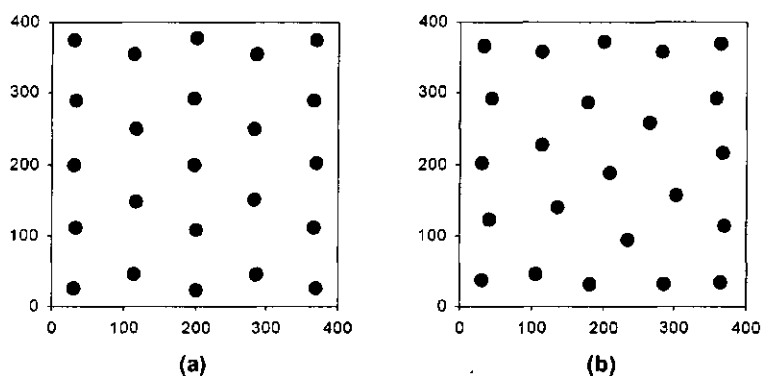


Figure 5.8. Influence of nugget on the optimised sampling scheme, for a spherical model with a nugget of 50 (a) and a nugget of 75 (b). The sill is 100 for both cases.

solution similar to that of the linear variogram, the Gaussian model yielded a different scheme, with observations much closer to the boundaries of the area.

5.3.3. Influence of range

Figure 5.7 shows the influence of the range on optimised sampling schemes for a spherical variogram. Figure 5.7a shows the optimised sampling scheme for a pure nugget effect, *i.e.* the range is 0. This variogram results in a random sampling scheme; each run of the optimisation will give a different result. As a pure nugget effect implies that no spatial correlation is present, each sampling scheme used for interpolation is equally inadequate. Figure 5.7b shows the optimised sampling scheme for a very short range of 20. This implies that no spatial correlation is present at distances larger than 20. The observations are placed at distances larger than 40 (twice the range). Within this constraint, the resulting sampling scheme is still random. For larger ranges, the sampling scheme will be similar to that produced using a linear variogram (Figure 5.3b). A real range of 0 (a pure nugget effect) is very unusual in soil science. More often, a modelled nugget effect stems from sampling at larger distances than the range. Therefore, optimal sampling schemes for a pure nugget effect are mainly of theoretical interest.

5.3.4. Influence of (relative) nugget effect

If the nugget is 0, the sill will have no effect on the sampling scheme. A sill of 200

will simply result in kriging variances twice as high as those using a sill of 100, but the optimised sampling scheme will not change. Therefore, influence of the sill only depends on the nugget/sill ratio. A spherical sampling scheme with nugget 0 and a sufficiently large range results in a point symmetric solution, similar to that derived using a linear variogram (Figure 5.3b). With a nugget/sill ratio of 0.5 (nugget is 50, sill is 100), this does not change (Figure 5.8a). However, when the nugget/sill ratio increases to 0.75 (nugget 75, sill 100), a somewhat different solution is reached (Figure 5.8b).

5.4. Discussion and conclusions

Two different optimisation criteria were formulated in this chapter, both dealing with optimising the quality of the kriging interpolator. Minimising the mean kriging variance results in other sampling schemes than minimising the maximum kriging variance.

Most variogram parameters influence optimal sampling schemes for minimal kriging variance. Therefore, use of the SSA algorithm almost always results in superior sampling schemes in terms of kriging variance. For cases with a relatively low nugget and sufficient range, the optimised sampling schemes for linear, exponential and spherical variograms were similar. Only the Gaussian variogram yielded a different scheme. It can be expected that other variogram models like the hole effect model, will also result in different sampling schemes.

The question remains to what extent the obtained solutions are truly the global optimum. This can not be proven. For simulated annealing, it has been proven that the global solution is always found, given infinite calculation time (Aarts and Korst, 1989). For realistic calculation times, this has not been proven. However, all optimised sampling schemes were reproduced in subsequent runs of the algorithm with different (random) starting points. This indicates that the solutions might indeed be the global optimum. More to the point of this study, it was shown that the sampling schemes performed better than the traditional equilateral grid.

This study considered univariate problems. Other problems might be multivariate. In those cases, several optimisation strategies might apply. Similar to the method presented in this paper, the cokriging variance might be minimised when a model of co-regionalisation can be defined. However, this requires much modelling of variogram and cross-variograms, and can only apply in cases with few variables. Van Groenigen *et al.* (*submitted*) (Chapter 6) and Broos *et al.* (*submitted*) showed how conditional probabilities of exceeding environmental threshold values of several contaminants can be pooled into one variable, indicating health risk.

Use of SSA for minimisation of the kriging variance will result in sampling schemes that explicitly take into account the nature of spatial dependence. The surveyor can

choose between minimising the overall kriging variance and the maximum kriging variance. Moreover, preliminary observations can be taken into account. Therefore, SSA should be able to provide a useful tool for surveyors in the future.

5.5. Acknowledgements

The author is very grateful to Harry-Jan Hendricks-Franssen for his modification of the OKB2D subroutine of GSLIB, which was used to check the results of the optimisations. Furthermore, he would like to thank Alfred Stein for his useful comments on the draft version of this manuscript.

Chapter 6

Optimising Multivariate Interpolation¹

Abstract

Effectiveness of regular sampling grids to collect multivariate contamination data in urban areas is often strongly reduced by buildings and boundary effects. In addition, earlier observations and knowledge on the history of the area may provide valuable information. This chapter presents a simulated annealing-based procedure to optimise the sampling scheme, taking sampling constraints and preliminary information into account. The sampling scheme is optimised using a spatial weight function, that allows to distinguish areas with different priorities. A case study in the Rotterdam harbour with five contaminants at two depths showed two subsequent sampling stages, in which two weight functions were applied. The first stage combined earlier observations and historical knowledge, with emphasis on areas with high priority. The resulting scheme showed a contamination at 17.4% of the samples, with 1.5% heavily contaminated. The second stage used probability maps of exceeding intermediate threshold values to guide additional sampling to possible hot-spots. This yielded 26.7% contaminated samples, with 16.7% being heavily contaminated. These included new locations that were not detected during the first stage. The proposed method allows to incorporate important preliminary information, and can be used as a valuable tool in environmental decision making.

¹ Based on: Van Groenigen, J.W., Pieters, G. and Stein, A. (submitted). *Optimising spatial sampling for multivariate contamination in urban areas*. *Environmetrics*.

6.1. Introduction

Characterisation of multivariate contamination for remediation of urban areas is a highly complicated problem in soil geostatistics. A surveyor has to balance carefully between scientific integrity, legal validity and practical attainability. Traditional tools of a soil surveyor such as photo interpretation and knowledge on soil/landform relationships are of limited use, as urban soils are seldom in their natural state. The spatial distribution of the contaminants is typically poorly related to soil type. Also, accessibility to the sampling sites is often constrained by buildings and roads. Geostatistical methods of handling many Regionalized Variables (ReV's) and their interactions are still complicated, involving much modelling and many assumptions (e.g. Deutsch and Journel, 1992). Finally, observations are usually costly, and the most should be made from limited facilities. Because of all these complications, optimising the sampling scheme is a crucial issue.

A geostatistical survey normally involves first estimation of a model for spatial dependence, usually expressed by the variogram. Secondly, this model can be used for optimal interpolation of an ReV. Past efforts of optimising spatial sampling for interpolation can therefore be split up in two, potentially conflicting strategies:

- i) *Strategies aiming at optimal estimation of variogram parameters.* Webster and Oliver (1992) found that for an isotropic (direction-independent) variogram, at least 150 sampling points on a regular grid are needed for estimation of the variogram, and that 225 usually give reliable estimations. Yfantis *et al.* (1987) found that a triangular grid yielded the most reliable estimation of the variogram, as compared to square and hexagonal grids. Russo (1984) and Warrick and Myers (1987) proposed an optimisation method for reproducing an *a priori* defined ideal distribution of point pairs for estimating the variogram. Van Groenigen and Stein (1998) (Chapter 3) proposed a more powerful algorithm for the same criterion. Russo and Jury (1988) found that this criterion yielded more reliable short range estimates of the variogram than systematic sampling.
- ii) *Strategies aiming at optimal spatial interpolation (often kriging).* Given the variogram, optimal grid spacing can be calculated given a required minimal accuracy (McBratney *et al.*, 1981; McBratney and Webster, 1983). More recently, methods were developed for selecting optimal locations of sampling points for interpolation, rather than optimal grid spacing. First, this was done using a list of possible sampling locations as a rather small discrete solution space (Sacks and Schiller, 1988). Van Groenigen *et al.* (1999) (Chapter 4) showed how this method could be extended to much larger problems with a continuous solution space.

Although these strategies have contributed much to improve the quality of sampling schemes, they were all essentially univariate. As most surveys aim at characterisation of more than one ReV, there is a clear need for multivariate sampling strategies.

Urban contamination studies typically offer much preliminary information. Often the historic use of the area is known, indicating possibly contaminated sites. As environmental legislation involves the use of different threshold values and requires different survey phases, preliminary data might be available on the area. Using these constraints and preliminary information on multivariate contaminants to optimise a sampling scheme is one of the main challenges in environmental soil surveys.

In this chapter, we will introduce a method for optimising spatial sampling in urban areas. We will show how we can make use of the preliminary information on the area, in the form of both historic information and preliminary observations. The optimisation method is an extension of the SSA algorithm presented in Van Groenigen and Stein (1998) (Chapter 3).

6.2. Theory and methods

6.2.1. Spatial Simulated Annealing

Spatial Simulated Annealing (SSA) is an optimisation method for spatial sampling schemes. SSA is especially beneficial in studies with many sampling constraints and preliminary observations, such as urban remediation studies (Van Groenigen and Stein, 1998) (Chapter 3). The algorithm considers constrained optimisation of a sampling scheme as a combinatorial optimisation problem. Consider a 2-dimensional region A and let the collection of all possible sampling schemes with n observations on A be denoted S^n . Then we define a fitness function $\phi(S) : S^n \rightarrow \mathfrak{R}^+$ that has to be minimised to optimise the sampling scheme. The fitness function can be any quantitative function of a sampling scheme, and does not necessarily have to be continuous. Optimisation starts with a randomly selected sampling scheme $S_0 \in S^n$, consisting of observation points $\bar{x}_0^1, \dots, \bar{x}_0^n$. At the first optimisation step, the observation points are transformed over a random vector $\bar{\delta}_0^n$ with only one element different from zero, yielding a new sampling scheme S_1 . This alternative sampling scheme is accepted with probability P_c , using the Metropolis criterion:

$$\begin{aligned}
 P_c(S_i \rightarrow S_{i+1}) &= 1, & \text{if } \phi(S_{i+1}) \leq \phi(S_i) \\
 P_c(S_i \rightarrow S_{i+1}) &= \exp\left(\frac{\phi(S_i) - \phi(S_{i+1})}{c}\right), & \text{if } \phi(S_{i+1}) > \phi(S_i).
 \end{aligned} \tag{6.1}$$

If S_1 has a lower fitness (*i.e.* if it is superior according to our criterion) than S_0 , it is accepted as a basis for further optimisation. If it has a higher fitness, it is accepted with a probability that is positively related to parameter c of Equation 6.1. The process then continues with the creation of a sampling scheme S_2 out of S_1 (if it was accepted) or S_0 . As the optimisation process evolves, c and the maximum length of $\bar{\delta}_1^n$ decrease, forcing the sampling scheme to 'freeze' in its optimal configuration, in a similar way as the physical annealing process of solids (Aarts and Korst, 1989). For a more detailed discussion of SSA, see Van Groenigen and Stein (1998) (Chapter 3).

As different surveys may have different aims, several optimisation criteria with corresponding fitness functions $\phi(S)$ have been used in the past. The so called MMSD-criterion minimises the expectation of the distance of an arbitrary point to its nearest observation point whereas the MEAN_OK-criterion minimises the mean ordinary kriging variance over the area of interest (Van Groenigen *et al.*, 1999) (Chapter 4). Figure 6.1 shows, for a very simple case, the difference between the two criteria. Figure 6.1a shows the optimised sampling scheme using the MMSD criterion for 23 sampling points. Figure 6.1b uses the MEAN_OK-criterion, for an anisotropic variable. The axis of minimum spatial variation is located in the 0° direction, with the axis for maximum variation perpendicular to it. The MEAN_OK-criterion accounts for the directional effect, placing more observations in the direction of maximum variation (Figure 6.1b)

Although the MEAN_OK-criterion is very useful if we want to optimise our prediction accuracy, it is not necessarily the best criterion in contamination studies. In such studies, accurately delineating hot-spots might be more important than elaborately mapping the whole area, including low-spots (Watson and Barnes, 1995). Therefore, we modified the MMSD-criterion in order to be able to distinguish between different degrees of priority within the areas.

6.2.2. The WMSD-criterion

The WMSD (Weighted Means of Shortest Distances)-criterion is a weighted version of the MMSD-criterion. The fitness function is extended with a location dependent weighing function $w(\bar{x}) : A \rightarrow \mathbb{R}^+$:

$$\min_S = \int_A w(\bar{x}) \|\bar{x} - V_S(\bar{x})\| d\bar{x} \quad (6.2)$$

where $V_S(\bar{x})$ denotes the coordinate vector of the sampling point nearest to \bar{x} . This function is estimated by:

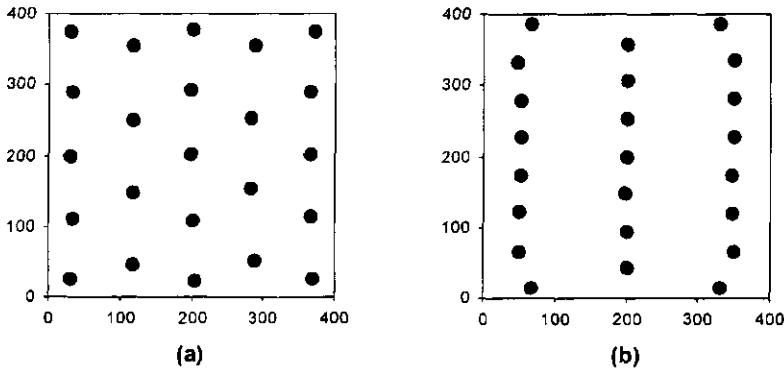


Figure 6.1. Optimised sampling schemes using the MMSD-criterion (a) and the MEAN_OK-criterion in the presence of anisotropy (b)

$$\phi_{\text{WMSD}}(\mathbf{S}) = \frac{1}{n_e} \sum_{j=1}^{n_e} w(\bar{x}_e^j) \|\bar{x}_e^j - V_s(\bar{x}_e^j)\| \quad (6.3)$$

where $\bar{x}_e^1, \dots, \bar{x}_e^{n_e}$ are the nodes of a fine evaluation grid over A .

As an example, a square area with 23 observations is subdivided into two parts (Figure 6.2). The upper half has weighing factor 2.0, the lower half a weighing factor 1.0. Hence $w(\bar{x}) = 1$ if $\bar{x} \in A_1$, and $w(\bar{x}) = 2$ for $\bar{x} \in A_2$. This leads to more intensive sampling in A_2 (15 observations) than in A_1 (8 observations).

Use of the weighing function $w(\bar{x})$ offers a flexible way of using prior knowledge on A and on the possible multivariate character of the survey. In the case of soil contamination, $w(\bar{x})$ can be used to express knowledge on the expected contamination in some parts of the area, e.g. by increasing sampling density on sites with higher expected contamination. In this chapter we present two ways of defining $w(\bar{x})$. The first is using historic knowledge on the severity and location of processes that caused the contamination. The second way of defining $w(\bar{x})$ is based on probability maps of exceeding threshold values using preliminary observations.

6.2.3. Probability maps using Indicator Kriging

Indicator kriging aims at predicting the conditional cumulative distribution function of a ReV at any unvisited location (Deutsch and Journel, 1992). If we consider a ReV $Z(\bar{x})$ with observations $z(\bar{x}^1), \dots, z(\bar{x}^n)$, this is achieved by indicator transforming the observations for cut off value z_c :

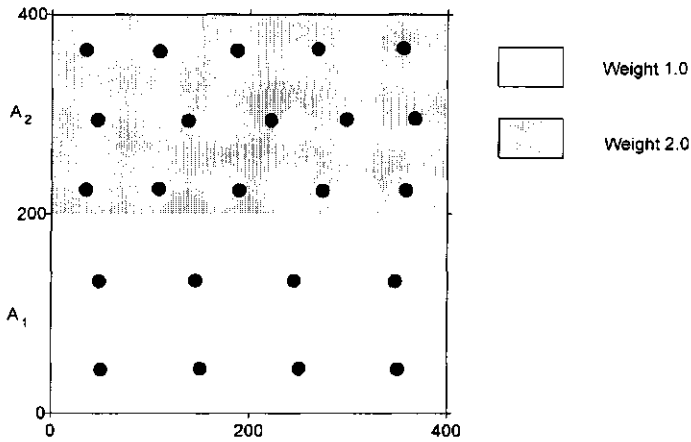


Figure 6.2. Optimised sampling scheme using the WMSD-criterion, with subareas with different weights..

$$i(\bar{x}^j, z_c) = \begin{cases} 1, & \text{if } z(\bar{x}^j) \leq z_c \\ 0, & \text{if } z(\bar{x}^j) > z_c \end{cases} \quad \forall \bar{x}^j \in S \quad (6.4)$$

These transformed variables are interpolated with kriging. The interpolated values, marked with a star, at unvisited location \bar{x}^0 can then be interpreted as a conditional probability:

$$i(\bar{x}^0, z_c)^* = \Pr\{Z(\bar{x}^0) \leq z_c | S\} \quad (6.5)$$

By repeating this procedure for several cut off values, a conditional cumulative distribution function can be estimated at each location by kriging of the indicator transformed variables.

This nonlinear transformation is especially useful when a skewed distribution of $Z(\bar{x})$ prohibits the use of linear geostatistical methods such as ordinary kriging (Journel, 1983). Goovaerts (1997b) showed that indicator techniques can be more effective in delineating a contamination than ordinary kriging. Van Groenigen *et al.* (1997) (Chapter 2) showed how probability maps produced by indicator kriging could guide additional sampling towards possibly contaminated sites. They found that such a multi stage sampling design led to more accurate site characterisation, with much smaller type-I errors, and comparable type-II errors, as compared to a rectangular grid and a random sampling scheme. However, additional sampling sites were chosen by simple random sampling. Using SSA, this approach can be optimised.

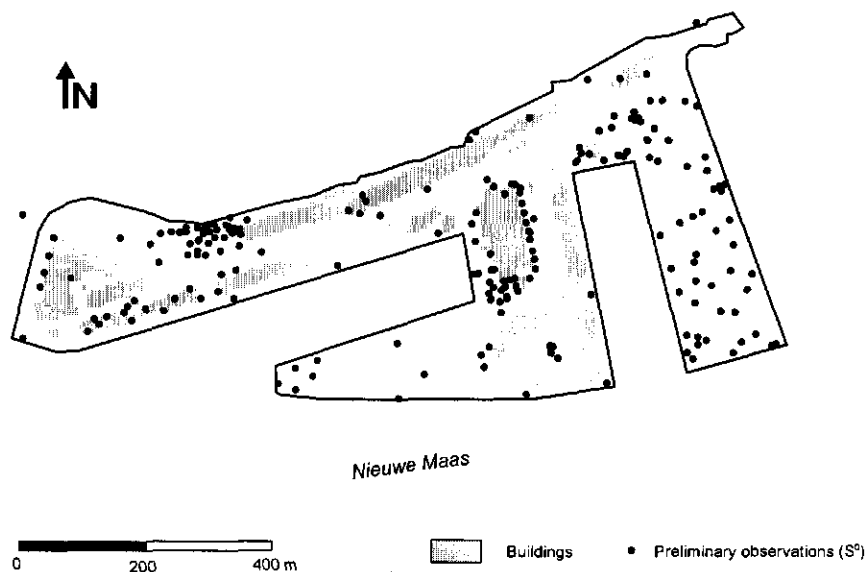


Figure 6.3. An overview of the research area with sampling constraints and preliminary observations (S^0).

6.3. Case-study: the Rotterdam Harbour

An optimal sampling scheme was developed for an urban contamination site A in the Rotterdam harbour, where many preliminary data and sampling constraints were met. The WMSD-criterion was selected to use this information to optimally locate additional observations.

In the Netherlands, legislation on soil remediation recognises four threshold values, defined for each potential contaminant. Sites with concentrations below the S-threshold are considered non-contaminated. Sites with concentrations above the I-threshold are severely contaminated. The $\frac{1}{2}(S+I)$ value is used as a threshold for further research. Both the S- and I-thresholds depend on the clay and organic matter content of the soil, recognising the influence of these parameters on the chemical behaviour of the soil. Additionally, concentrations above the BAGA-threshold require more expensive action (Anonymous, 1994; Anonymous, 1995).

6.3.1. Description of study area

The study area is located along the river Maas, in the old harbour near the centre of Rotterdam (Figure 6.3). As harbour activities are withdrawing from this location, the old area increasingly allows more urban uses such as housing and office building. The size of the area is approximately 30 ha.

The harbour area developed around 1900, due to increasing industrial activity around Rotterdam. On the eastern pier, a stevedores firm and several silos and warehouses settled. A passenger terminal was located on the western pier. A power plant was located at the entrance area near the western pier, and a gas plant west of that. With modernisation of the merchant fleet after the second world war, harbour activities moved to areas further away from the city centre, leaving the site with a variety of light industrial and commercial activities (Pieters *et al.*, 1996).

The area is to be converted into a mixed housing, recreational and commercial area within the next 10 years (City of Rotterdam, 1997). Therefore, a survey of the area on possible contamination due to former industrial activity had to be carried out.

The soil consists of a sandy layer overlaying the original Holocene clays. This layer was applied for heightening and varies in depth from around 5 m on the piers, to 2 m on the north west part of the area. No peat was found within 5 m depth.

6.3.2 Preliminary data and sampling constraints

Based upon historic research and future plans, different priority areas for sampling were delineated. We distinguished five historical factors:

- i) filling in of harbour
- ii) heightening of surface level
- iii) type of industrial activities
- iv) boundaries of ownership parcels
- v) historic remediations

This information was combined with maps showing the urgency of the survey for future activities, yielding a priority map. Classification ranged from 1 (low priority due to no expected contamination and/or no immediate building plans) to 4 (high expected contamination and/or immediate building plans) (Pieters *et al.*, 1996). The area *A* can be subdivided according to these classifications in sub-areas A_1, \dots, A_4 , respectively (Figure 6.4).

Much data on contamination was available from preliminary surveys and earlier remediations, yielding a total of 201 observation locations. Three heavy metals (Pb, Cu

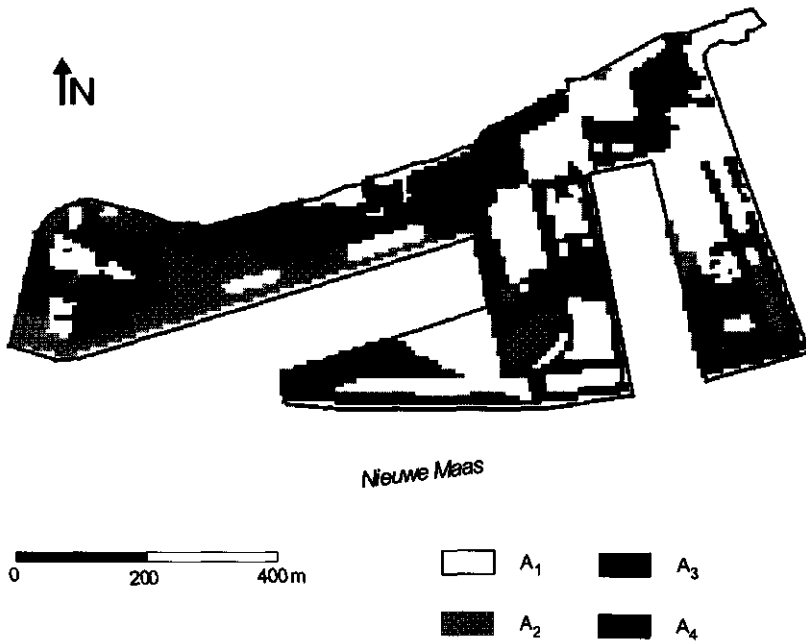


Figure 6.4. Priority map of sampling, based upon expected contamination and urgency of remediation ranging from A_1 (low priority) to A_4 (high priority).

and Zn) and 2 carbohydrates (mineral oil and PAH's) were considered, as they were expected to constitute the main contamination. Not all variables were sampled at each of the 201 locations, resulting in a different number of observations for each contaminant. This preliminary sampling scheme is denoted by S^0 . As data were collected at different depths, for practical reasons two layers were considered for analyses: 0-1 m (L_1) and 1-5 m (L_2), respectively. The observations of S^0 are strongly clustered, with high concentrations of points on the western pier and around the former gas- and electricity plants. As several buildings in the area were of architectural or historical value, sampling was constrained by both boundaries and built-on areas (Figure 6.3).

6.3.3. Step 1: additional sampling using historic information

As a first step, 80 additional sampling points were selected, to cover the areas that were left unsampled by S^0 . To optimise such an additional sampling scheme S^1 , the weight function $w_{S^1}(\bar{x})$ was defined. We assigned to the four subareas

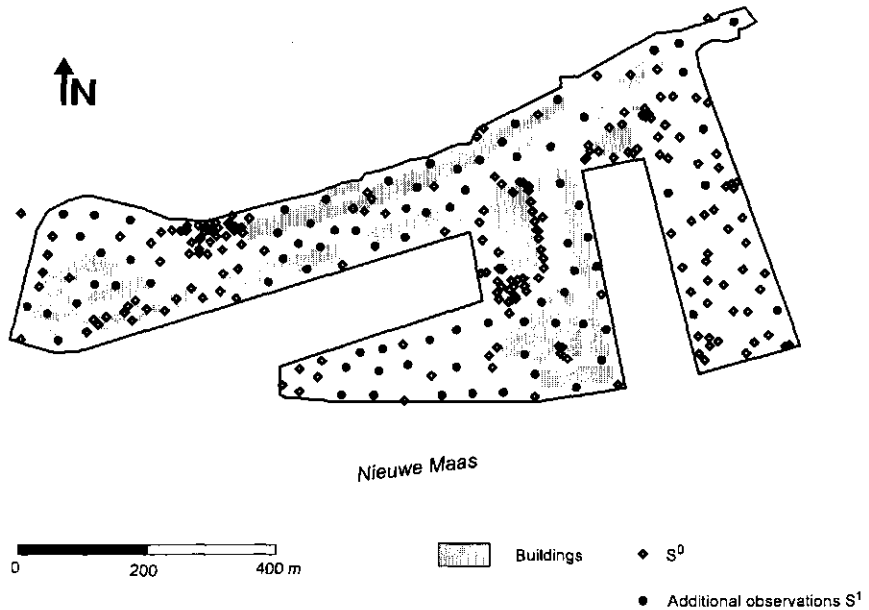


Figure 6.5. Additional sampling scheme based on priority map using WMSD-criterion (S^1).

A_1, \dots, A_4 different weights, with the highest weight for the subarea with the highest probability of being contaminated and the highest priority (A_4). After comparing different weight distributions, it was decided by expert judgement to use the weight vector (1.0, 1.5, 2.0, 3.0). By fine tuning after several runs of the SSA algorithm this vector was chosen, as it provided the most satisfactory distribution of points over the different subareas. This weight vector resulted in a coverage of the area that reflected the severity of the different priority area.

6.3.4. Step 2: additional sampling using probability maps

After getting a first overview of the extent of the contamination by ($S^0 \cup S^1$), a second stage of sampling focussed on areas with a high probability of being contaminated, as calculated from the observations. Here, a second weight function $w_{S^2}(\bar{x})$ was used, based on indicator kriging. After choosing a threshold value, probability maps of all 5 contaminants at both depths were calculated, denoted $i_1(\bar{x}, z_1)^*$, ..., $i_{10}(\bar{x}, z_{10})^*$, respectively. As we were interested in the probability of any of the variables reaching high levels of contamination, we used the maximum of all

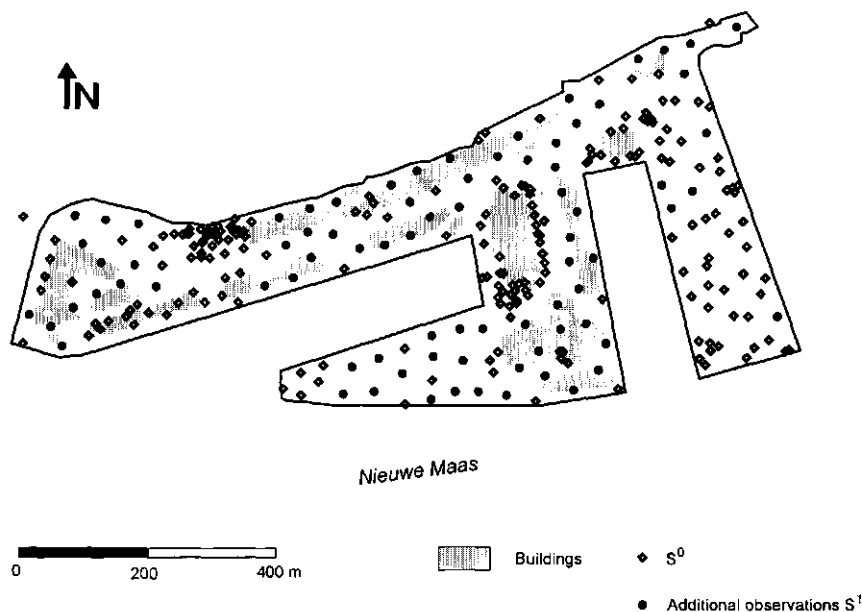


Figure 6.6. *Additional sampling scheme using MMSD-criterion.*

probabilities:

$$w_{S^2}(\bar{x}) = \max(i_1(\bar{x}, z_1)^*, \dots, i_{10}(\bar{x}, z_{10})^*) \quad (6.6)$$

This weight function was used to select the additional sampling scheme S^2 .

6.4. Results

Figure 6.5 shows the optimised additional sampling scheme. As a comparison, figure 6.6 shows an alternative without the weight function, using the MMSD-criterion. Use of priority areas in the design of S^1 influenced the sampling schemes, with more emphasis on areas with higher priority than on areas with lower priority. The WMSD-criterion places 46 and 18 observations at the most urgent priority areas A_3 and A_4 , respectively (Figure 6.5). Using the MMSD-criterion, these numbers are only 32 and 7 (Figure 6.6). Both figures show that sampling constraints are honoured.

Table 6.1 describes the data set for the total data after the first sampling round ($S^0 \cup S^1$). The distribution of all contaminants is skewed, with only a few outliers

Table 6.1. Descriptive statistics for S^0 and S^1 ($mg\ kg^{-1}$).

Poll.	Layer	n	Mean	St.dev.	Max	Min
Zn	1	161	129	127	1300	10
	2	127	102	208	1500	10
Pb	1	162	73	129	1350	1
	2	138	71	260	2700	1
Oil	1	148	50	110	950	2
	2	113	94	419	2700	2
PAH	1	142	6	17	161	0
	2	111	42	284	2756	0
Cu	1	161	30	61	580	0
	2	156	35	197	2300	0

above the **I**- and **BAGA**-thresholds. These contaminated locations are shown in Figure 6.7 by smaller and larger circles, respectively. The **I**-threshold was exceeded at 35 locations for at least one of the contaminants. Three of these 35 were above the **BAGA**-threshold. These last three were closely located to the former gas and electricity plants.

To select the 30 points of S^2 , the exceedance probability of the **I**-threshold would be a logical choice. However, the high proportion of nugget effect in most indicator variograms (Table 6.2) indicated that spatial variation was extremely high. Therefore, because local hot-spots could still have been missed by sampling schemes S^0 and S^1 , we applied the $\frac{1}{2}(S+I)$ -threshold, for which indicator variograms were calculated (Table 6.2).

As an example, Figures 6.8 and 6.9 show the interpolated probability maps of mineral oil in the second layer, and Zinc in the first layer, respectively. The contamination by mineral oil is only very minor (with only 3 observations above the $\frac{1}{2}(S+I)$ threshold). The probability that Zinc exceeds this threshold in most parts of the area is much larger.

Figure 6.10 shows the combined probability map $w_{S^2}(\bar{x})$. Probabilities of exceeding the $\frac{1}{2}(S+I)$ threshold exceed 0 everywhere, but clear hot-spots are located around the

Table 6.2. Spherical indicator variograms for all contaminants for threshold value $\frac{1}{2}(S+I)$ ($mg\ kg^{-1}$).

Poll.	Layer	Nugget	Sill	Range
Cu	1	1.0	0.0	-
	2	1.0	0.0	-
Pb	1	0.54	0.46	100.0
	2	1.0	0.0	-
Oil	1	0.39	0.61	85.0
	2	0.45	0.55	76.0
PAH	1	1.0	0.0	-
	2	1.0	0.0	-
Zinc	1	0.72	0.28	86.0
	2	1.0	0.0	-

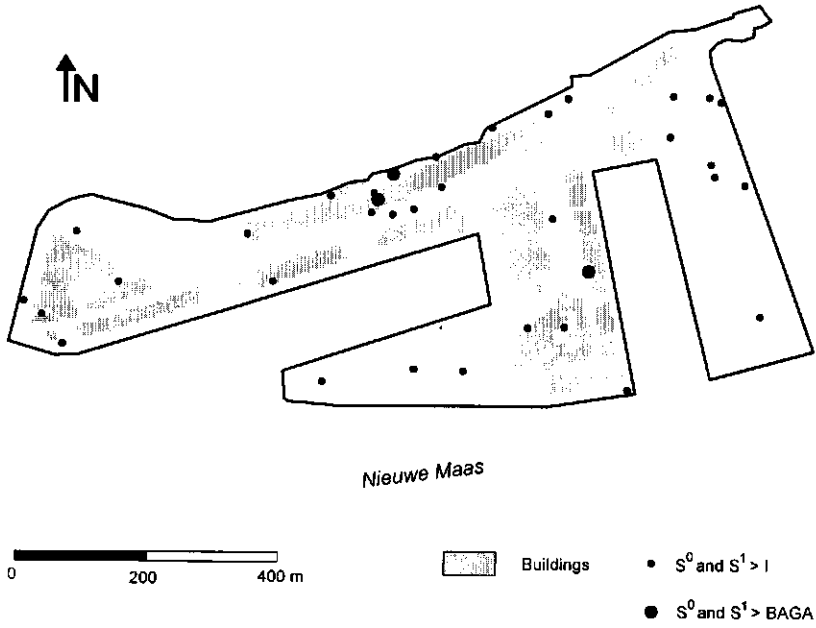


Figure 6.7. Location of contaminated areas.

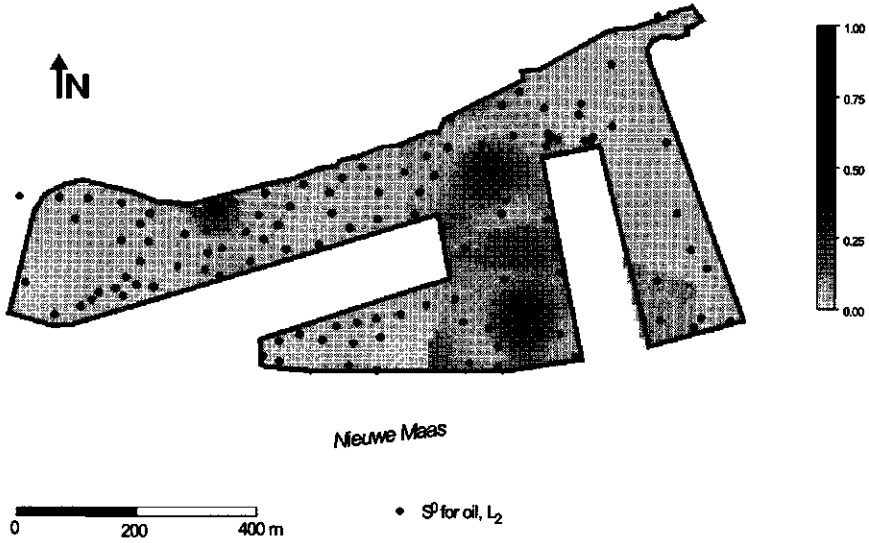


Figure 6.8. Probability of exceeding the $\frac{1}{2}(S+I)$ threshold for mineral oil in L_2 .

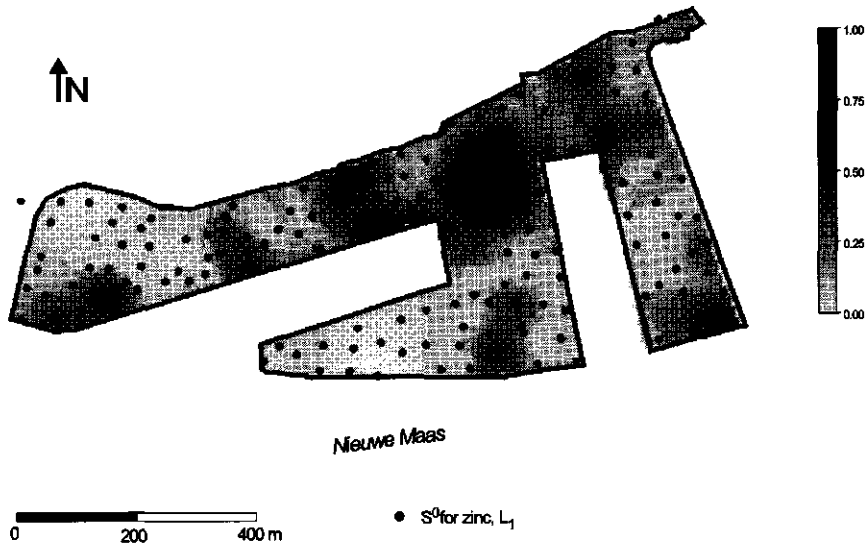


Figure 6.9. Probability of exceeding the $\frac{1}{2}(S + I)$ threshold for Zinc in L_1

former electricity and gas-plants. Using this map, 30 additional sampling points (S^2) were determined (Figure 6.11). The observations are located close to the former plants, and at the northern part of the eastern pier.

Table 6.3 presents descriptive statistics for S^2 . As not all analyses were taken in both layers L_1 and L_2 , the total number of observations is not 30 but 20 and 26, respectively. The observations exceeding the I- or BAGA-thresholds are indicated by the smaller and larger triangles, respectively (Figure 6.12). Table 6.4 compares the percentage of pollution of $S^0 \cup S^1$ with that of S^2 . The percentage of observations

Table 6.3. Descriptive statistics for S_2 (mg kg^{-1}).

Poll.	layer	n	mean	st.dev.	max	min
Zn	1	20	128	68	290	5
	2	26	128	99	396	20
Pb	1	20	99	100	380	13
	2	26	223	590	2800	13
Oil	1	20	51	52	240	20
	2	26	314	1425	7300	20
PAH	1	20	24	49	198	0
	2	26	2	3	10	0
Cu	1	20	47	77	350	5
	2	26	33	39	160	5

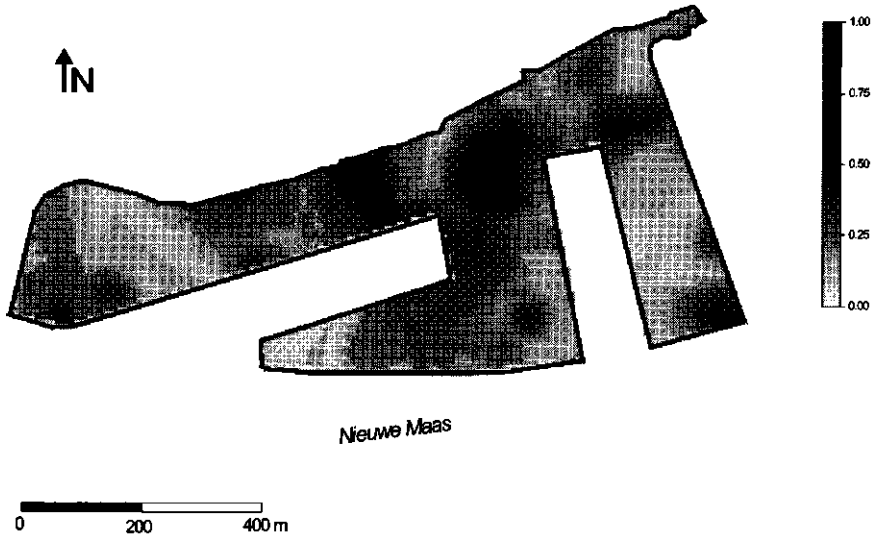


Figure 6.10. *Maximum probability of exceeding the $\frac{1}{2}(S + I)$ threshold for any of the contaminants and both depths.*

above the I-threshold increases from 17.4% to 26.7%, respectively. For the BAGA-threshold these numbers are 1.5% and 16.7%. Apart from several observations of S^2 being relatively close to already observed contaminations, some new sites above the BAGA-threshold were discovered. These are located at the northern part of the eastern pier, and the southern part of the entrance area.

6.5. Discussion and conclusions

In this chapter, we showed the use of SSA in optimising a spatial sampling scheme for a multivariate contamination in an urban area. The WMSD-criterion can use any spatial weight function $w(\bar{x})$ to focus sampling on areas with high priority and reduce sampling on areas with low priority. Two examples represent different stages in the case study. At the first stage, the weight function $w_{S^1}(\bar{x})$ differentiated between areas of priority. The optimised sampling scheme includes existing observations, and sampling constraints are acknowledged. At the second stage probability maps of exceeding an environmental threshold value were combined into a single weight map $w_{S^2}(\bar{x})$. The sampling scheme thus derived detected several hot-spots that were left unnoticed before (Table 6.4).

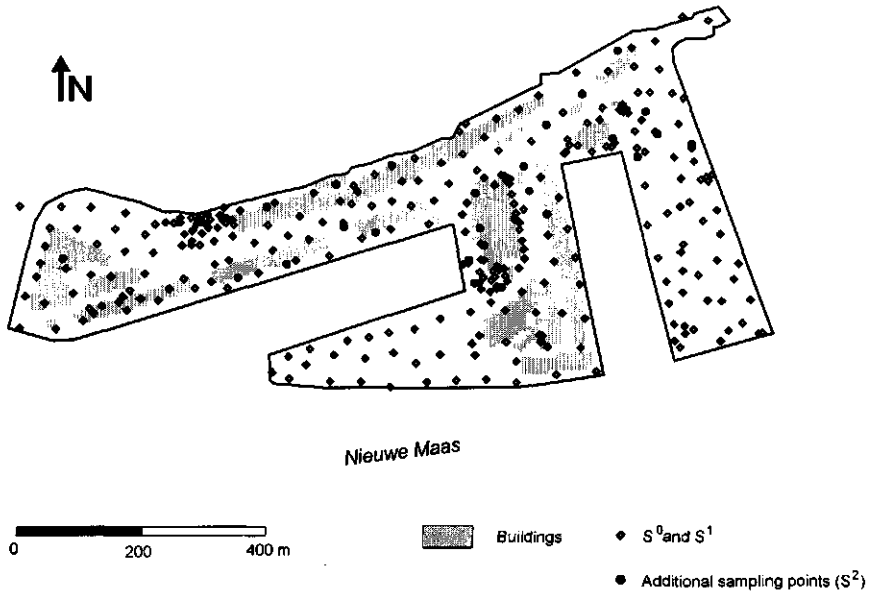


Figure 6.11. *Additional sampling scheme based on maximum probability of exceedance (S^2).*

A critical issue in the presented methods is the setting of the weighing factors. In the first stage the (semi quantitative) priority values have been transformed into weighing factors using expert knowledge. During the second stage, probabilities of exceeding the threshold values have been used as weighing factors. These choices are to a certain extent arbitrarily, and more research should be dedicated to the best setting of these weighing factors. On the other hand, the possibility of attaching priorities to certain areas may prove a valuable tool in decision making processes.

In theory, it would improve the straightforwardness of the optimisation algorithm to use a combined, two-part fitness function. For example a weighted average may be considered:

$$\phi_{\text{WSUM}}(S_i) = \lambda_1 \phi_1(S_i) + \lambda_2 \phi_2(S_i) \quad (6.7)$$

Table 6.4. *Percentage of observations above the different threshold values, for the two different sampling stages.*

Sampling scheme	n	> T (%)	> I (%)	> BAGA (%)
$S^0 \cup S^1$	201	31.8	17.4	1.5
S^2	30	43.3	26.7	16.7

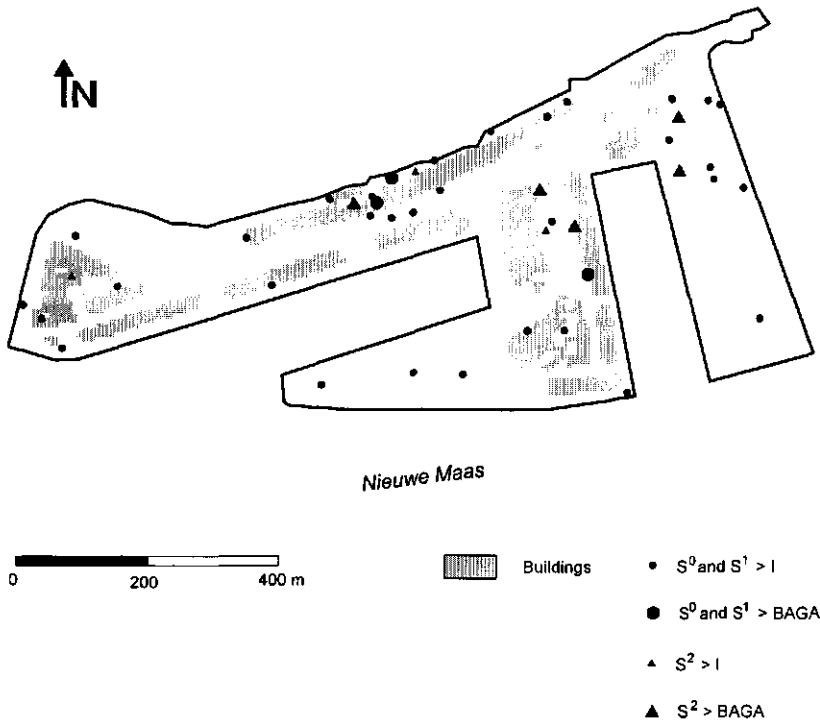


Figure 6.12. Location of heavy and very heavy contaminated areas.

However, this would lead to problems with the weights derived from indicator kriging (IK). IK does not really provide probabilities of exceeding threshold values, but *predictions* of these probabilities. As the accuracy of these predictions is not accounted for (contrary to ordinary kriging, where the kriging variance can be used for that purpose), it is essential that the sampling scheme is adequate for interpolation before applying IK. Therefore, a phased approach was necessary in this study, where the first sampling stage was used to create a sampling scheme that covered the area satisfactorily.

Table 6.4 and Figure 6.12 indicate the efficiency of the second sampling scheme in detecting contaminated sites. Although validation with stochastic simulations (e.g. Goovaerts, 1997b) would be preferable, simulating multivariate stochastic fields still poses serious problems. Modelling of spatial dependence would involve estimation of a large number of (cross)-variograms. This problem only increases in magnitude if several indicator variables are used. Deutsch and Journel (1992) mention different solutions to this problem. These solutions, however, require either all variables to be

known at most observation points, or one of the variables to be known at each simulated point. Ignoring co-regionalisation and simulating the contaminants separately would lead to a huge overestimation of the contaminated area. Therefore, a validation based on stochastic simulations was not tried.

The proposed method of handling the multivariate problem does not require the use of a model of co-regionalisation, thus avoiding elaborate cross variogram calculations. Although this approach is somewhat pragmatic, it is very valuable in practice. Furthermore, we should like to emphasise that this method only applies to optimisation of the sampling scheme. Afterwards, data processing can proceed with any form of kriging, cokriging, simulation or other geostatistical procedure.

This chapter focussed on optimisation of spatial sampling for geostatistical analysis. Recently, a discussion arose in the soil science community on the benefits of geostatistics as compared to classical statistics (e.g. Brus and De Gruijter, 1997; Brus and De Gruijter, 1993). Using classical sampling theory, sampling schemes should always have a random component (e.g. a random origin of a sampling grid in systematic sampling), as all parts of the area should have an inclusion probability larger than 0. Therefore, optimising point locations as has been done in this paper is only allowed in geostatistics. Several optimisation methods using classical sampling theory have been proposed in the past (e.g. Stevens, 1997; Domburg *et al.*, 1997). Although we recognise the significant role that classical sampling theory should play in spatial statistics, we think that geostatistics can be more flexible in dealing with sampling constraints and earlier measurements, as has been demonstrated in the case study.

The main requirement of a sampling strategy in the practice of urban multivariate contamination studies is flexibility and robustness. In this paper we showed that SSA is able to handle the deviations, sampling constraints and preliminary observations. Furthermore, we showed that by distinguishing between priority areas, SSA can be a helpful tool in decision making processes.

6.6. Acknowledgements

The authors would like to thank Coen van Tooren for his advice and support during the starting phase of this study, and an anonymous reviewer for some very useful remarks and suggestions on the notation. Furthermore, we would like to thank the city of Rotterdam for having us use the data for this publication. Finally, contributions of Maaïke Broos to the discussions on multivariate simulations are gratefully acknowledged.

Chapter 7

Establishing Soil/Yield Relations in Precision Agriculture Studies¹

Abstract

In this chapter it is shown how yield maps may be used to optimise soil sampling for precision agriculture in a low-tech environment. The proposed method is applied in an on-farm study in Niger. Using a cheap, low-tech scoring technique, yield maps of millet were produced. Yield varied from 0 to 2500 kg ha⁻¹. Subsequently, the Spatial Simulated Annealing (SSA) algorithm was used to optimise three sampling schemes. Scheme 1 optimised coverage over the whole area. Scheme 2 covered the whole yield range. Scheme 3 covered the low-producing areas. Using correlation coefficients, scheme 2 found significant correlations between 5 variables and yield. Scheme 1 found only one significant correlation. Using multivariate regression of yield on soil variables, scheme 2 explained 70% of the variation in yield. For scheme 1, this was only 37%. Differences between scheme 3 and scheme 1 proved to be significant for distance to shrubs, micro-relief, pH-H₂O and CEC. From this study, we concluded that shrubs are the main factor influencing millet yield by means of catching eroded materials and improving soil fertility. The possibilities of planting shrubs to improve soil fertility should be investigated. Variograms of micro-relief and yield suggested that spatial correlation is largely confined to distances of 3 to 5 m.

¹ Based on: Van Groenigen, J.W., Gandah, M. and Bouma, J. (submitted). *Soil sampling strategies for precision agriculture research under Sahelian conditions*. Soil Science Society of America Journal.

7.1. Introduction

In recent years, an increased interest in soil variability on the field scale may be observed. Precision agriculture has become an important research topic in soil science. As one of the major quantitative techniques of dealing with spatial variability, geostatistics plays an important part in these developments. Well-known geostatistical interpolation techniques like ordinary kriging and indicator kriging have been applied for optimal interpolation of point observations (e.g. Van Uffelen *et al.*, 1997; Stein *et al.*, 1997). Also, stochastic simulation techniques (e.g. Gómez-Hernández and Srivastava, 1990; Goovaerts, 1997a) offer interesting perspectives in scenario studies.

Apart from these applications and adaptations of existing geostatistical tools, precision agriculture poses some more specific challenges to geostatisticians. One of these is the increasing availability of maps of auxiliary data that can be potentially of help for purposes such as mapping of soil properties or yield prediction. Examples of such auxiliary data are maps of soil tillage resistance (Van Bergeijk and Goense, 1997), Remote Sensed imagery (Booltink and Verhagen, 1997), and yield maps collected using low-tech (Stoorvogel, 1995) or high-tech (Bouma, 1997) approaches.

In this chapter, we focus on the question how (relatively cheap) auxiliary data can be used to optimise collection of soil samples for (expensive) chemical analysis.

The developed methodology was applied in a case study in Niger. In a farmers' field, yield predictions were made using a quick, low-tech scoring technique. These yield maps were used to guide sampling of soil properties to potentially interesting sites. Special attention was paid to the specific possibilities and constraints of precision agriculture in a low-tech environment.

7.2. Materials and Methods

7.2.1. Study area

The study area is located on an on-farm millet field near the village of Tchigo Tagui, approximately 80 km north east of the city of Niamey, western Niger. The field is located on a laterite plateau with eolian sand deposits. The soils can be classified as plinthustalfs (Soil Survey Staff, 1996). The rainy season is from May to September, and mean annual rainfall is around 480 mm (Sivakumar *et al.*, 1993). The soils are generally characterised by a high spatial variability in terms of chemical and physical fertility. Bouma *et al.* (1996) mention five reasons for this: *i*) micro-relief and crusting, causing crucial redistribution of water infiltration (Gaze *et al.*, 1997), *ii*) termites, that may enrich the soil locally, *iii*) local effects of trees and shrubs, *iv*) differences associated

with landscape position, and *v*) soil fertility gradients around villages. As the area is confined to one farmers' field relatively far from the village, only the first three aspects play an important role in this study.

Brouwer *et al.* (1993) reported a yield stabilising effect of this micro-variability on similar fields. In relatively wet years, yields were highest at sites with a high micro-relief, due to catchment of fertile eroded particles. However, in dry years these spots were non-productive due to lack of water. In these cases the lower areas ensured a subsistence level of millet yield.

Since both fertility and water supply are extremely low, even moderate spatial variability of these can have profound influence on crop yields. Stein *et al.* (1997) found a yield range of 0 to 2885 kg ha⁻¹, measured on 5 x 5 m blocks. They were able to explain 30 % of the yield variability by multiple regression on soil variables. Gandah *et al.* (1998) used the same support size for yield sampling, and were able to explain only 5-28 % of the yield variability by regression.

These results, combined with observations in the field, suggested that the support size of yield sampling should be smaller to capture most of the soil variability. In this study the support size was on a hill level, *i.e.* each planting hill (consisting of 2-6 individual millet tillers) was evaluated using a low-tech scoring technique in order to get a crude estimation of the yield. This data was used to optimise the sampling scheme.

7.2.2. Spatial Simulated Annealing

Spatial Simulated Annealing (SSA) is an algorithm that was designed for optimisation of spatial sampling schemes (Van Groenigen and Stein, 1998) (Chapter 3). Its features include the incorporation of preliminary observations, and taking into account sampling constraints and boundaries. Furthermore, SSA allows the use of several quantitative optimisation criteria, among them minimisation of the ordinary kriging variance and estimation of the experimental variogram (Van Groenigen *et al.*, 1999) (Chapter 4). In this study, use was made of the MMSD-criterion, which aims at an even spreading of the observations over the area of interest. This is done by minimising the expectation of the distance between an arbitrarily chosen location, and its nearest observation point. This leads to the following minimisation function:

$$\min_S \int_A \|\bar{x} - V_S(\bar{x})\| \quad (7.1)$$

where S denotes the sampling scheme, A represents the areas of interest, and \bar{x} is a random location vector in A . $V_S(\bar{x})$ represents the location vector of the nearest sampling point $\bar{x}_s \in S$ to \bar{x} . As this minimisation function can generally not be solved

analytically, we estimated it using the following function:

$$\phi_{\text{MMSD}}(\mathbf{S}) = \frac{1}{n_c} \sum_{j=1}^{n_c} \|\bar{\mathbf{x}}_c^j - \mathbf{V}_S(\bar{\mathbf{x}}_c^j)\| \quad (7.2)$$

where the location vectors $\bar{\mathbf{x}}_c^1, \dots, \bar{\mathbf{x}}_c^{n_c}$ denote the nodes of a fine raster grid over \mathbf{A} . The optimisation process starts with an initial sampling scheme \mathbf{S}_0 , consisting of randomly drawn locations over \mathbf{A} . Subsequently, an alternative sampling scheme \mathbf{S}_1 is derived from \mathbf{S}_0 by a transformation of one of the sampling locations over a random vector. The probability of \mathbf{S}_1 being accepted as a basis for further optimisation depends on the Metropolis criterion:

$$\begin{aligned} P_c(\mathbf{S}_i \rightarrow \mathbf{S}_{i+1}) &= 1, & \text{if } \phi(\mathbf{S}_{i+1}) \leq \phi(\mathbf{S}_i) \\ P_c(\mathbf{S}_i \rightarrow \mathbf{S}_{i+1}) &= \exp\left(\frac{\phi(\mathbf{S}_i) - \phi(\mathbf{S}_{i+1})}{c}\right), & \text{if } \phi(\mathbf{S}_{i+1}) > \phi(\mathbf{S}_i). \end{aligned} \quad (7.3)$$

This criterion ensures that occasionally also inferior solutions are accepted, thereby avoiding prematurely ending of the optimisation process in local minima. As the process continues, parameter c decreases. In this way, the sampling scheme 'freezes' into a more optimal configuration.

For an extensive discussion on SSA and the implementation of the MMSD-criterion, see Van Groenigen and Stein (1998) (Chapter 3).

7.2.3. Soil sampling

Three different soil sampling schemes were applied, in order to test the benefits of the different approaches. Samples were taken at three depths (0-0.1 m, 0.1-0.2 m and 0.2-0.4 m) and were analysed for pH-H₂O, texture, P-Bray, CEC and OM. In order to make the case study realistic in terms of financial constraints usually met in precision agriculture research, each sampling scheme consisted of 27 observations only. The sampling schemes are listed below:

- i) The first sampling scheme (\mathbf{S}_1) aims at even spreading of the sampling points over the total area (the three plots). For 9 locations per plot, SSA yielded a regular square grid.
- ii) Sampling scheme \mathbf{S}_2 aims at covering the whole range of yields, in order to establish

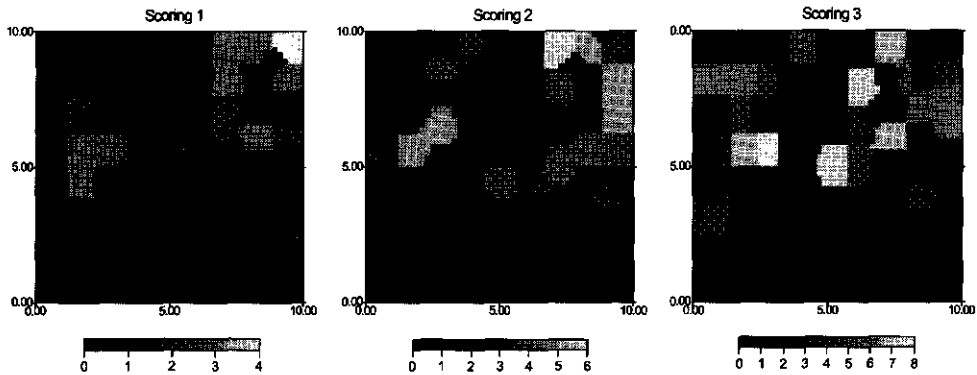


Figure 7.1. Results of the three scorings of plot 3, presented using Thiessen polygons.

clearer relationships between soil characteristics and yield. This was done by stratifying the area according to yield, and distributing a number of samples optimally over each of the strata using SSA.

- iii) Sampling scheme S_3 focuses specifically on low-producing areas. This may be especially useful for precision agriculture purposes when a detailed survey of low-spots is desired for remedial action. To define low-producing areas, use was made of the threshold value of $250 \text{ kg ha}^{-1} \text{ yr}^{-1}$. This threshold was mentioned in Stein *et al.* (1997) as the minimum yield required for a family of 10 persons with 8 ha of cropping land.

In addition, a detailed micro-relief map was produced, as former studies showed a strong influence of topography on yield (Stein *et al.*, 1997; Brouwer *et al.*, 1993). Using a level, observations were collected on a square $1\text{m} \times 1\text{m}$ grid. Since earlier studies suggested an influence of shrubs on yield variability, a variable 'distance to nearest shrub' was included.

7.3. Results

7.3.1. Yield maps

In order to get a cheap, low-tech prediction of millet yield, a semi-quantitative scoring technique as presented in Buerkert *et al.* (1995) was used. Scoring was performed at July 4th, August 8th and September 4th 1997 on 3 plots of $10\text{m} \times 10\text{m}$ each. Scoring values were assigned to individual hills, ranging from 0 (no plants at all) to 8 (maximal

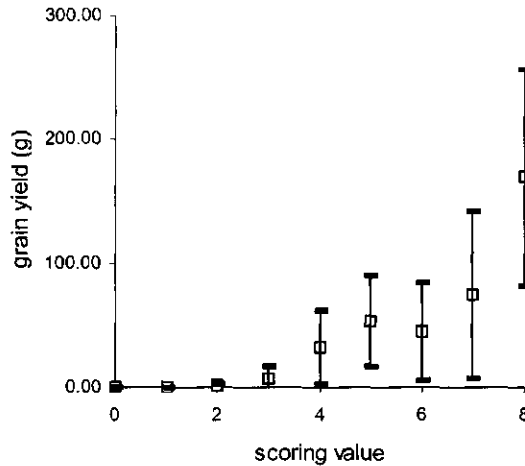


Figure 7.2. Relationship between scoring value and yield (mean and standard deviation), measured per hill.

aerial biomass). The millet was harvested around November 7th.

Figure 7.1 shows the scoring results of plot 3 at the three scoring dates. Scoring data are represented using Thiessen polygons, as the semi-quantitative character prohibits direct interpolation using kriging. Since the discriminative features were rather limited at the earlier scoring dates, scoring ranges differed from 0-4 (July 4th), 0-6 (August 8th) to 0-8 (September 4th).

After the last scoring, several hill samples were collected to establish the relationship between scoring and yield. For each of the 9 scoring classes of September 4th, 8 hills were harvested, the heads were threshed in a traditional manner and weighted separately. The results are shown in Figure 7.2. From this figure, it can be concluded that the scoring should not be used directly as a quantitative measure that is linearly related to the yield, but should be calibrated using real yield measurements. Scoring values of 0 to 3 resulted all in a yield very close or equal to 0. Scoring classes 5 to 7 did not result in a significant difference in yield.

Using this calibration data, the scoring data per hill were transformed into estimated yield data. This was interpolated using ordinary kriging, resulting in the predicted yield maps shown in Figure 7.3a-c. Table 7.1 shows the variogram parameters for the yield. These can be characterised by a relative high nugget effect (54 % and 48 % for plots 2 and 3, respectively). Furthermore, ranges of 3.4 and 4.2 m confirm suppositions of high variability at very short distances.

A survey of any shrubs on the plots was made, as previous studies showed that shrubs can have a considerable influence on soil and yield variability (Brouwer *et al.*,

Table 7.1. Variogram models for the yield estimates and micro-relief.

Variable	Plot no.	Variogram model	Nugget	c / slope	range / r [m]
yield (kg/ha)	1	linear	140800	21120	-
yield (kg/ha)	2	exponential	159300	134991	3.42
yield (kg/ha)	3	exponential	165200	182000	4.15
relief (m)	1	linear	0.030	0.023	-
relief (m)	2	spherical	0.012	0.020	3.78
relief (m)	3	spherical	0.001	0.073	4.27

1993). The position of the shrubs is also shown in Figure 7.3a-c.

7.3.2. Soil data

Table 7.1 shows fitted variogram models for micro-relief, measured at the 1 m x 1 m grid. Starting at the boundaries of the plots, this resulted in a data set of 121 observations for each plot. Although the variograms differ much more in character than those of the yield predictions, ranges are still shorter than 5 m. Figure 7.3d-f

Table 7.2. Descriptive statistics of measured variables for all three sampling schemes.

Variable	Depth	Scheme 1			Scheme 2			Scheme 3		
		Mean	SD	CV	Mean	SD	CV	Mean	SD	CV
pH-H ₂ O	0-0.1	5.1	0.3	0.06	5.1	0.2	0.05	4.8	0.3	0.05
	0.1-0.2	5.0	0.3	0.06	5.0	0.2	0.05	4.8	0.2	0.04
	0.2-0.4	4.9	0.3	0.07	4.9	0.2	0.05	4.7	0.2	0.04
CEC, cmol kg ⁻¹	0-0.1	1.08	0.30	0.28	1.21	0.33	0.27	1.29	0.22	0.17
	0.1-0.2	1.32	0.57	0.43	1.30	0.30	0.23	1.32	0.34	0.25
	0.2-0.4	1.74	0.14	0.42	1.74	0.45	0.26	1.81	0.44	0.24
Sand, %	0-0.1	89.9	2.3	0.03	89.9	2.1	0.02	89.2	2.2	0.02
	0.1-0.2	87.5	3.4	0.04	87.7	2.3	0.03	87.2	2.3	0.03
	0.2-0.4	82.6	5.2	0.06	82.8	5.0	0.06	82.0	5.2	0.06
Silt, %	0-0.1	4.0	0.7	0.19	4.0	1.0	0.25	4.1	0.9	0.22
	0.1-0.2	4.0	0.9	0.24	4.0	0.8	0.20	4.0	0.7	0.18
	0.2-0.4	3.5	0.9	0.26	3.5	0.9	0.26	3.7	0.9	0.24
Clay, %	0-0.1	6.1	1.8	0.30	6.1	1.6	0.26	6.8	1.8	0.27
	0.1-0.2	8.6	3.1	0.36	8.4	1.9	0.23	8.8	2.1	0.24
	0.2-0.4	14.0	4.9	0.35	13.9	4.7	0.34	14.4	4.8	0.33
OM, g kg ⁻¹	0-0.1	0.28	0.09	0.31	0.29	0.08	0.26	0.25	0.05	0.20
	0.1-0.2	0.27	0.01	0.19	0.27	0.06	0.22	0.25	0.06	0.25
	0.2-0.4	0.25	0.05	0.21	0.27	0.06	0.21	0.28	0.07	0.24
PBray, mg kg ⁻¹	0-0.1	3.8	1.3	0.33	3.5	1.3	0.36	3.6	1.7	0.47
	0.1-0.2	3.0	1.3	0.43	2.5	1.2	0.48	2.5	1.3	0.51
	0.2-0.4	1.9	0.5	0.24	1.5	0.7	0.43	1.5	1.0	0.66
Relief, m	-	0.048	0.004	0.48	0.053	0.021	0.40	0.034	0.019	0.56
Distance, m	-	2.64	1.47	0.56	2.12	1.76	0.83	3.30	1.30	0.40

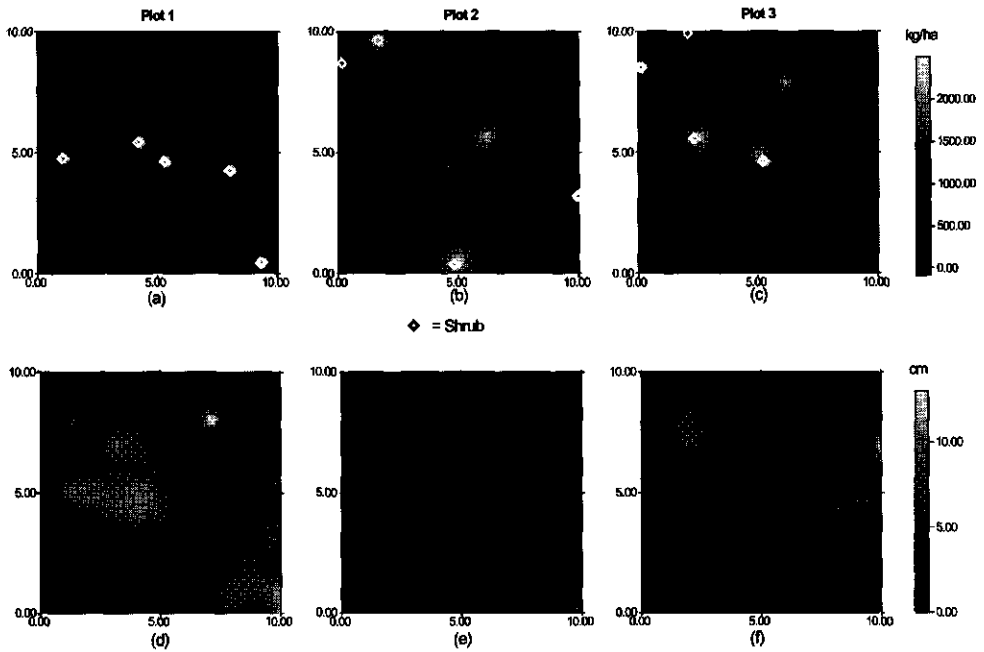


Figure 7.3. Interpolated yield estimates with position of shrubs (a-c) and interpolated micro-topography (d-f).

shows the interpolated micro-relief for the three plots using ordinary kriging. The maximum difference in micro-relief within a plot is 0.13 m (measured at plot 1).

The three sampling schemes were established as follows:

- i) Sampling scheme S_1 is shown in figure 7.4a-c. As it is not designed using the yield data, it misses several important features, notably all hot-spots of plot 2. Table 7.2 summarises the soil data of S_1 . Texture varies from sand in the upper soil, to loamy sand (due to clay illuviation) in the subsoil. This results in an increasing CEC with depth. Furthermore, organic matter content and P-Bray are decreasing with depth. Coefficients of Variation (CV's) are reasonably low ($< 50\%$) for all variables, except for distance to the nearest shrub (56%)
- ii) Sampling scheme S_2 is shown in figure 7.4d-f. As this scheme aims at covering a wide range of yields, most of the hot-spots are sampled. All low producing areas (e.g. the lower left corner of plot 3) are covered. Most descriptive statistics of S_2 are close to those of S_1 . However, standard deviation of relief is much higher in S_2 , and mean distance to shrubs is much smaller.
- iii) The area with a predicted yield lower than the threshold of $250 \text{ kg ha}^{-1} \text{ yr}^{-1}$ was delineated in the three plots, and sampling points of S_3 were evenly distributed.

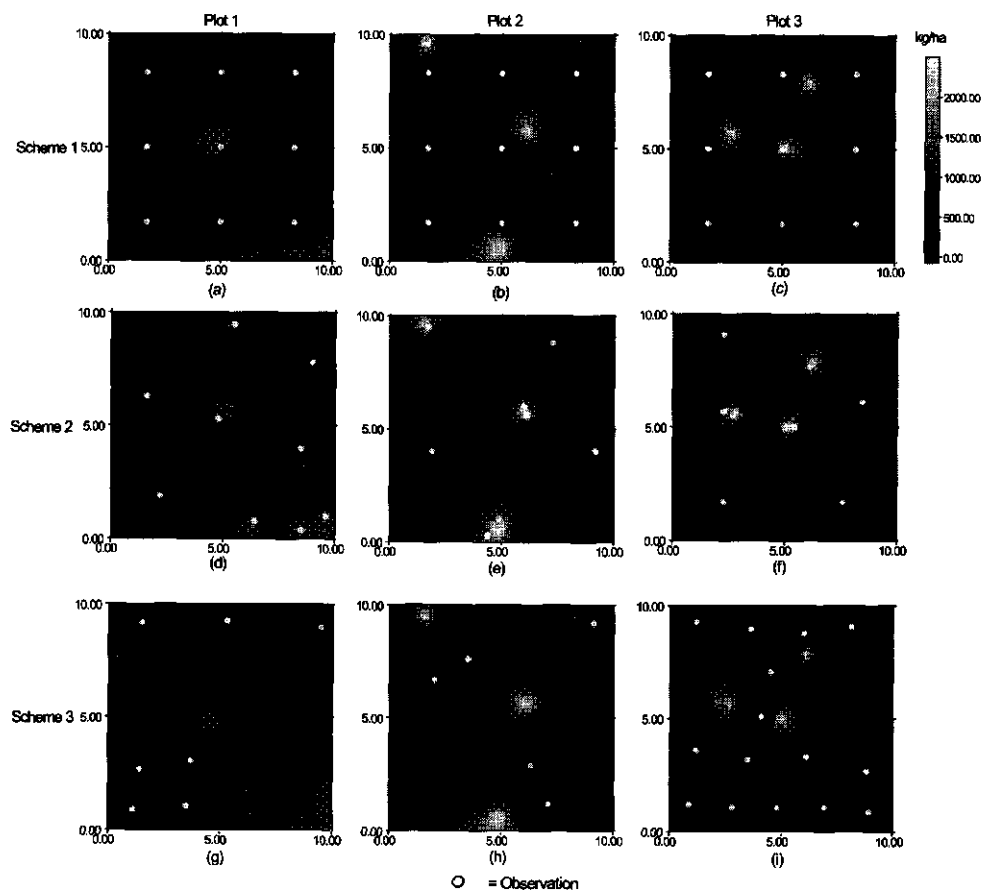


Figure 7.4. *The three sampling schemes for the three plots, superimposed on the yield estimates*

Figure 7.4g-i shows the resulting sampling scheme. As the predicted yield in plot 3 is considerably lower than in plots 1 and 2, most of the observations (15) are taken there. Descriptive statistics are presented in Table 7.2. pH-H₂O is lower at all depths, as compared to S₁. Furthermore, texture of S₃ is slightly heavier, organic matter content is lower and distance to shrubs is higher.

7.3.3. Correlation and regression analysis

To test the performances of S₁ and S₂ in assessing soil-yield relations, we calculated correlation coefficients between yield and soil variables. Yield at sampling locations

Table 7.3. Correlation coefficients for millet yield and soil variables.

Depth m	Relief	Distance m	pH-H ₂ O	CEC cmol kg ⁻¹	Sand	Silt %	Clay	OM g kg ⁻¹	P-Bray mg kg ⁻¹
<u>Sampling Scheme 1</u>									
0-0.1	0.258	-0.464*	-0.160	-0.086	0.047	-0.060	-0.031	0.190	0.160
0.1-0.2	-	-	0.054	0.007	-0.083	0.214	0.028	-0.028	-0.221
0.2-0.4	-	-	0.081	0.004	-0.190	0.069	0.188	-0.112	-0.293
<u>Sampling Scheme 2</u>									
0-0.1	0.328	-0.451*	0.200	0.239	-0.141	-0.128	0.255	0.361	0.264
0.1-0.2	-	-	0.141	0.005	-0.189	0.342*	0.095	0.557**	0.217
0.2-0.4	-	-	0.052	0.184	-0.215	0.097	0.208	0.080	0.102

Levels of significance: * = 0.10, * = 0.05, ** = 0.01

was predicted using ordinary kriging. For S_1 , only distance to shrubs is significant with $\alpha = 0.05$, and no other variables show a significant correlation with yield (Table 7.3). The negative correlation of yield with distance to shrubs can be explained by the relative higher fertility (in terms of OM and P-Bray) closer to the shrubs.

A stepwise linear regression was performed to find significant relations for S_1 and S_2 . For yield, the Lilliefors test for normality was not rejected (with $\alpha = 0.05$). The selected model for S_1 was:

$$\text{Yield} = 789.6 - 137.1 * \text{Dist} + 253.2 * \text{Silt}[0.1 - 0.2] - 452.1 * \text{PBray}[0.2 - 0.4] \quad (7.4)$$

where $\text{Silt}[0.1-0.2]$ denotes the silt content in the second layer. This significant model (with $\alpha = 0.05$) yielded an adjusted r^2 of 0.372. These results are only slightly better than those from Stein *et al.* (1997), who found an unadjusted r^2 of 0.377 in a much less detailed survey.

For S_2 , one correlation coefficient was highly significant ($\alpha = 0.01$), another one significant at $\alpha = 0.05$, and three more significant with $\alpha = 0.10$. The significant correlation coefficient for distance to shrubs confirms the findings of S_1 . Micro-relief and OM were positively correlated to yield. This can probably be explained by the higher relief around shrubs, due to catchment of eroded material. In the second layer silt was significantly correlated, and OM highly significant.

The selected regression model for S_2 was:

$$\text{Yield} = 18728 + 12702 * \text{Relief} - 257.9 * \text{Silt}[0.1 - 0.2] - 135.9 * \text{PBray}[0.1 - 0.2] + 9239 * \text{OM}[0.1 - 0.2] - 219.6 * \text{Sand}[0.1 - 0.2] \quad (7.5)$$

Table 7.4. Results from a student t-test between S_1 and S_3 . Probabilities that both populations are similar.

Depth	Relief	Distance	pH-H ₂ O	CEC	Sand	Silt	Clay	OM	P-Bray
0-0.1	0.015*	0.070*	0.002**	0.009**	0.168	0.616	0.127	0.119	0.545
0.1-0.2	-	-	0.003**	0.612	0.947	0.960	0.923	0.267	0.235
0.2-0.4	-	-	0.053*	0.749	0.529	0.496	0.468	0.051*	0.299

Levels of significance: * = 0.10, * = 0.05, ** = 0.01

This model yielded an adjusted r^2 of 0.705, which is much higher than that obtained by S_1 . It may be noticed that in this model distance to shrubs is not included, despite its significant correlation coefficient (Table 7.3). This might be explained by its highly significant correlation to micro-relief. In fact, univariate regression of yield gives a highly significant model for distance to shrubs:

$$\text{Yield} = 1704.6 - 218.8 * \text{Dist} \quad (7.6)$$

Although this model is highly significant, it only explains 25 % of the yield variability. Therefore, the multivariate regression was preferred.

7.3.4. Student t-test

A Student t-test was performed to show significant differences in soil variables between S_1 and S_3 . Table 7.4 shows the probabilities that the means of the two populations are similar. Three variables showed highly significant differences, whereas four additional ones showed significant differences. The significance of distance to shrubs and relief coincide with findings of the regression analysis. The CEC[0-0.1] is significantly higher for S_3 . This may be related to the relative enrichment with finer particles due to erosion at the low-producing areas. Finally, pH-H₂O is highly significant different at all depths, reflecting the general depletion of the low producing area.

7.4. Discussion and Conclusions

Using the SSA algorithm, three different sampling schemes of limited size were constructed. Scheme S_1 , which aimed at optimal spreading of the sampling points over the area, yielded a significant (negative) correlation of distance to shrubs with yield. Scheme S_2 , which aimed at coverage of all the extremes in yield, showed significant correlations for four additional variables. Moreover, using multivariate regression analyses the explained yield variation increased from 37% using S_1 , to 70% using S_2 .

We concluded therefore that the approach taken for constructing S_2 should be preferred for relating yield to soil variables in case of limited observations.

Differences between S_3 , which aimed specifically at low-producing areas, and S_1 were used to detect the main limiting variables for yield. Distance to shrubs and micro-relief were probably the most important factors in this. Other significant variables (higher CEC and lower pH- H_2O in S_3) were probably related to the effects of soil erosion, due to absence of shrubs.

These findings are in line with existing management techniques practised by local farmers. Shrubs are seen as a valuable asset, useful in locally improving soil fertility by catchment of airborne eroded particles, and by enrichment with organic matter from the shrub. At the start of the growing season, shrubs are trimmed to limit competition with the millet crops. Since rainfall at the start of the 1997 growing season was relatively good, this is also in line with the yield stabilising effect of soil variability, as reported by Brouwer (1996). In drier years, the yield pattern may be reversed. However, in order to optimise the yield stabilising effect of the shrubs, planting of shrubs should be considered as a low-tech soil management practice. More research should be dedicated to optimal placement of such shrubs.

The sampling schemes used in this study were deliberately kept small in order to stay close to financial constraints generally met in precision agriculture research. Although costs of laboratory analysis is still out of reach for the marginal farmers in the study area, we think that the developed techniques can be of great value to researchers dealing with the highly variable soils of the Sahelian zone. Moreover, the type of auxiliary data used to optimise sampling is typical for precision agriculture, and developed techniques may be applied in other types of precision agriculture. As an example, remote sensed data related to crop yield might be treated in the same way as the predicted yield maps in this study.

Since the number of observations was small, relatively simple statistical techniques were used to relate yield to soil variables. In our opinion, this represents the practical constraints of precision agriculture rather well. If a higher number of observations can be collected, geostatistics can be applied for interpolation of soil variables. Auxiliary data like yield maps may then serve as co-variable in cokriging.

7.5. Acknowledgements

The authors are very grateful to Joost Brouwer and Alfred Stein for their advice on many aspects of this study. Furthermore, we would like to thank Niek van Duivenbooden and Charles Biolders for their assistance at ICRISAT Sahelian Centre, and Tjerk Kuster for his part in the scoring. Finally, we are very grateful to Hassan Ousmane for assistance during fieldwork.

Chapter 8

Sampling Strategies for Effective Variogram Estimation¹

Abstract

This chapter compares various sampling strategies for variogram estimation. Sampling strategies are tested using stochastic simulations. In the first part, a regular grid is compared to a sampling scheme that optimises the point pair distribution for variogram estimation. This yielded unbiased experimental variograms. However, the fluctuation of the experimental variograms was much lower with a regular grid. We concluded from this that the point pair distribution alone is not a useful optimisation criterion for variogram estimation. In the second part, additional observations selected for optimal point pair distribution were compared with randomly drawn additional observations. The random observations resulted in much higher standard deviations at shorter distances. We concluded from this that for additional short distance observations the point pair distribution is a very useful optimisation criterion. In the third part, we focused on optimal variogram use. A sampling grid of 81 observations was completed after preliminary estimation of the variogram with 19 additional observations for minimal kriging variance. The scheme was compared to a regular grid of 100 observations. For an exponential underlying variogram without nugget effect, the use of the phased sampling scheme reduced the mean squared kriging error from 0.39 [unit]² to 0.31 [unit]², and the maximum squared kriging error from 6.05 [unit]² to 4.24 [unit]². For a spherical underlying variogram with a nugget effect of 33%, mean squared kriging error did not change and maximum squared kriging error decreased from 15.98 [unit]² to 11.52 [unit]². We concluded that minimisation of the squared kriging error is often more relevant than accurate estimation of the variogram. Taking samples just outside the area improved the quality of the prediction in terms of both kriging variance and actual squared kriging error.

¹Based on: Van Groenigen, J.W., Mainam, F. and Stein, A. (submitted). *Sampling strategies for effective variogram estimation*. European Journal of Soil Science.

8.1. Introduction

Establishing the nature of spatial correlation is often the starting point for a geostatistical study. Information on spatial correlation is needed in many geostatistical algorithms such as kriging and Multi Gaussian simulation. Other algorithms such as indicator kriging and cokriging require information on spatial (cross-)correlation of more than one variable. This spatial correlation is expressed in the variogram or some related statistic such as the co-variogram or cross-variogram.

The isotropic (direction-independent) variogram is defined as:

$$\gamma(\mathbf{h}) = \frac{1}{2} \mathbf{E} \left[\{Z(\bar{\mathbf{x}}) - Z(\bar{\mathbf{x}} + \mathbf{h})\}^2 \right] \quad (8.1)$$

where $Z(\bar{\mathbf{x}})$ denotes the value of regionalized variable (ReV) Z at location $\bar{\mathbf{x}}$. This variogram can be estimated from observations $z(\bar{\mathbf{x}})$:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2n(\mathbf{h})} \sum_{i=1}^{n(\mathbf{h})} \{z(\bar{\mathbf{x}}_i) - z(\bar{\mathbf{x}}_i + \mathbf{h})\}^2 \quad (8.2)$$

where $n(\mathbf{h})$ is the number of point pairs separated by distance \mathbf{h} (plus or minus an interval). Normally, the experimental variogram is calculated for nc lag classes to which subsequently a variogram model is fitted.

In geostatistics, a distinction is usually made between two variograms:

- i) *The variogram of the underlying stochastic model*, of which the 'reality' that is observed and sampled by the soil surveyor is considered to be just one realisation. This is called the *underlying* (Webster and Oliver, 1992) or *ergodic* (Brus and De Gruijter, 1994) variogram. In this chapter we will use the term underlying variogram. The underlying variogram is reproduced over a large number of realisations.
- ii) *The variogram of the realisation* that is actually observed in practice, which can deviate from the underlying variogram. This variogram is called the *local* (Webster and Oliver, 1992) or *non-ergodic* (Brus and De Gruijter, 1994) variogram. We will use the term local variogram in this chapter.

In the literature, optimisation of spatial sampling for variogram estimation has derived some attention. However, no consensus has been reached on the criterion to be used to assess the quality of sampling schemes for variogram estimation.

In general, three approaches towards optimising sampling for variogram estimation

can be distinguished:

- i) *Optimising experimental variogram estimation.* If nothing is known on the spatial distribution of a ReV, all that can be manipulated is the location of the sampling points. This has led to formulation of optimisation criteria for the optimal point pair distribution of a sampling scheme. This was first done by Warrick and Myers (1987), who presented a Monte Carlo algorithm for optimisation of the point pair distribution. Russo and Jury (1988) tested a slightly modified version of this procedure for estimation of the covariogram on several realisations of a stochastic field. They concluded that the covariogram was more accurately estimated at shorter distances, as compared to a regular grid. Van Groenigen *et al.* (1999) (Chapter 4) used this optimisation criterion to improve their sampling scheme for estimation of spatial correlation at short distances. Brus and De Gruijter (1994) modified this criterion for optimisation of the estimation accuracy of the local variogram. Taylor and Burrough (1986) and McBratney and Webster (1986a) used stochastic simulations to assess the quality of the experimental variogram. Webster and Oliver (1992) and Shafer and Varljen (1990) did the same using multiple samples from a single realisation of the stochastic field.
- ii) *Optimising the fitting of the experimental variogram.* When the structure of spatial correlation is known, sampling can be optimised for estimation of the variogram (D-optimality). Procedures for this have been developed and applied by Rasch (1990), Pettitt and McBratney (1993), Zimmerman and Homer (1991) and Van Groenigen and Stein (1998) (Chapter 3). Morris (1991) estimated the accuracy of the variogram by assuming a variogram model.
- iii) *Optimising the effectiveness of the variogram.* This approach does not look at the accuracy of the (modelled) variogram itself, but on its use in further geostatistical analysis. Russo and Jury (1988) compared the resulting kriging variance (KV) of their optimisation procedure for variogram estimation with that of a regular sampling grid. However, this seems an awkward criterion since the KV is relative to the variogram, and therefore does not represent any errors in variogram estimation (Deutsch and Journel, 1998). Gascuel-Oudou and Boivin (1994), by resampling from a large data set, found that the actual squared kriging error (SKE) was lower than the calculated KV. Laslett *et al.* (1987) and Bregt *et al.* (1991) reported much higher SKE's than the calculated kriging variances. Brus (1993) showed how the sampling error of the local variogram could be included in the kriging error.

As the variogram is generally not known in practice (and does not have to be estimated if it is), the second group of optimisation criteria is mainly of theoretical interest. Therefore, this chapter focuses on the first and third group. Using the SSA-algorithm for optimisation of spatial sampling schemes, the merits of these optimisation

criteria for variogram estimation will be examined. Special emphasis will be placed on the practical value of the developed sampling schemes.

Webster and Oliver (1992) tested the effectiveness of sampling grids of different sizes for estimation of the experimental variogram. Since they resampled a single realisation of a stochastic field using different origins of the sampling grid, their results are valid for the local variogram. In this study, we test sampling strategies for estimation of the underlying variogram, and therefore we use more than one realisation of the stochastic fields for two of the examples. The last example will focus on effective characterisation of one realisation.

8.2. Materials and methods

8.2.1. Spatial Simulated Annealing

Spatial Simulated Annealing (SSA) was developed as a general purpose optimisation method for sampling in geostatistical studies (Van Groenigen and Stein, 1998) (Chapter 3). Its features include optimisation of sampling schemes at the point level, incorporation of preliminary observations and honouring of sampling constraints. Starting with a random initial sampling scheme, the optimisation algorithm generates random distortions in the sampling scheme. The quality of the new scheme is assessed using a quantitative optimisation criterion. Improvements in the sampling scheme are accepted, while deteriorations are sometimes (and with decreasing probability as the optimisation advances) accepted to avoid local minima. The SSA algorithm allows to handle different quantitative optimisation criteria. Those used in this study are briefly presented below. For a detailed discussion of SSA, see Van Groenigen and Stein (1998) (Chapter 3).

8.2.2. The WM-criterion

The WM-criterion, formulated by Warrick and Myers (1987), optimises the fit of the realised point pair distribution of the sampling scheme to an *a priori* defined, ideal distribution. The criterion is formulated as the sum of squared differences between the realised and the ideal point pair distribution:

$$\sum_{i=1}^{nc} (\zeta_i^* - \zeta_i)^2 \quad (8.3)$$

where nc is the number of lag classes of the variogram, and ζ_i and ζ_i^* are the realised and ideal number of point pairs in the i^{th} lag class, respectively. In this chapter, the ideal distribution is defined as even distribution of the point pairs over the chosen variogram lags, i.e. $\zeta_i^* = \zeta_j^*, \forall i, j \in 1, \dots, nc$.

8.2.3. The MAX_OK-criterion

The MAX_OK criterion minimises the maximum ordinary kriging variance of the sampling scheme, using a variogram model as input. Using a fine raster of evaluation points $\bar{x}_{e,1}, \dots, \bar{x}_{e,nc}$ over the whole research area, the criterion is formulated as:

$$\max(\sigma_{OK}^2(\bar{x}_{e,j} | S), \quad \forall \bar{x}_{e,1}, \dots, \bar{x}_{e,nc}) \quad (8.4)$$

where $\sigma_{OK}^2(\bar{x}_{e,j} | S)$ denotes the ordinary kriging variance at the j^{th} evaluation point $\bar{x}_{e,j}$ using sampling scheme S . Van Groenigen (*submitted*) (Chapter 5) showed that sampling schemes optimised using this criterion are affected considerably by the choice of the variogram parameters and (to a lesser extent) the kriging neighbourhood.

8.2.4. Stochastic simulation

The accuracy of the sampling schemes will be assessed below using stochastic simulation. The advantage of using stochastic simulations instead of actual measurements is two-fold:

- i) The 'reality' is fully known. Therefore, we are able to quantify actual SKE's without constructions like jack-knifing.
- ii) Multiple realisations of the stochastic field can be generated. Therefore, both the underlying variogram and the local variogram can be calculated. In reality, only the local variogram can be known (Journel, 1985; Brus and De Gruijter, 1993).

In this study, we used the Multi Gaussian simulation algorithm to generate the stochastic fields (Deutsch and Journel, 1998). This algorithm was chosen for its theoretical transparency.

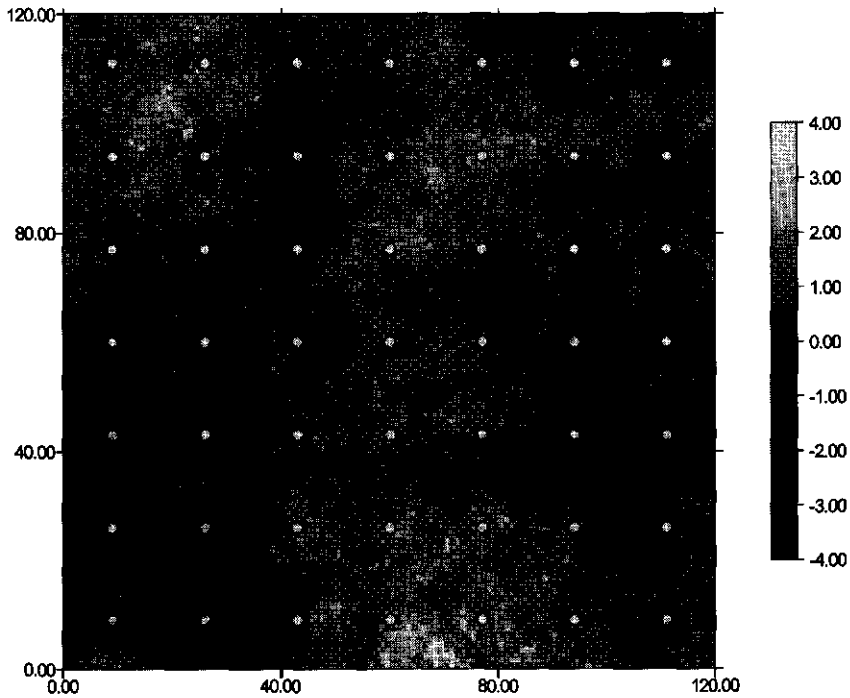


Figure 8.1. *Simulated field using the spherical underlying variogram, with a regular sampling grid of 49 observations*

8.3. Sampling schemes

Below, several sampling strategies are tested for their value in variogram estimations using stochastic simulations.

8.3.1. Regular grid vs. WM

In order to test the usefulness of the WM-criterion for variogram estimation, 40 different realisations of two underlying variogram models as used by Webster and Oliver (1992) were constructed using Multi Gaussian simulation (Deutsch and Journel, 1998). The first model is defined by a spherical global variogram:

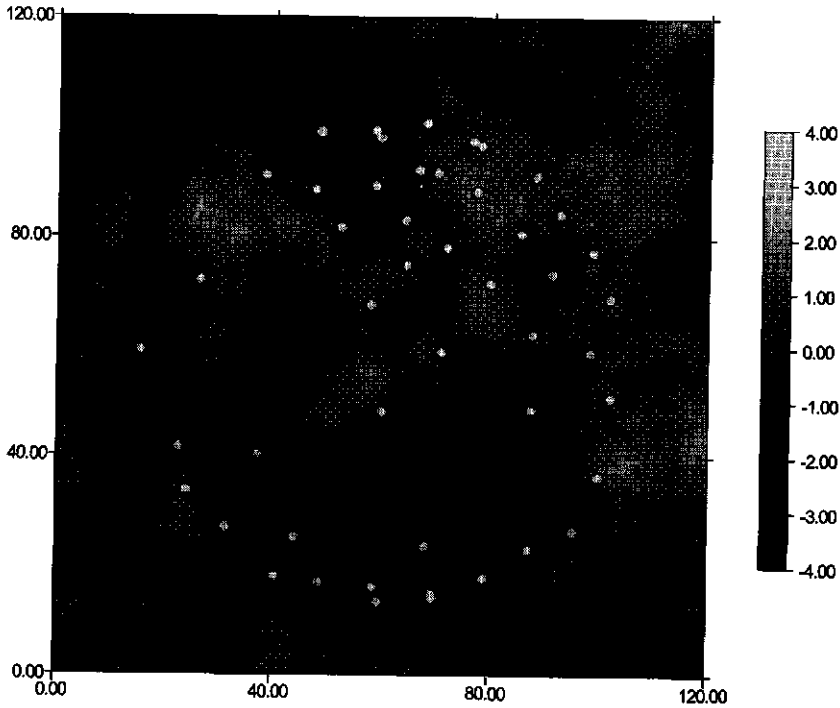


Figure 8.2. Simulated field using the exponential underlying variogram, with a WM-optimised sampling scheme of 49 observations.

$$\gamma(\mathbf{h}) = 0.333 + 0.667 \left\{ \frac{3}{2} \cdot \left(\frac{\mathbf{h}}{50} \right) - \frac{1}{2} \cdot \left(\frac{\mathbf{h}}{50} \right)^3 \right\} \text{ for } 0 < \mathbf{h} \leq \mathbf{a},$$

$$\gamma(\mathbf{h}) = 1 \quad \text{for } \mathbf{h} > \mathbf{a}, \quad (8.5)$$

$$\gamma(0) = 0.$$

This represents a variogram with nugget = 0.333, sill = 1 and effective range = 50. The second underlying variogram model is exponential and has similar sill and (effective) range parameters, but no nugget effect:

$$\gamma(\mathbf{h}) = 1 - \exp\left(-\frac{\mathbf{h}}{16}\right) \quad (8.6)$$

Table 8.1. Parameters of the sampling schemes and the experimental variograms, with sample size (n) and number of lag classes (nc).

Grid		Variogram	
n	Spacing	nc	Lag size
25	20	4	20
36	18	5	16
49	17	5	16
64	15	6	12
81	13	7	10
100	12	8	10

Figures 8.1 and 8.2 show one realisation of the spherical and exponential underlying variograms, respectively. Both fields consist of $120 \times 120 = 14400$ grid cells. All realisations had variances close to 1, average values close to 0 and no significant skewness. Because of the absence of a nugget effect, the exponential field is much smoother than the spherical one.

All 80 realisations were sampled using both a regular square grid and a WM-optimised scheme for different sample sizes. Table 8.1 lists all sample sizes and the corresponding number of lags and lag sizes for the experimental variogram. Figure 8.1 shows the grid scheme for a sample size of 49. Distances to the boundaries of the area are half the grid spacing. Figure 8.2 shows the WM-optimised scheme for the same sample size. Noteworthy are the nearly circular structure of the sampling scheme, and the absence of observations from large parts of the area (such as the corners).

The performances of the sampling schemes for the estimation of the underlying variogram were investigated. Calculations for all 80 simulated fields and all 12 sampling schemes resulted in 960 experimental variograms. Statistics of the variation in these variograms assessed the performances of the sampling schemes.

8.3.2. Combining regular grid and WM

In many geostatistical surveys based upon a regular sampling grid, surveyors may be interested in spatial correlation at shorter distances than the grid spacing. This may be because no apparent spatial correlation is detected using the grid, suggesting that spatial correlation is confined to shorter distances (e.g. Oliver and Webster, 1987). In other studies, more information of the behaviour of the variogram near the origin might be necessary for optimal interpolation (the SKE and KV are considerably influenced by the modelled nugget effect). Goovaerts (1997a) randomly selected additional observations at fixed distances of grid nodes. McBratney and Webster (1986b) followed a nested strategy. Van Groenigen and Stein (1998) (Chapter 3) showed how

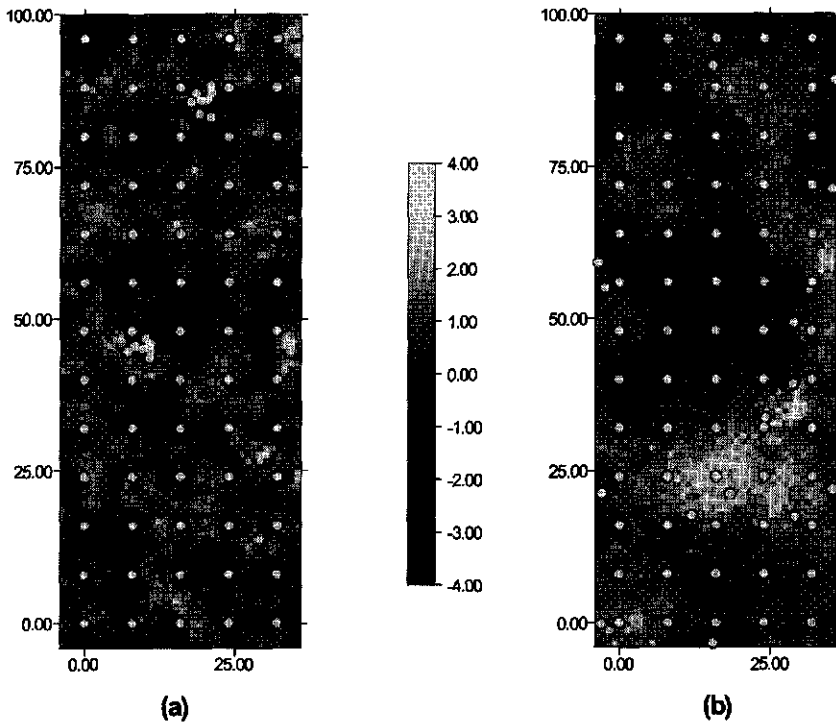


Figure 8.3. Two simulated fields and the sampling grid for the Cameroon study. Additional samples optimised using SSA were drawn from a realisation of a spherical underlying variogram with a range of 5 m (a), and randomly drawn from a realisation of a spherical underlying variogram with a range of 20 m (b).

SSA with the WM-criterion could be used for upgrading existing sampling grids for variogram estimation at short distances. The usefulness of such a sampling scheme for variogram estimation will be investigated below.

Figure 8.3 shows the sampling plot of a geostatistical erosion study conducted in the Northern parts of Cameroon (Mainam and Zinck, 1998). To assess the susceptibility of different soils to water erosion, sampling plots were chosen along two different slopes. The size of the plots is 40 m x 104 m, with the longer part in the direction of steepest descent. To get an overview of the general degree of erosion over the plot, an initial sampling grid with 8 m spacing was decided upon. This yielded an initial sample size of 65. Fifteen additional observations were selected for estimation of the short range spatial variation using SSA with the WM-criterion. The performance of the proposed sampling scheme was investigated using stochastic simulations before actual sampling took place.

Two different underlying variogram models were selected for stochastic simulation.

Both are spherical models with nugget = 0.25 and sill = 1.0. The first model has a range of 5 m (shorter than the lag spacing), and the second model a range of 20 m. For both models, 20 realisations were generated. Figures 8.3a and 8.3b show realisations of the first and second model, respectively.

SSA with the WM-criterion aimed at distributing the additional point pairs evenly over two lag classes of 1-3 m and 3-6 m, respectively. To reduce the risk of failure of this additional sampling scheme due to local anomalies such as rock outcrops *etc.*, the extra points were divided into two different groups of 7 and 8 observations. These two groups were placed independently from each other. This resulted in two clusters of points in different parts of the area (Figure 8.3a). The clustering of additional points using the WM-criterion corresponds with the findings of earlier studies (Velthof *et al.*, submitted; Van Groenigen *et al.*, 1999) (Chapter 4). As a comparison, a second sampling scheme distributed the additional observations randomly over the area. Figure 8.3b shows this scheme.

Experimental variograms were calculated for both sampling schemes and for all 40 realisations. Statistics of these variograms assessed the quality of the WM-optimised additional sampling scheme for variogram estimation.

8.3.3. Variogram use: minimising the prediction error

The third optimisation method deals with the effectiveness of the estimated variogram, rather than with its accuracy. Following Gascuel-Oudoux and Boivin (1994) and Stein and Corsten (1991), the SKE of the interpolation based upon the estimated variogram was used to assess the performances of the sampling schemes.

Optimisation strategy and evaluation criteria were chosen with two sampling issues in mind:

- i) The variogram may be estimated during a first survey stage, whereas a second stage may be conducted for purposes of interpolation. In practice, these stages are seldom totally independent, *i.e.* the data from the second stage is often used to improve the estimation of the variogram. Conversely, data from the first stage is often used during interpolation.
- ii) Kriging variance, providing an indication of the SKE, depends upon the chosen variogram model. Therefore, errors from variogram estimation and modelling are not accounted for. In addition, validity of the kriging variance as a measure of prediction accuracy depends on the intrinsic hypothesis of geostatistics. In case of non-stationarity and/or heteroscedasticity, the kriging variance loses part of its value. In those cases the SKE is a more direct measure of accuracy. Unfortunately, SKE can only be assessed *a posteriori*, and can therefore not serve as a direct

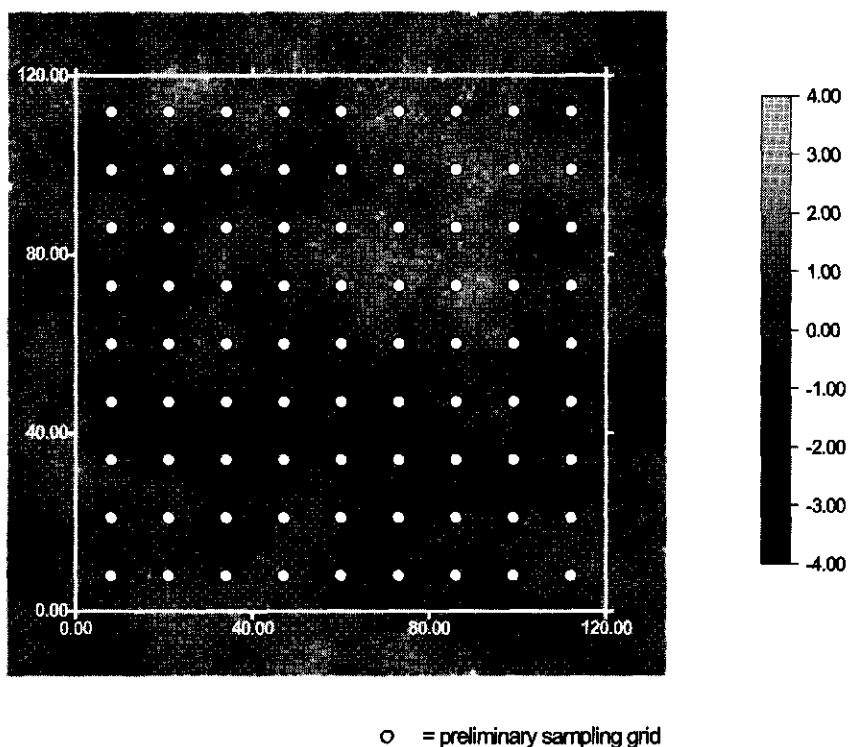


Figure 8.4. Preliminary sampling grid on the spherical simulated field.

optimisation criterion.

As a data source, one realisation of each of the two underlying variograms presented by Webster and Oliver (1992) and presented earlier in this chapter (Figures 8.1 and 8.2) were randomly drawn. The total sample size was equal to 100 to allow a comparison with a regular grid (10 x 10). A two-stage sampling strategy was applied, where the first stage approximately estimated the variogram. Using the preliminary variogram we then optimised allocation of additional observations for minimal kriging variance with SSA. More formally, if the total sample size is denoted n , the sampling strategy and the resulting interpolation are:

- i) The sampling scheme of the first sampling stage S^1 consists of n_1 observations, chosen on a regular grid covering the whole area, with $n_1 < n$. The data of S^1 are collected.
- ii) Using the data of S^1 we calculated a preliminary experimental variogram $\hat{\gamma}^1(h)$ and

Table 8.2. Standard deviations of experimental variograms per lag for grid schemes and WM-optimised schemes, calculated over 40 realisations of two underlying variograms.

n		Lag number							
		1	2	3	4	5	6	7	8
<i>Exponential model</i>									
25	grid	0.20	0.36	0.41	0.44	-	-	-	-
	wm	0.18	0.28	0.40	0.64	-	-	-	-
36	grid	0.15	0.24	0.35	0.41	0.50	-	-	-
	wm	0.14	0.29	0.42	0.39	0.46	-	-	-
49	grid	0.14	0.23	0.30	0.35	0.38	-	-	-
	wm	0.17	0.28	0.33	0.38	0.46	-	-	-
64	grid	0.11	0.14	0.20	0.28	0.33	0.34	-	-
	wm	0.12	0.23	0.29	0.41	0.53	0.64	-	-
81	grid	0.07	0.11	0.17	0.23	0.28	0.30	0.33	-
	wm	0.10	0.23	0.30	0.36	0.49	0.59	0.74	-
100	grid	0.08	0.12	0.18	0.22	0.27	0.31	0.33	0.33
	wm	0.09	0.17	0.26	0.35	0.47	0.61	0.70	0.62
<i>Spherical model</i>									
25	grid	0.22	0.32	0.33	0.35	-	-	-	-
	wm	0.24	0.30	0.38	0.31	-	-	-	-
36	grid	0.15	0.22	0.30	0.32	0.29	-	-	-
	wm	0.19	0.26	0.33	0.37	0.43	-	-	-
49	grid	0.12	0.20	0.24	0.27	0.30	-	-	-
	wm	0.15	0.20	0.30	0.30	0.33	-	-	-
64	grid	0.11	0.16	0.17	0.22	0.26	0.28	-	-
	wm	0.15	0.19	0.24	0.32	0.30	0.34	-	-
81	grid	0.09	0.10	0.13	0.18	0.21	0.23	0.23	-
	wm	0.12	0.20	0.28	0.34	0.36	0.39	0.49	-
100	grid	0.08	0.12	0.13	0.15	0.18	0.19	0.22	0.23
	wm	0.11	0.17	0.21	0.28	0.30	0.31	0.33	0.33

fitted it to a variogram model $\gamma^1(\mathbf{h})$ using a weighted least squares regression procedure.

- iii) An additional sampling scheme S^2 of sample size $n_2 = n - n_1$ is obtained by optimising for minimal kriging variance. This is done by SSA with the MAX_OK criterion, using $\gamma^1(\mathbf{h})$ to fill the variance matrices in the kriging equations. The data for S^2 are collected.
- iv) A new experimental variogram $\hat{\gamma}^2(\mathbf{h})$ is calculated from $S^1 \cup S^2$ and a new model $\gamma^2(\mathbf{h})$ is fitted.
- v) Using $\gamma^2(\mathbf{h})$ and the data from $S^1 \cup S^2$, an interpolated surface is constructed using ordinary kriging.

To ensure a reasonable estimation of $\gamma^1(\mathbf{h})$, n_1 should be sufficiently large. In this study with the simulated field of 120 x 120 unit coordinates we took $n_1 = 81$ samples on a 9 x 9 grid, with a grid spacing of 13 units. Earlier studies showed that

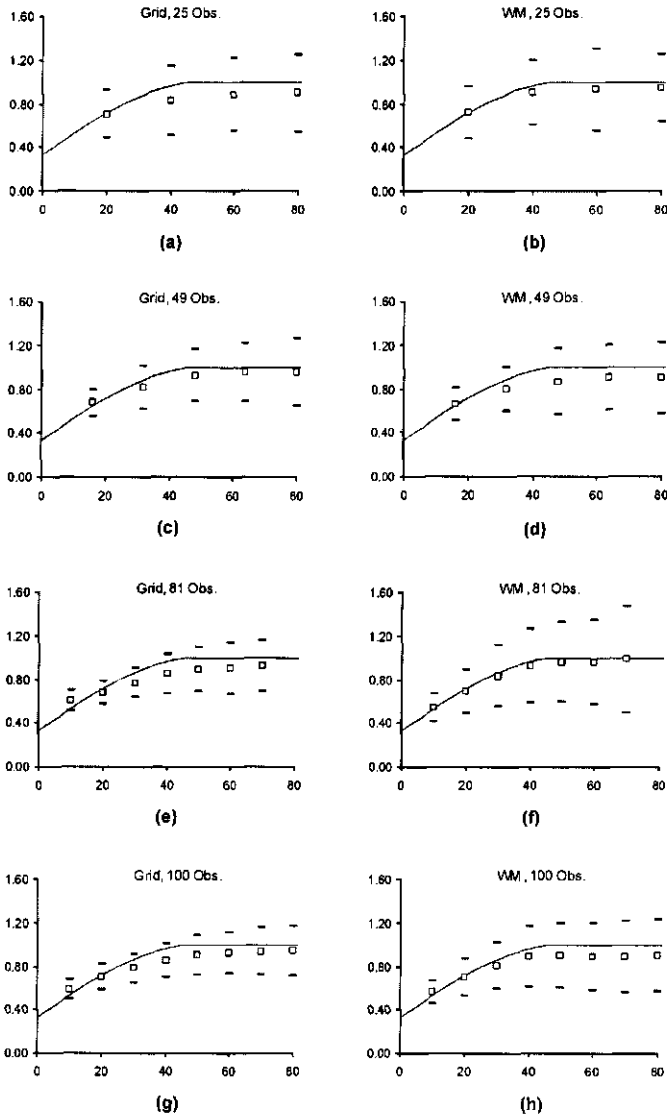


Figure 8.5. Comparison of a regular grid with the WM-criterion for the spherical model. Experimental variograms with standard deviations and the underlying variogram for several sampling sizes over 40 realisations.

optimisation of sampling schemes for minimal KV can involve samples close to or even over the boundaries of the area (Van Groenigen *et al.*, 1999) (Chapter 4). Therefore we also considered a buffer area of 15 units outside the actual research area. Figure 8.4 shows the simulated field with buffer area for the spherical underlying variogram with

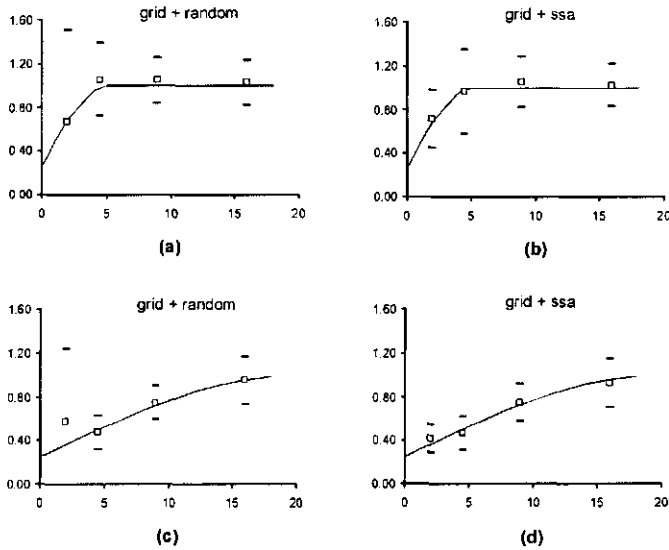


Figure 8.6. Comparison of a regular grid with the WM-criterion for the two spherical models in the Cameroon study over 20 realisations. Experimental variograms with standard deviations and the global variogram for a short range (a,b) and a long range (c,d).

nugget = 0.333, sill = 0.667 and range = 50 units. Although the mean and maximum SKE's were only calculated for the 120 x 120 unit research area, additional observations were allowed to be placed outside of this area.

For comparison, a second sampling scheme of sample size 100 was set up using a 10 x 10 square grid with a grid spacing of 12 units. Both stochastic fields were sampled using both sampling strategies. Subsequently, interpolated surfaces were predicted using ordinary kriging, and the mean and maximum SKE's were calculated.

8.4. Results

8.4.1. Regular grid vs. WM-criterion

Table 8.2 summarises the variogram calculations for all 80 realisations of the underlying variograms. Since the bias was close to zero (*i.e.* the underlying variogram was reproduced over all the realisations), only the standard deviations of the experimental variograms are presented for each lag class. The WM-optimised sampling scheme scores occasionally better than the sampling grid (notably with very small sample

Table 8.3. Parameters of the underlying variogram and the fitted variogram models for a sampling grid of 100 observations and a two stage sampling scheme with 81 and 100 observations, respectively.

		Variogram parameters		
Model		Nugget	Sill/slope	(Eff.) range
<i>Exponential</i>				
<i>Underlying</i>	<i>exp.</i>	0.000	1.000	50
<i>Grid</i>	<i>exp.</i>	0.396	0.804	57.6
<i>Grid + SSA 1</i>	<i>sph.</i>	0.372	0.720	58.2
<i>Grid + SSA 2</i>	<i>sph.</i>	0.060	0.696	30.6
<i>Spherical</i>				
<i>Underlying</i>	<i>sph.</i>	0.333	1.000	50
<i>Grid</i>	<i>lin.</i>	0.408	0.016	-
<i>Grid + SSA 1</i>	<i>lin.</i>	0.432	0.016	-
<i>Grid + SSA 2</i>	<i>lin.</i>	0.504	0.013	-

sizes). However, the sampling grid is superior in terms of variogram fluctuation, in particular with increasing sampling density.

Figure 8.5 shows some of these results graphically for the spherical underlying variogram and several sample sizes. The continuous line represents the underlying variogram over the 40 realisations of the spherical model, showing that bias (distance between line and squares) is close to zero. The standard deviations are much larger for the WM-optimised schemes, occasionally reaching more than twice the value of those for the sampling grid.

8.4.2. Combining regular grid and WM

The variograms for the 40 realisations of the two underlying variograms are shown in Figure 8.6. Since the width of the sampling area was only 40 m, no variogram values were calculated for distances exceeding 20 m, to avoid bias. Four lags were evaluated. Following the WM-optimisation, the centres of the two smallest lags were 2 m and 4.5 m. The two largest lags centred on once and twice the grid spacing, 8 m and 16 m.

Figure 8.6 shows that the standard deviation of the shortest lag (1-3 m) is much higher for the grid sampling scheme, although at larger distances the experimental variograms are very similar. The standard deviations for this scheme are 0.84 and 0.67 for the first and second spherical model. With the WM-optimised sampling schemes this reduces to 0.26 and 0.12, respectively. This improvement was achieved by specifically optimising the additional observations for the used variogram lags. There is no significant bias for both of the sampling schemes.

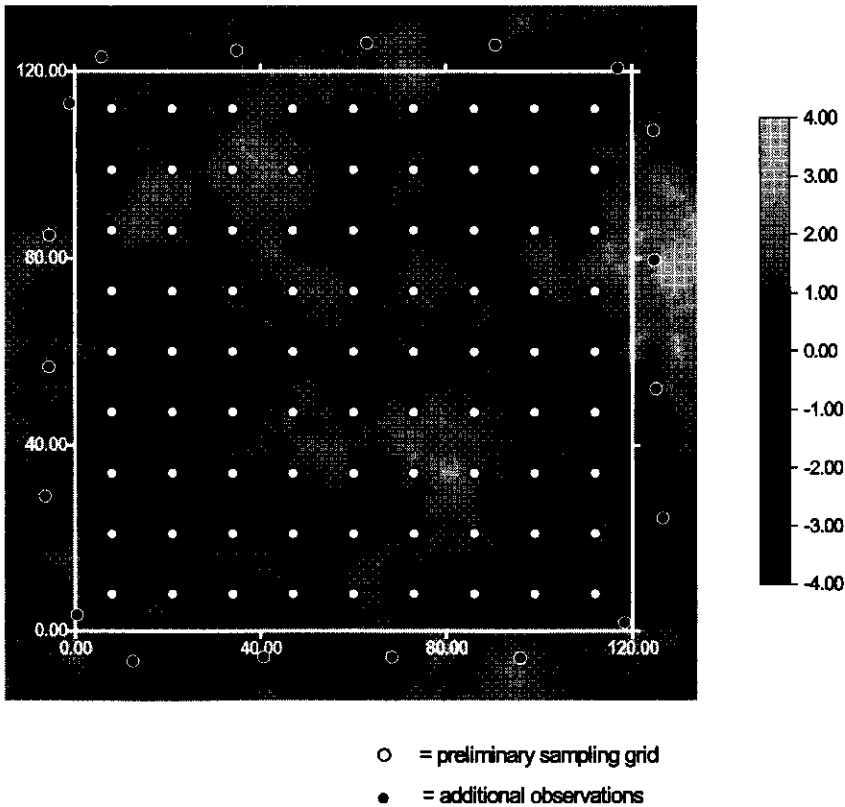


Figure 8.7. Optimisation of 19 additional sampling points for minimal kriging variance to a grid of 81 points on the exponential field. The additional sampling points are allowed outside the research area of 120 x 120 units.

8.4.3. Minimising the squared kriging error

Table 8.3 lists the underlying variogram models of the two stochastic fields, together with the modelled experimental variograms of the two sampling grids. The modelled variograms poorly relate to the underlying variogram. For the exponential stochastic field, both the 10 x 10 sampling grid and the first stage of the two-stage sampling scheme model a nugget effect of 0.396 and 0.372, respectively, while no nugget effect is present in the underlying variogram. For the spherical stochastic field, both sampling schemes overestimate the nugget effect. In addition, the experimental variogram for the exponential stochastic field is modelled as spherical for the first stage of the two-stage scheme. For the spherical stochastic field the variogram for both schemes is modelled as linear. This divergence is partly caused by the sampling scheme, as it is still

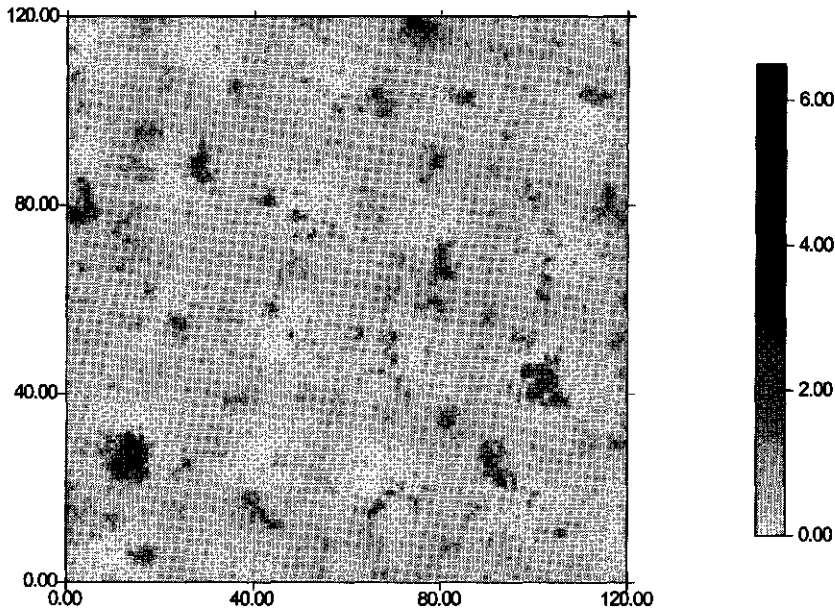


Figure 8.8. Squared kriging error (in unit²) for the two-stage sampling scheme on the exponential field, number of observations is 100.

inadequate for precisely estimating the variogram. However, since only one realisation of the stochastic model is sampled also differences between the underlying and local variogram play a role.

Using these modelled variograms, we optimised the additional sampling scheme of 19 observations for minimal KV using SSA. Figure 8.7 shows the exponential field with the preliminary sampling grid of 81 observations, together with the optimised additional sampling scheme. All 19 observations are placed outside or close to the borders of the actual sampling area. The two-stage scheme for the spherical field showed similar results.

For both simulated fields, the additional samples were collected and new experimental variograms were calculated and modelled using all data ($S^1 \cup S^2$). Table 8.3 shows the parameters for the modelled variograms. In comparison with the first stage, the nugget effect for the exponential model is much closer to that of the underlying variogram. The modelled range, however, is only 30.6, while the effective range of the underlying variogram is 50. The modelled variogram for the spherical field does not improve considerably as compared to the first sampling stage.

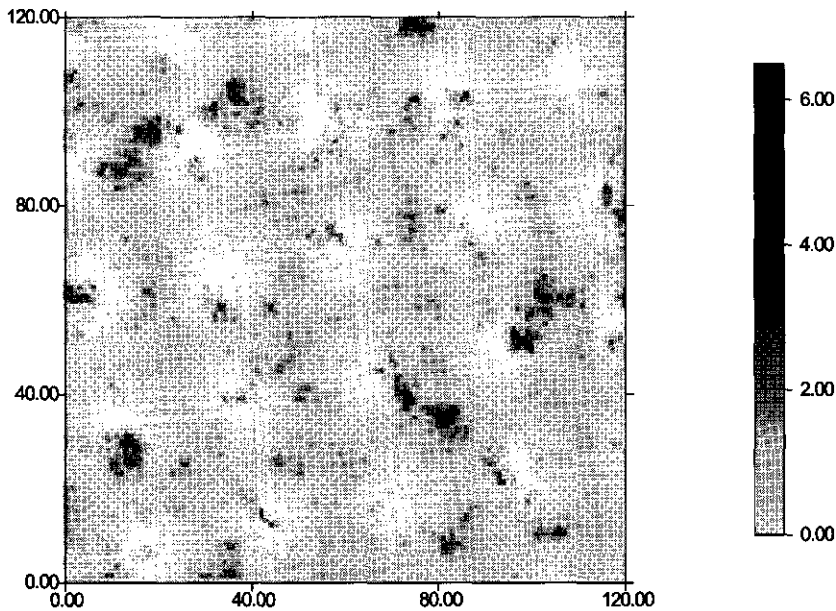


Figure 8.9. Squared kriging error ($[\text{unit}]^2$) for the sampling grid on the exponential field, number of observations is 100.

Figures 8.8 and 8.9 show the SKE of the exponential field, using the modelled variogram for the two-stage sampling scheme and the 10×10 sampling grid, respectively. Using the two-stage sampling scheme, the mean SKE decreased from $0.39 [\text{unit}]^2$ to $0.31 [\text{unit}]^2$. The maximum SKE decreased from $6.05 [\text{unit}]^2$ to $4.24 [\text{unit}]^2$. For the spherical field, the maximum SKE decreased from $15.98 [\text{unit}]^2$ to $11.52 [\text{unit}]^2$, while the mean SKE was $0.31 [\text{unit}]^2$ for both schemes.

Figures 8.8 and 8.9 show that the improvement in terms of SKE does not only occur in areas close to the boundaries of the area. Since the kriging neighbourhood was set equal to 14 observations, also locations closer to the centre were influenced by observations taken outside the borders of the research area.

8.5. Discussion

Although the use of stochastic simulations for testing the sampling schemes has some important advantages (which were mentioned in section 8.2.4.), also some drawbacks for this method can be observed. The Multi Gaussian simulation algorithm

produces stationary fields based on the intrinsic hypothesis. Therefore errors that can come up in practice due to nonstationarity or heteroscedasticity are not accounted for. However, the developed methods can also be applied in studies dealing with indicator variograms, which are much less sensitive to non-stationarity.

For similar reasons, minimisation of the kriging variance remains an optimisation criterion that will draw criticism in practice. The kriging variance only depends on the variogram, the configuration of observations and size of kriging neighbourhood. Therefore, local discontinuities *etc.* are not accounted for. In this study, we tried to avoid this problem by evaluating the resulting squared prediction errors. It was shown that optimising the additional sampling scheme for minimal kriging variance resulted in better predictions in both terms of both mean and maximum squared kriging errors.

The two-stage procedure for minimising squared kriging error as it was presented in this chapter can in principle be extended to more sampling stages. In that way, the sampling scheme can be fine tuned to the variogram. However, the first sampling stage should always be large enough to allow a reasonable estimation of the variogram. It was shown in this chapter that a number of 81 observations can be enough for this purpose.

The use of SSA with both WM and MAX_OK criteria will probably be even more beneficial in cases of pronounced anisotropy. Van Groenigen *et al.* (1999) (Chapter 4) showed that taking into account anisotropy can considerably reduce the kriging variance.

Use of SSA is also useful in terms of flexibility. In order to allow fair comparison between square grids and SSA, only sample sizes that are squared integers (*e.g.* 49, 81, 100) were used in this chapter. In contrast, SSA is able to handle all sample sizes.

8.6. Conclusions

In this chapter, three different optimisation strategies for estimation of the variogram were tested. The first strategy, which optimised the distribution of point pairs over the lag classes of the experimental variogram using the WM-criterion, was clearly inadequate as compared to regular grids of the same sample size. It led to much higher fluctuations of the experimental variogram, especially at larger distances. The second strategy combined a regular grid with extra observations taken at short distances using the WM-criterion. This improved estimation of variogram values at short distances, as compared to a sampling scheme of the same sample size. The third strategy aimed at optimal variogram use. It was shown that a roughly estimated variogram in the first sampling stage could lead to better predictions in terms of both mean and maximum squared kriging errors. For variogram estimation, we recommend the second strategy. For geostatistical interpolation studies we recommend the third.

8.7. Acknowledgements

The authors wish to thank Jaap de Gruijter for his part in discussions related to the content of this chapter.

Chapter 9

Conclusions and Further Research

9.1. Introduction

The research presented in chapters 2 to 8 dealt with a wide variety of optimisation criteria and case studies, most of which used the same general optimisation algorithm (Spatial Simulated Annealing), which therefore formed a coherent factor. The technical details about the algorithm and the optimisation criteria varied somewhat from one study to another and were presented piecemeal in the separate chapters. In this chapter, the main conclusions that can be drawn from these chapters will be presented integrally. Since the primary purpose of the study was the development of new optimisation tools, rather than the specific research questions dealt with in the case studies, the conclusions will focus on the methodology developed and on its use in future studies.

The purposes of the study as formulated in section 1.3. will be revisited, and briefly checked against the research presented. Finally, some recommendations for sampling and for further research are given.

9.2. Purposes of the study

Section 1.3. presented several purposes for this study. Below they are listed, with the main conclusions related to these purposes:

- i) *Formulation of a range of optimisation criteria that honour a wide variety of aims in soil-related surveys.*

In this thesis, a wide range of optimisation criteria has been formulated or drawn from the literature. Chapter 3 introduced the Warrick Myers (WM) criterion from the literature, aiming at optimal estimation of the variogram by reproducing an *a priori*

defined ideal point pair distribution (Equation 3.8). The MMSD (Minimisation of the Means of the Shortest Distances) criterion was formulated in the same chapter, with the aim of evenly spreading the observations over the area of interest (Equation 3.7). This criterion was generalised into the WMSD (Weighted Means of Shortest Distances) criterion, which was formulated in chapter 6, allowing the use of a weight function to direct more attention to certain sub-areas and thus leaving more room for decision-making (Equation 6.3).

In chapters 4 and 5, two additional optimisation criteria were introduced, aiming at minimisation of the ordinary kriging variance. The MEAN_OK criterion aims at minimising the mean kriging variance over the area of interest, whereas the MAX_OK criterion minimises the maximum occurring kriging variance over the area (Equation 4.5; Equation 5.7).

Table 9.1 lists these criteria, together with the required input and some remarks on their use in case studies.

- ii) *Development of an optimisation algorithm for spatial sampling that is able to handle these different optimisation criteria.*

The Spatial Simulated Annealing (SSA) algorithm was formulated and further developed in the course of this thesis. It is an algorithm specifically designed for optimising spatial sampling schemes. It is flexible in handling different optimisation criteria, and insensitive to local minima. All the formulated optimisation criteria were implemented in SSA, and more can be added in the future. Chapter 3 described the main outlines of the algorithm, and gave some examples using the MMSD and WM optimisation criteria. In chapters 4 to 6, the other optimisation criteria were formulated and implemented in the SSA algorithm for use in specific case studies. A preliminary version of the software for using the SSA algorithm, which I programmed in the course of my study, is available at <http://www.itc.nl/~groenig>. This site includes some example files and will be continually updated in the future.

- iii) *Incorporation of ancillary data such as co-related imagery, historic knowledge and expert knowledge in the sampling strategy.*

One of the main achievements of this thesis is the development of tools to incorporate many types of ancillary data in the sampling strategy. Chapter 2 presented a phased sampling strategy in which each sampling stage used the preceding sampling stage(s) to provide ancillary data. Using Indicator Kriging, the probability of exceeding

Table 9.1. Characteristics and required input of optimisation criteria implemented in Spatial Simulated Annealing (black) and future optimisation criteria (grey).

Criterion	Aim	Use in Spatial Studies	Required Input (besides map)	Remarks	Chapter
MMSD	Minimisation of the Mean Shortest Distance of an arbitrary point to the sampling scheme	Even spreading of points over area, especially with many sampling constraints. Filling up 'holes' in existing sampling schemes	-	-	3, 4, 6, 7
WMMSD	Weighted Minimisation of the Shortest Distance of an arbitrary point to the sampling scheme	Same as MMSD, but with possibility to distinguish between different grades of priority within the area	Spatial weight function, based upon historic knowledge and/or probability maps	-	6
WM	Formulated by Warrick and Myers. Reproduction of a target histogram of point pair distribution	Improving the accuracy of the experimental variogram, especially at shorter distances	Target histogram for point pair distribution. Number, size and direction of lags	Can only be used as supplement to an existing sampling scheme	3, 4, 8
MEAN_OK	Minimisation of the MEAN Ordinary Kriging variance	Minimisation of the overall error in spatial interpolation	-	-	4, 5
MAX_OK	Minimisation of the MAXimum Ordinary Kriging variance	Minimisation of the maximum error in spatial interpolation	Variogram of the ReV, size of kriging neighbourhood, anisotropy	Intrinsic hypothesis must hold	5, 8
MEAN_CO	Minimisation of the MEAN COkriging variance	Minimisation of the overall error in spatial interpolation, when many observations on co-related variables are available	(Co)variograms of the ReV's, size of kriging neighbourhood, anisotropy	See MEAN_OK. Number of (Co)variograms increases dramatically with number of ReV's	-

an environmental threshold value was used to direct sampling in subsequent stages. This procedure was combined with SSA in chapter 6.

The SSA algorithm allows the optimisation of spatial sampling schemes using many other types of ancillary data. Preliminary observations are used as an integrated, albeit static, part of the sampling scheme. This was presented and illustrated with some examples in chapter 3. In chapters 4 and 5, it was shown how the preliminary observations can be included into the optimised sampling scheme for minimal kriging variance. Chapter 6 showed how a large number of observations on multivariate contamination can be included in an optimised sampling scheme. It also shows how historical knowledge on land use can be incorporated into the sampling strategy.

One of the most attractive properties of SSA is that it can honour sampling constraints that are very common in (environmental) soil survey. Chapter 3 showed how SSA distinguishes between sampling constraints that are simply delineating the area of interest (e.g. boundaries of the area) and constraints that frustrate the sampling effort (e.g. buildings under which contamination should be assessed). In chapter 6, this property was used in a soil contamination study within a complex shaped study area with many buildings that prohibited sampling.

iv) Comparison of the performances of the developed optimisation algorithms with established sampling strategies.

Wherever possible, the optimisation algorithms were compared with more conventional sampling schemes. The use of probability maps in a phased sampling strategy to guide additional sampling to the more interesting areas (chapter 2) was compared with both a regular grid and a random sampling scheme using stochastic simulation (Table 2.2). Results clearly showed that a two-stage sampling approach outperformed both conventional schemes in terms of efficiency of derived action (remediation). The two-stage sampling scheme characterised 70% of the area correctly as either polluted or not polluted, as opposed to 55% by the regular grid (Table 2.3).

Most optimisation criteria for SSA were compared with conventional sampling schemes. Optimisation by SSA using the WM criterion for optimising variogram estimation was compared with a Monte Carlo optimisation from the literature in chapter 3. It was shown that SSA outperformed the Monte Carlo optimisation for all test studies in terms of the WM criterion, in one case reaching almost a complete solution. In chapter 8, the validity of the WM criterion as an optimisation criterion for variogram estimation was tested. Compared with a regular grid in a simulation study, WM-optimised schemes showed higher fluctuations in experimental variograms, especially at larger distances (Figure 8.5). WM-optimised schemes for observations in addition to a preliminary, systematic sampling scheme showed better results, especially at shorter

distances. It was therefore concluded that the WM criterion can be used for optimising additional observations for short-distance variogram estimation only.

Sampling schemes designed with the MMSD criterion were compared with an equilateral triangular grid in chapter 3. The optimised schemes showed a small but distinct improvement of 2.4% in a simple example (Figure 3.1). This improvement increased considerably to about 30% when more sampling constraints and preliminary observations were added (Table 3.1). In chapter 7, different uses of the MMSD criterion were compared with each other for optimally establishing soil/yield relationships. The use of an approximate yield map considerably improved the establishment of significant relations, thereby proving the potential benefits of using ancillary data in the sampling strategy (Tables 7.3 and 7.4.)

The performances of the WMSD criterion (the MMSD criterion combined with a spatial weight function) could not be compared with other schemes, since there is no good equivalent for this in conventional sampling strategies.

The two optimisation criteria for minimisation of the kriging variance (the MEAN_OK and MAX_OK criteria) were compared with a triangular grid in chapters 4 and 5 (Figure 5.3). In a simple example. The triangular grid was inferior in terms of both criteria. Performances of the MEAN_OK criterion improved with around 1.5%, while the MAX_OK criterion improved around 38.5% (Table 5.1). Furthermore, the MEAN_OK and MAX_OK criteria were compared with each other in chapter 5. The optimised schemes showed considerable differences, indicating that a clear choice has to be made for one of the criteria in surveys (Table 5.2). Table 9.1 summarises the aims and prerequisites for the various optimisation criteria formulated in this thesis.

In summary, the developed methodology performed better than conventional schemes in terms of all the optimisation criteria except the WM criterion. The extent of the improvements varied widely, depending on the optimisation criterion, the size of the sampling scheme and sampling constraints within the area. However, it should be emphasized that optimisation with SSA is especially beneficial in areas where boundaries are intricate and sampling constraints and preliminary observations are abundant.

v) *Application of developed optimisation techniques in practical soil sampling studies.*

Throughout the thesis, optimisation techniques I developed have been applied in a variety of case studies. In chapter 2, the developed phased sampling approach was applied to a lead contamination study in the city of Schoonhoven. Results of this case study were tested using stochastic simulation, showing that the phased sampling approach yielded better results than conventional techniques (Table 2.3).

This phased sampling approach, combined with SSA, was applied in a multivariate

pollution study in the city of Rotterdam (chapter 6). Knowledge on sampling constraints and historical knowledge on the area were incorporated in the sampling strategy. This highly intricate case study clearly showed the flexibility of the used methodology.

In chapter 4, SSA with the MEAN_OK criterion was applied in a study on a river terrace in Thailand. It was shown how a preliminary sampling scheme could be used to estimate the variogram, and subsequently used for completing the sampling scheme for minimisation of the kriging variance.

Chapter 8 showed that these optimised sampling schemes for minimal kriging variance also resulted in smaller squared prediction errors.

Chapter 7 showed how co-related imagery (yield maps) can assist in optimising the assessment of relations between soil and yield. This was done in a case study on precision agriculture for zero-input millet farming in Niger.

9.3. Recommendations for sampling

Out of the different optimisation criteria and case studies presented in this thesis, several practical and some more theoretical recommendations for sampling can be derived. They will be formulated below.

One of the more general conclusions is that the routinely applied regular sampling grid is almost never the optimal sampling scheme in terms of the formulated optimisation criteria. There are some exceptional cases where the combination of the shape of sampling area and the number of observations results in a regular grid. An example of this is shown in Figure 7.4a, where a square area with 9 observations results in a 3 x 3 square sampling grid. Another example can be seen on one of the figures shown on the back of the cover, where an equilateral triangular sampling area with 10 observations results in an equilateral triangular sampling grid. However, such cases are relatively rare. Case studies will often look more like Figure 6.3, with intricate boundaries, sampling constraints and preliminary observations. In such cases, use of SSA will considerably improve the efficiency of the survey, and the quality of its outcome. *Therefore, the use of SSA for dealing with cases of this type is recommended.*

Another recommendation is to very carefully consider what the exact objective of the study is. Although this may seem self-evident, the studies in this thesis showed that even very small differences in optimisation criteria could result in very different optimised sampling schemes. An example of this is shown in Figure 5.6, where two sampling schemes were optimised for minimal kriging variance (MEAN_OK criterion). The variograms for both schemes had similar nuggets, sills and (effective) ranges. Yet, the resulting sampling schemes differed considerably in outlook. In this respect, the choice between minimising the mean kriging variance and the maximum kriging variance is also crucial.

If optimal interpolation using kriging is the aim, taking observations just outside the area may considerably improve accuracy of the results. This is shown in Figure 8.7, where some observations taken outside the area improved both kriging variance and the actual squared kriging error. It should be emphasised that this is only true in situations where the boundaries of the sampling area do not coincide with those of the ReV in question. As an example, it may be very worthwhile to sample just outside a sample plot if the surrounding field has had the same treatment. However, it would be very unwise if the sample area is a mapping unit of a soil map based on sharply delineated physiographic units.

A final recommendation is to be very careful in using the point pair distribution as a measure of accuracy for variogram estimation. Figure 8.5 showed that this criterion (WM criterion) on its own is inferior compared with a regular grid. It should only be used to 'upgrade' an existing data set that roughly covers the whole area, for additional observations at short distances. An example of this is presented in Figure 8.6.

9.4. Recommendations for further research

In this thesis, SSA has been developed as a basis algorithm for optimisation of spatial sampling schemes. Although many optimisation criteria have already been formulated, the SSA algorithm leaves room for many more to be formulated in the future. Most importantly, an optimisation criterion dealing with cokriging will be formulated in the near future (Table 9.1). This will be especially useful in case studies where a large amount of co-related secondary data is available to improve the prediction accuracy of the primary variable. An obvious example of such secondary data is remotely sensed imagery. A drawback of cokriging is that it involves much modelling of (co)variograms, and that a model of co-regionalisation should be assumed. Space/time geostatistics offers many opportunities for new optimisation criteria, and a start has been made by Stein *et al.* (1998). Finally, cost models should be included in the optimisation criteria.

References

- Aarts, E. and Korst, J. (1989). *Simulated annealing & Boltzmann machines - a stochastic approach to combinatorial optimization and neural computing*. John Wiley & Sons, New York.
- Armstrong, M.P. and Densham, P.J. (1990). *Database organisation strategies for spatial decision support systems*. International Journal of GIS 4: 3-20.
- Anonymous. (1994). *Circulaire interventiewaarden bodemsanering, Ministerie van VROM*. Staatsuitgeverij, The Hague.
- Anonymous. (1995). *Leidraad bodembescherming, Ministerie van VROM*. Staatsuitgeverij, The Hague.
- Bardossy, A., Bogardi, I. and Kelly, W.E. (1990a). *Kriging with imprecise (fuzzy) variogram, Part I - Theory*. Mathematical Geology 22: 63-79.
- Bardossy, A., Bogardi, I. and Kelly, W.E. (1990b). *Kriging with imprecise (fuzzy) variogram, Part II - Application*. Mathematical Geology 22: 81-94.
- Bierkens, M.F.P. and Burrough, P.A. (1993a). *The indicator approach to categorical soil data I. Theory*. Journal of Soil Science 44: 361-368.
- Bierkens, M.F.P. and Burrough, P.A. (1993b). *The indicator approach to categorical soil data. II. Application to mapping and land use suitability analysis*. Journal of Soil Science 44: 369-381.
- Bohachevsky, I., Johnson, M.E. and Stein, M.L. (1986). *Generalized simulated annealing for function optimization*. Technometrics 28: 209-217.
- Booltink, H.W.G. and Verhagen, J. (1997). *Integration of remote sensing, modelling and field measurements towards an operational decision support system for precision agriculture*. In: Stafford, J.V. (ed.) Precision Agriculture '97. Bios, Oxford, pp. 921-929.
- Boucneau, G., Van Meirvenne, M., Thas, O. and Hofman, G. (1998). *Integrating properties of soil map delineations into ordinary kriging*. European Journal of Soil Science 49: 213-229.
- Bouma, J. (1997). *Precision agriculture: introduction to the spatial and temporal variability of environmental quality*. In: CIBA Foundation, CIBA Foundation Symposium 210. Wiley, Chichester, pp. 5-13.
- Bouma, J., Verhagen, J., Brouwer, J. and Powell, J.M. (1996). *Using systems approaches for targeting site specific management on field level*. In: Kropff, M.J., Teng, P.S., Aggerwd, P.K., Bouma, J., Bouman, B.A.M., Jones, J.W. and Van Laar, H.H. (eds.) Applications of systems approaches at the field level. Kluwer, Dordrecht, pp. 25-36.
- Bregt, A.K., McBratney, A.B. and Wopereis, M.C.S. (1991). *Construction of isolinear maps of soil attributes with empirical confidence limits*. Soil Science Society of America Journal 55: 14-19.
- Broos, M.J., Pieters, G., Van Tooren, C.F. and Stein, A. (submitted). *Quantification of the effects of spatially varying environmental contaminants into a financial model for soil remediation*. Journal of Environmental Management.

- Brouwer, J. (1996). *Water and nutrients alternate in limiting agricultural production in the Sabel*. In: Centre National de la Recherche Scientifique et de la Technologie, Proceedings of the 1st international conference of the West and Central African Soil Science Association. Ouagadougou, Burkina Faso.
- Brouwer, J., Fussel, L.K. and Herrmann, L. (1993). *Soil and crop growth micro-variability in the West African semi-arid tropics: a possible risk reducing factor for subsistence farmers*. Agriculture, Ecosystems and Environment 45: 229-238.
- Brus, D.J. (1993). *Incorporating models of spatial variation in sampling strategies for soil*. Ph.D. thesis. Pudoc, Wageningen.
- Brus, D.J. (1994). *Improving design-based estimation of spatial means by soil map stratification*. Geoderma 62: 233-246.
- Brus, D.J. and De Gruijter, J.J. (1993). *Design-based versus model-based estimates of spatial means. Theory and application in environmental soil science*. Environmetrics 4: 123-152.
- Brus, D.J. and De Gruijter, J.J. (1994). *Estimation of non-ergodic variograms and their sampling variance by design-based sampling strategies*. Mathematical Geology 26: 437-454.
- Brus, D.J. and De Gruijter, J.J. (1997). *Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion)*. Geoderma 80: 45-49.
- Buerkert, A., Stern, R.D. and Marchner, H. (1995). *Post stratification clarifies treatment effects on pearl millet growth in the Sabel*. Agronomic Journal 87: 752-763.
- Burgess, T.M. and Webster, R. (1984). *Optimal sampling strategies for mapping soil types - II. Risk functions and sampling intervals*. Journal of Soil Science 35: 655-665.
- Burgess, T.M., Webster, R. and McBratney, A.B. (1981). *Optimal interpolation and isarithmic mapping of soil properties. IV Sampling strategy*. Journal of Soil Science 32: 643-659.
- Burrough, P.A. (1986). *Principles of geographical information systems for land resources assessment*. Oxford University Press, New York.
- Burrough, P.A. (1992). *Development of intelligent geographical information systems*. International Journal of GIS 6: 1-11.
- Burrough, P.A. (1993). *Principles of geographical information systems for land resources assessment*. Oxford University Press, New York.
- Cerny, V. (1985). *Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm*. Journal of Optimization Theory and Applications 45: 41-51.
- Christakos, G. and Olea, R.A. (1992). *Sampling design for spatially distributed hydrogeologic and environmental processes*. Advances in Water Resources 15: 219-237.
- City of Rotterdam. (1997). *Masterplan Schiehaven-Mullerpier*. City of Rotterdam, Rotterdam.
- Cochran, W.G. (1977). *Sampling techniques*. Wiley, New York.
- Corsten, L.C.A. and Stein, A. (1994). *Nested sampling for estimating spatial semi-variograms compared to other designs*. Applied Stochastic Models and Data Analysis 10: 103-122.
- Cressie, N.A.C. (1991). *Statistics for spatial data*. John Wiley & Sons, New York.
- Csillag, F., Kertész, M. and Kummert, Á. (1996). *Sampling and mapping of heterogeneous surfaces: multi-resolution tiling adjusted to spatial variability*. International Journal of GIS 10: 851-875.
- Davis, L. (1990). *Genetic algorithms and simulated annealing*. Morgan Kaufmann Publishers, Los Altos.
- De Gruijter, J.J. and Ter Braak, C.J.F. (1990). *Model free estimation from spatial samples: a reappraisal of classical sampling theory*. Mathematical Geology 22: 407-415.

- Deutsch, C.V. and Cockerman, P. (1994). *Practical considerations in the applications of simulated annealing to stochastic simulation*. *Mathematical Geology* 26: 67-72.
- Deutsch, C.V. and Journel, A.G. (1992). *GS-LIB - Geostatistical software library and user's guide*. Oxford University Press, Oxford.
- Deutsch, C.V. and Journel, A.G. (1998). *GS-LIB - Geostatistical software library and user's guide*. Oxford University Press, Oxford.
- Domburg, P. (1994). *A Knowledge-based system to assist in the design of soil survey schemes*. Wageningen Agricultural University, Wageningen.
- Domburg, P., De Gruijter, J.J. and Brus, D.J. (1994). *A structured approach to designing soil survey schemes with prediction of sampling error from variograms*. *Geoderma* 62: 151-164.
- Domburg, P., De Gruijter, J.J. and van Beek, P. (1997). *Designing efficient soil survey schemes with a knowledge-based system using dynamic programming*. *Geoderma* 75: 183-201.
- FAO. (1994). *Soil map of the world - revised legend with corrections*. ISRIC, Wageningen.
- Flatman, G.T. and Yfantis, E.A. (1984). *Geostatistical strategy for soil sampling: the survey and the census*. *Environmental Monitoring and Assessment* 4:335-349.
- Gandah, M., Bouma, J., Brouwer, J. and Van Duivenbooden, N. (1998). *Use of a scoring technique to assess the effect of field variability on yield of pearl millet grown on three Alfisols in Niger*. *Netherlands Journal of Agricultural Science* 46: 39-52.
- Gascuel-Oudou, C. and Boivin, P. (1994). *Variability of variograms and spatial estimates due to soil sampling: a case study*. *Geoderma* 62: 165-182.
- Gaze, S.R., Simmonds, L.P., Brouwer, J. and Bouma, J. (1997). *Measurement of surface redistribution of rainfall and modelling its effect on water balance calculations for a millet field on sandy soil in Niger*. *Journal of Hydrology* 188-189: 267-284.
- Geman, S. and Geman, D. (1984). *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*. *IEEE transactions on pattern analysis and machine intelligence* 6: 721-741.
- Goldstein, L., and Waterman, M.S. (1987). *Mapping DNA by stochastic relaxation*. *Advanced Applied Mathematics* 8: 194-207.
- Gómez-Hernández, J.J. and Srivastava, R. (1990). *ISIM3D: and ANSI-C three dimensional multiple indicator conditional simulation program*. *Computers and Geosciences* 16: 395-440.
- Goovaerts, P. (1996). *Stochastic simulation of categorical variables using a classification algorithm and simulated annealing*. *Mathematical Geology* 28: 909-921.
- Goovaerts, P. (1997a). *Geostatistics for natural resources evaluation*. Oxford University Press, Oxford.
- Goovaerts, P. (1997b). *Kriging vs. stochastic simulation for risk analysis in soil contamination*. In: Soares, A., Gómez-Hernández, J.J. and Froidevaux, R. (eds.) *geoENV I - geostatistics for environmental applications*. Kluwer, Dordrecht, pp. 247-258.
- Haslett, J., Wills, G. and Unwin, A. (1990). *SPIDER - An interactive statistical tool for the analysis of spatially distributed data*. *International Journal of GIS* 4: 285-296.
- Hendricks-Franssen, H.J.W.M. and Gómez-Hernández, J.J. (1997). *Impact of random function choice on groundwater mass transport modelling*. In: Soares, A., Gómez-Hernández, J.J. and Froidevaux, R. (eds.) *geoENV I - geostatistics for environmental applications*. Kluwer, Dordrecht, pp. 101-110.
- Hesse, P.R. (1971). *A textbook of soil chemical analysis*. John Murry, London.
- Heuvelink, G.B.M. and Bierkens, M.F.P. (1992). *Combining soil maps with interpolations from point observations to predict quantitative soil properties*. *Geoderma* 55: 1-15.

- Heuvelink, G.B.M., Burrough, P.A. and Stein, A. (1989). *Propagation of errors in spatial modelling with GIS*. International Journal of GIS 3: 303-322.
- Isaaks, E.H. and Srivastava, R.M. (1989). *An introduction to applied geostatistics*. Oxford University Press, Oxford.
- Journal, A.G. (1983). *Nonparametric estimation of spatial distributions*. Mathematical Geology 15: 445-468.
- Journal, A.G. (1985). *The deterministic side of geostatistics*. Mathematical Geology 17: 1-15.
- Journal, A.G. (1989). *Fundamentals of geostatistics in five lessons*. American Geophysical Union, Washington.
- Journal, A.G. and Huijbrechts, Ch.J. (1978). *Mining geostatistics*. Academic Press, London.
- Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983). *Optimisation by simulated annealing*. Science 220: 671-680
- Kollias, V.J. and Voliotis, A. (1991). *Fuzzy reasoning in the development of geographical information systems FRSS: a prototype soil information system with fuzzy retrieval capabilities*. International Journal of GIS 5: 209-223.
- Laslett, G.M., McBratney, A.B., Pahl, P.J. and Hutchinson, M.F. (1987). *Comparison of several spatial prediction methods for soil pH*. Journal of Soil Science 38: 325-341.
- Mainam, F. and Zinck, J.A. (1998). *Erodibility assessment of selected soils in Northern Cameroon using a field rainfall simulator*. Proceedings of the World Congress of Soil Science, 20-26 August, 1998, Montpellier. CD-ROM.
- Matheron, G. (1973). *The intrinsic random functions and their applications*. Advances in Applied Probability 5: 439-468.
- McBratney, A.B. and Webster, R. (1983). *Optimal interpolation and isarithmic mapping of soil properties V. Co-regionalization and multiple sampling strategy*. Journal of Soil Science 34: 137-162.
- McBratney, A.B., Webster, A. and Burgess, T.M. (1981). *The design of optimal sampling schemes for local estimation and mapping of regionalized variables: 1. Theory and method*. Computer and Geosciences 7: 331-334.
- McBratney, A.B. and Webster, R. (1981). *The design of optimal sampling schemes for local estimation and mapping of regionalized variables: 2. Program and examples*. Computers and Geosciences 7: 335-365.
- McBratney, A.B. and Webster, R. (1986a). *Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates*. Journal of Soil Science 37: 617-639.
- McBratney, A.B. and Webster, R. (1986b). *Combined nested and linear sampling for determining the scale and form of spatial variation of regionalized variables*. Geographical Analysis 18: 225-242.
- Morris, M.D. (1991). *On counting the number of data pairs for semi-variogram estimation*. Mathematical Geology 23: 929-943.
- Odeh, I.O.A., McBratney, A.B. and Chittleborough, D.J. (1990). *Design of optimal sampling spacings for mapping soil using fuzzy k-means and regionalised variable theory*. Geoderma 47: 93-122.
- Okx, J.P., Heuvelink, G.B.M. and Grinwis, A.W. (1990). *Expert knowledge and (geo)statistical methods: complementary tools in soil pollution research*. In: Arendt, F., Hinsenfeld, M. and Van den Brink, W.J. (eds) Contaminated soil '90. Kluwer, Dordrecht, pp. 729-736.
- Okx, J.P., Hordijk, L. and Stein, A. (1996). *Managing soil remediation problems*. Environmental Science and Pollution Research 3: 229-235.

- Oliver, M.A. and Webster, R. (1987). *The elucidation of soil pattern in the Wyre Forest of the West Midlands, England. II: Spatial distribution*. Journal of Soil Science 38: 293-307.
- Petitgas, P. (1993). *Use of disjunctive kriging to model areas of high pelagic fish density in acoustic fisheries surveys*. Aquatic Living Resources 6: 285-298.
- Pettitt, A.N. and McBratney, A.B. (1993). *Sampling designs for estimating spatial variance components*. Applied Statistics 42: 185-209.
- Pieters, G. van Tooren, C. and Martens, J. (1996). *Integraal bodemonderzoek Schiehaven-Mullerpier. Situatie per 1-11-1996*. IM Engineering office, Rotterdam.
- Rasch, D. (1990). *Optimum experimental designs in nonlinear regression*. Communications on Statistical Theory and Methods 19: 4789-4806.
- RIVM, (1991). *Nationale milieuverkenning 2 - 1990-2010*. RIVM, Bilthoven.
- Russo, D. (1984). *Design of an optimal sampling network for estimating the variogram*. Soil Science Society of America Journal 48: 708-716.
- Russo, D. and Jury, W.A. (1988). *Effect of sampling network on estimates of the covariance function of stationary fields*. Soil Science Society of America Journal 52: 1228-1234.
- Sacks, J. and Schiller, S. (1988). *Spatial designs*. In: Gupta, S.S. and Berger, J.O. (eds.) Statistical decision theory and related topics IV, Vol. 2. Springer-Verlag, New York, pp. 385-399.
- Seneta, E. (1981). *Non-negative matrices and Markov chains, 2nd ed.* Springer verlag, New York.
- Seo, D.J., Krajewski, W.F. and Bowles, D.S. (1990). *Stochastic interpolation of rainfall data from rain gages and radar using cokriging 1. Design of experiments*. Water Resources Research 26: 469-477.
- Shafer, J.M. and Varljen, M.D. (1990). *Approximation of confidence limits on sample semivariograms from single realisations of spatially correlated random fields*. Water Resources Research 2: 1787-1802.
- Sivakumar, M.V.K., Maidoukia, A. and Stern, R.D. (1993). *Agroclimatology of West Africa: Niger*. ICRISAT, Niamey.
- Soil Survey Staff. (1996). *Keys to Soil Taxonomy (7th edition)*. USDA, Blacksburg.
- Stein, A., Brouwer, J. and Bouma, J. (1997). *Methods for comparing spatial variability patterns of millet yield and soil data*. Soil Science Society of America Journal 61: 861-870.
- Stein, A. and Corsten, L.C.A. (1991). *Universal kriging and cokriging as a regression procedure*. Biometrics 47: 1227-1239.
- Stein, A., Hoogerwerf, M. and Bouma, J. (1988a). *Use of soil-map delineations to improve (co-)kriging of point data on moisture deficits*. Geoderma 43:163-177.
- Stein, A., Staritsky, I.G., Bouma, J. and Van Groenigen, J.W. (1995). *Interactive GIS for environmental risk assessment*. International Journal of GIS 5: 509-525.
- Stein, A., Van Dooremolen, W. Bouma, J. and Bregt, A.K. (1988b). *Cokriging point data on moisture deficit*. Soil Science Society of America Journal 52: 1418-1423.
- Stein, A., Van Groenigen, J.W., Jeger, M. and Hoosbeek, M.R., (1998). *Space-time statistics for environmental and agricultural related phenomena*. Environmental and Ecological Statistics 5: 155-172.
- Stevens, D.L. (1997). *Variable density grid-based sampling designs for continuous spatial populations*. Environmetrics 8: 167-195.
- Stoorvogel, J.J. (1995). *Linking GIS and models: structure and operationalisation for a Costa Rican case study*. Geoderma 43: 19-29.

- Taylor, C.C. and Burrough, P.A. (1986). *Multiscale sources of spatial variation in soil. III. Improved methods for fitting the nested model to one-dimensional semi-variograms*. *Mathematical Geology* 18: 811-821.
- Thompson, S.K. (1992). *Sampling*. John Wiley & Sons, New York.
- Thompson, S.K. and Seber, G.A.F. (1996). *Adaptive Sampling*. John Wiley & Sons, New York.
- Van Bergeijk, J. and Goense, D. (1997). *Soil tillage resistance as a tool to map soil type differences*. In: Stafford, J.V. (ed.) *Precision Agriculture '97*. BIOS, Oxford, pp. 605-616.
- Vanderbilt, D. and Louie, S.G. (1984). *A Monte-Carlo simulated annealing approach to optimisation over continuous variables*. *Journal of Computational Physics* 36: 259-271.
- Van Groenigen, J.W. (submitted). *The influence of variogram parameters on optimal sampling schemes for kriging*. *Geoderma*.
- Van Groenigen, J.W., Pieters, G. and Stein, A. (submitted). *Optimising spatial sampling for multivariate contamination in urban areas*. *Environmetrics*.
- Van Groenigen, J.W., Siderius, W. and Stein, A. (1999). *Constrained optimisation of soil sampling for minimisation of the kriging variance*. *Geoderma* 87: 239-259.
- Van Groenigen, J.W. and Stein, A. (1998). *Constrained Optimisation of Spatial Sampling using Continuous Simulated Annealing*. *Journal of Environmental Quality* 27: 1078-1086.
- Van Groenigen, J.W., Stein, A. and Zuurbier, R. (1997). *Optimisation of environmental sampling using interactive GIS*. *Soil Technology* 10: 83-97.
- Van Tooren, C.F. (1993). *Guess-field methode; gebruik van voorkennis bij bodemverontreinigings onderzoek*. CHO-TNO rapporten en nota's 30: 83-97.
- Van Uffelen, C.G.R., Verhagen, J. and Bouma, J. (1997). *Comparison of simulated crop yield patterns for site specific management*. *Agricultural Systems* 54: 207-222.
- Velthof, G.L., Van Groenigen, J.W., Gebauer, G., Pietrzak, S., Jarvis, S.C., Pinto, M., Corre, W. and Oenema, O. (submitted). *Temporal stability of spatial patterns of nitrous oxide fluxes from a sloping fertilized grassland soil*. *Journal of Environmental Quality*.
- Warrick, A.W. and Myers, D.E. (1987). *Optimisation of sampling locations for variogram calculations*. *Water Resources Research* 23: 496-500.
- Watson, A.G. and Barnes, R.J. (1995). *Infill sampling criteria to locate extremes*. *Mathematical Geology* 27: 589-608.
- Webster, R. and Oliver, M.A. (1989). *Optimal interpolation and isarithmic mapping of soil properties: VI. Disjunctive kriging and mapping the conditional probability*. *Journal of Soil Science* 40: 497-512.
- Webster, R. and Oliver, M.A. (1990). *Statistical methods in soil and land resource survey*. Oxford University Press, Oxford.
- Webster, R. and Oliver, M.A. (1992). *Sample adequately to estimate variograms of soil properties*. *Journal of Soil Science* 43: 177-192.
- Winkels, H.J. and Stein, A. (1997). *Optimal cost-effective sampling for monitoring and dredging of contaminated sediments*. *Journal of Environmental Quality* 26: 933-946.
- Yfantis, E.A., Flatman, G.T. and Behar, J.V. (1987). *Efficiency of kriging estimation for square, triangular, and hexagonal grids*. *Mathematical Geology* 19: 183-205.
- Zimmerman, D.L. and Homer, K.E. (1991). *A network design criterion for estimating selected attributes of the semi-variogram*. *Environmetrics* 2: 425-441.

Bibliography

Principal Author

- Van Groenigen, J.W. (1997). *Spatial simulated annealing for optimising sampling – different optimisation criteria compared*. In: Soares, A., Gómez-Hernandez, J. and Froidevaux, R. (eds.) *geoENVI-Geostatistics for environmental applications*. Kluwer, Dordrecht, pp. 351-362.
- Van Groenigen, J.W. (submitted). *The influence of variogram parameters on optimal sampling schemes for kriging*. *Geoderma*.
- Van Groenigen, J.W., Gandah, M. and Bouma, J. (submitted). *Soil sampling strategies for precision agriculture research under Sahelian conditions*. *Soil Science Society of America Journal*.
- Van Groenigen, J.W., Mainam, F. and Stein, A. (submitted). *Sampling strategies for effective variogram estimation*. *European Journal of Soil Science*.
- Van Groenigen, J.W., Pieters, G. and Stein, A. (submitted). *Optimising spatial sampling for multivariate contamination in urban areas*. *Environmetrics*.
- Van Groenigen, J.W., Siderius, W. and Stein, A. (1999). *Constrained optimisation of soil sampling for minimisation of the kriging variance*. *Geoderma* 87: 239-259.
- Van Groenigen, J.W. and Stein, A. (1998). *Constrained Optimisation of Spatial Sampling using Continuous Simulated Annealing*. *Journal of Environmental Quality* 27: 1078-1086.
- Van Groenigen, J.W., Stein, A. and Zuurbier, R. (1997). *Optimisation of environmental sampling using interactive GIS*. *Soil Technology* 10: 83-97.

Other publications

- Stein, A., Bastiaanse, W.G.M., de Bruin, S., Cracknell, A.P., Curran, P.J., Fabbri, A.G., Gorte, B.G.H., Van Groenigen, J.W., Van der Meer, F.D. and Saldaña, A. (1998). *Integrating spatial statistics and remote sensing*. *International Journal of Remote Sensing* 19: 1793-1814.
- Stein, A., Staritsky, I.G., Bouma, J. and Van Groenigen, J.W. (1995). *Interactive GIS for environmental risk assessment*. *International Journal of GIS* 5: 509-525.
- Stein, A. and Van Groenigen, J.W. (1997). *Spatial and temporal statistical procedures to support decision making for smart farming*. In: ten Berge, H.F.M. and Stein, A. (eds.) *Model-based decision support in agriculture*. C.T. de Wit Graduate school for Production Ecology, Wageningen, pp. 125-129.

- Stein, A., Van Groenigen, J.W., Jeger, M. and Hoosbeek, M.R., (1998). *Space-time statistics for environmental and agricultural related phenomena*. Environmental and Ecological Statistics 5: 155-172.
- Velthof, G.L., Van Groenigen, J.W., Gebauer, G., Pietrzak, S., Jarvis, S.C., Pinto, M., Corre, W. and Oenema, O. (submitted) *Temporal stability of spatial patterns of nitrous oxide fluxes from a sloping fertilized grassland soil*. Journal of Environmental Quality.

Summary

Aims

This thesis aims at the development of optimal sampling strategies for geostatistical studies. Special emphasis is on the optimal use of ancillary data, such as co-related imagery, preliminary observations and historic knowledge. Although the object of all studies is the soil, the developed methodology can be used in any scientific field dealing with geostatistics.

In summary, the objectives of this study were:

- i)* Formulation of a range of optimisation criteria that honour a wide variety of aims in soil-related surveys.
- ii)* Development of an optimisation algorithm for spatial sampling that is able to handle these different optimisation criteria.
- iii)* Incorporation of ancillary data such as co-related imagery, historic knowledge and expert knowledge in the sampling strategy.
- iv)* Comparison of the performances of the developed optimisation algorithms with established sampling strategies.
- v)* Application of developed optimisation techniques in practical soil sampling studies.

Outline of major tools

Chapter 2 shows how a phased sampling procedure can optimise environmental risk assessment. Using indicator kriging, probability maps of exceeding environmental threshold levels are used to direct subsequent sampling. The method is applied in a lead-pollution study in the city of Schoonhoven, The Netherlands. It is tested using stochastic simulations, and results are compared to conventional sampling schemes in terms of type-I and type-II errors. The phased sampling schemes have much lower type-I errors than the conventional sampling schemes with comparable type-II errors. They predict almost 70% of the area correctly (polluted or not-polluted), as compared to 55% by conventional schemes.

Chapter 3 introduces the spatial simulated annealing (SSA) algorithm as a general,

flexible optimisation method for spatial sampling. Sampling schemes are optimised at the point level, taking into account sampling constraints and preliminary observations. Different optimisation criteria can be handled. SSA is demonstrated using two optimisation criteria from the literature. The first (the MMSD criterion) aims at even spreading of points over the area. The second (WM criterion) optimises the realised point pair distribution for variogram estimation. For several examples it is shown that SSA is superior to conventional sampling strategies. Improvements up to 30% occur for the first criterion, while an almost complete solution is found for the second criterion. SSA is flexible in adding extra criteria.

Optimising sampling for spatial interpolation

Chapter 4 introduces the MEAN_OK algorithm in SSA, which aims at minimisation of the mean ordinary kriging variance over the research area. It is applied on texture and phosphate content on a river terrace in Thailand. First, sampling is conducted for estimation of the variogram. The variograms thus obtained are used to optimise additional sampling for minimal kriging variance using SSA. This reduces kriging variance of sand percentage from 28.2 to 23.7 (%)². The variograms are used subsequently in a geomorphologically similar area. Optimised sampling schemes for anisotropic variables differ considerably from isotropic ones. Size of kriging neighbourhood has a small but distinct effect on the schemes. The schemes are especially efficient in reducing high kriging variances near boundaries of the area.

Chapter 5 further explores the possibilities of minimising kriging variance using SSA. Next to the MEAN_OK criterion, the MAX_OK criterion is introduced, which minimises maximum kriging variance. Both criteria are compared to a regular grid. Using SSA, the mean kriging variance reduces from 40.64 [unit]² to 39.99 [unit]². The maximum kriging variance reduces from 68.83 [unit]² to 53.36 [unit]². An additional sampling scheme of 10 observations is optimised for an irregularly scattered data set of 100 observations. This reduces the mean kriging variance from 21.62 [unit]² to 15.83 [unit]², and maximum kriging variance from 70.22 [unit]² to 34.60 [unit]². The influence of variogram parameters on the optimised sampling schemes is investigated. A Gaussian variogram produces a very different sampling scheme than an exponential variogram with similar nugget, sill and (effective) range. A very short range results in random sampling schemes, with observations separated by distances larger than twice the range. For a spherical variogram, magnitude of the relative nugget effect does not effect the sampling schemes, except for high values.

Chapter 6 introduces the WMSD criterion into SSA, which optimises sampling using a spatial weight function. This allows distinguishing between different areas of priority. A multivariate contamination study in the Rotterdam harbour with five

contaminants at two depths shows two subsequent sampling stages with two spatial weight functions. The first stage combines earlier observations and historic knowledge, with emphasis on areas with high priority. The resulting scheme shows a contamination at 17.4% of the samples, with 1.5% heavily contaminated. The second stage uses probability maps of exceeding intermediate threshold values to guide additional sampling to possible hot spots. This yields 26.7% contaminated samples, with 16.7% heavily contaminated. These include new locations that were not detected during the first stage. The WMSD criterion can be used as a valuable tool in decision making processes.

Optimising sampling for model estimation

Chapter 7 focuses on the use of ancillary data to optimise sampling for precision farming research. Using a cheap, low-tech scoring technique yield maps were predicted for millet in an on-farm study in Niger. Yield varied from 0 to 2500 kg ha⁻¹. Subsequently, SSA was used to optimise three different sampling schemes. Scheme 1 optimised coverage of the whole area. Scheme 2 covered the whole yield range, and scheme 3 covered the low producing areas. Using correlation coefficients, scheme 2 found significant correlations between 5 variables and yield. Scheme 1 found only one significant correlation. Using multivariate regression of yield on soil variables, scheme 2 explained 70% of the yield variation. For scheme 1 this was only 37%. Differences between scheme 3 and scheme 1 proved to be significant for distance to shrubs, micro-relief, pH-H₂O and CEC. From this study we concluded that shrubs are the main factor influencing yield by catching eroded particles and improving soil fertility. In general, we concluded that the sampling strategy of scheme 2 should be recommended for establishing yield/soil relations. Variograms of micro-relief and yield suggested that spatial correlation is largely confined to distances of 3 to 5 m.

Chapter 8 evaluates a number of sampling strategies for variogram estimation. In the first part, a regular grid is compared to a sampling scheme that optimises the point pair distribution for variogram estimation. This yields unbiased experimental variograms. However, the fluctuation of the experimental variograms is much lower with a regular grid. We concluded from this that the point pair distribution alone is not a useful optimisation criterion for variogram estimation. In the second part, additional observations selected for optimal point pair distribution are compared with randomly drawn additional observations. The random observations result in much higher standard deviations at shorter distances. We concluded from this that for additional short distance observations the point pair distribution is a very useful optimisation criterion. The third part focusses on optimal variogram use. A sampling grid of 81 observations is completed, after preliminary estimation of the variogram, with 19 additional observations for minimal kriging variance. The scheme is compared to a regular grid

of 100 observations. For an exponential field without nugget effect, the use of the phased sampling scheme reduces the mean squared kriging error from 0.39 [unit]² to 0.31 [unit]², and the maximum squared kriging error from 6.05 [unit]² to 4.24 [unit]². For a spherical field with a nugget effect of 33%, mean squared kriging error does not change and maximum squared kriging error decreases from 15.98 [unit]² to 11.52 [unit]². We concluded that minimisation of the squared kriging error is often more relevant than accurate estimation of the variogram. Taking samples just outside the area improved the quality of the prediction in terms of both kriging variance and squared kriging error.

Samenvatting

Doelstellingen

Dit proefschrift richt zich op de ontwikkeling van optimale bemonsteringsstrategieën voor geostatistische studies. Er is speciale aandacht voor het optimale gebruik van voorinformatie, zoals gerelateerd kaartmateriaal, eerdere observaties en historische kennis. Alhoewel alle case studies afkomstig zijn uit de bodemkunde, kunnen de ontwikkelde strategieën gemakkelijk worden aangepast voor andere wetenschappelijke velden waar geostatistiek een rol speelt.

De belangrijkste doelen van de studie kunnen als volgt worden samengevat:

- i)* Formuleren van een aantal optimalisatiecriteria die recht doen aan de grote verscheidenheid aan doelen in bodem gerelateerde studies.
- ii)* Ontwikkelen van een optimalisatie algoritme voor ruimtelijke bemonstering dat in staat is om verschillende optimalisatiecriteria te onderscheiden.
- iii)* Integreren van voorinformatie zoals gerelateerd kaartmateriaal, eerdere observaties and historische kennis in de bemonsteringsstrategie.
- iv)* Vergelijken van de ontwikkelde bemonsteringsstrategieën met conventionele strategieën.
- v)* Toepassen van de ontwikkelde optimalisatie technieken in praktische bodembemonsterings studies.

Ontwikkelde methodologie

Hoofdstuk 2 laat zien hoe een gefaseerde bemonsteringsprocedure ruimtelijke risicoinventarisatie kan optimaliseren. Kaarten met overschrijdingskansen van drempelwaarden worden geconstrueerd met behulp van Indicator Kriging. Deze worden vervolgens gebruikt om bemonstering in volgende fasen aan te sturen. De methode wordt toegepast in een loodverontreinigingsstudie in Schoonhoven, en wordt getest door middel van stochastische simulatie. De resultaten van de bemonsteringsstrategieën worden uitgedrukt in termen van type-I en type-II fouten. De gefaseerde aanpak resulteert in veel lagere type-I fouten dan die van conventionele bemonsteringsschema's,

terwijl type-II fouten gelijk blijven. De gefaseerde aanpak voorspelt bijna 70% van het gebied correct (verontreinigd of niet verontreinigd), en de conventionele aanpak slechts 55%.

Hoofdstuk 3 introduceert het Spatial Simulated Annealing (SSA) algoritme. SSA is een breed toepasbaar en flexibel algoritme voor optimalisatie van ruimtelijke bemonstering. Bemonsteringsschema's worden geoptimaliseerd op het punt niveau, waarbij rekening wordt gehouden met praktische bemonsteringsbeperkingen en eerdere observaties. Verschillende optimalisatiecriteria kunnen worden gehanteerd. Het gebruik van SSA wordt geïllustreerd aan de hand van twee, uit de literatuur gehaalde criteria. Het eerste (MMSD) criterium richt zich op het gelijkmatig spreiden van de punten over het gebied. Het tweede (WM) criterium optimaliseert de gerealiseerde puntenpaar verdeling voor het experimentele variogram. Door middel van verschillende voorbeelden wordt aangetoond dat SSA superieur is in vergelijking met conventionele bemonsteringsstrategieën. De prestaties van SSA zijn tot 30% beter voor het eerste criterium, terwijl voor het tweede criterium een bijna complete oplossing wordt gevonden. SSA biedt veel mogelijkheden voor het implementeren van nieuwe optimalisatie criteria.

Bemonsteringsoptimalisatie voor ruimtelijke interpolatie

In hoofdstuk 4 wordt SSA uitgebreid met het MEAN_OK criterium, dat zich richt op minimalisatie van de gemiddelde ordinary kriging variantie. Dit criterium wordt toegepast in een bemonsteringsoptimalisatie voor textuur en fosfaat gehalte op een rivierterras in Thailand. Als eerste wordt een bemonstering uitgevoerd voor het schatten van de variogrammen. Deze worden gebruikt voor het optimaliseren van een additioneel bemonsteringsschema voor minimale kriging variantie. De kriging variantie voor het percentage zand wordt zo teruggebracht van 28.2 tot 23.7 (%)². De variogrammen worden ook gebruikt voor bemonsteringsoptimalisatie in een ander, geomorfologisch gelijk gebied. Optimale bemonsteringsschema's voor anisotrope variabelen verschillen sterk van die voor isotrope variabelen. De grootte van de kriging neighbourhood heeft een klein maar duidelijk effect op de bemonsteringsschema's. De schema's reduceren vooral de kriging variantie aan de randen van het gebied.

Hoofdstuk 5 werkt de mogelijkheden van kriging variantie minimalisatie verder uit. Naast het MEAN_OK wordt ook het MAX_OK criterium geïntroduceerd, dat zich richt op minimalisatie van de maximum kriging variantie. Beide criteria worden vergeleken met een regelmatig grid. SSA brengt de gemiddelde kriging variantie terug van 40.64 [unit]² tot 39.99 [unit]². De maximum kriging variantie daalt van 68.83 [unit]² naar 53.36 [unit]². Een additioneel bemonsteringsschema van 10 observaties wordt geoptimaliseerd voor een onregelmatige data set van 100 observaties. Dit reduceert de

kriging variantie van 21.62 [unit]² naar 15.83 [unit]², en de maximum kriging variantie van 70.22 [unit]² tot 34.60 [unit]². Verder wordt de invloed van variogram parameters op het optimale bemonsteringsschema onderzocht. Een Gaussisch variogram model resulteert in een totaal ander bemonsteringsschema dan een exponentieel model met dezelfde nugget, sill en (effectieve) range. Een zeer korte range resulteert in random bemonsteringsschema's, met observaties op een minimum afstand van tweemaal de range. Voor een sferisch variogram heeft de hoogte van het relatieve nugget effect geen effect op het optimale bemonsteringsschema, behalve voor hoge waarden.

Hoofdstuk 6 introduceert het WMSD criterium in SSA, dat bemonsteringsschema's optimaliseert door middel van een ruimtelijke gewichtsfunctie. Dit maakt het mogelijk om onderscheid te maken tussen verschillende prioriteitsgebieden. Dit criterium wordt toegepast op een multivariate verontreinigingsstudie in de haven van Rotterdam met vijf verontreinigende stoffen op twee dieptes. De bemonsteringsstrategie bestaat uit twee fasen met twee verschillende ruimtelijke gewichtsfuncties. In de eerste fase worden eerdere observaties en historische informatie gecombineerd met informatie over de prioriteit van sanering. De metingen van het resulterende schema zijn voor 17.4% verontreinigd, waarvan 1.5% zwaar verontreinigd. In de tweede fase worden overschrijdingskanskaarten van relatief lage drempelwaarden gebruikt om additionele bemonstering te leiden naar mogelijke hot spots. Dit resulteert in 26.7% verontreinigde monsters, waarvan 16.7% zwaar verontreinigd. Hier was een aantal locaties bij die niet in de eerste fase waren aangemerkt als verontreinigd. Het WMSD criterium kan grote waarde hebben bij het ondersteunen van beslissingen.

Bemonsteringsoptimalisatie voor modelschattingen

Hoofdstuk 7 richt zich op het optimaliseren van bemonstering in de precisie landbouw met gebruik van co-gelateerde data. Met behulp van een goedkope, low-tech methode worden oogstkaarten voorspeld van gierst in Niger. Oogst varieert van 0 tot 2500 kg ha⁻¹. SSA wordt gebruikt voor het optimaliseren van 3 verschillende bemonsteringsschema's. Schema 1 optimaliseert verdeling van de observaties over het hele gebied. Schema 2 bestrijkt alle oogstwaarden, en schema 3 bestrijkt alleen de laagproducerende gebieden. Schema 2 resulteert in vijf significante correlaties tussen oogst en bodemvariabelen. Schema 1 vindt slechts één significante correlatie. Schema 2 verklaart 70 % van de oogst-variantie door middel van multivariate lineaire regressie. Schema 1 verklaart slechts 37 %. Verschillen tussen schema 3 en schema 1 zijn significant voor afstand tot struiken, micro-relief, pH-H₂O en CEC. Wij concludeerden dat struiken de belangrijkste factor zijn in de oogstvariabiliteit. Dit komt door het opvangen van geërodeerde deeltjes enerzijds, en het lokaal verbeteren van de bodemvruchtbaarheid anderzijds. Wij concludeerden verder dat de bemonsteringsstrategie van schema 2 moet

worden aanbevolen voor het onderzoeken van oogst/bodem relaties, in het bijzonder bij een laag aantal observaties. De variogrammen van micro-reliëf en oogst suggereren verder dat ruimtelijke correlatie grotendeels beperkt blijft tot afstanden van 3 tot 5 m.

Hoofdstuk 8 evalueert een aantal bemonsteringsstrategieën voor het schatten van variogrammen. Het eerste gedeelte vergelijkt een regelmatig grid met een bemonsterings-schema dat de puntenpaar verdeling optimaliseert. Dit resulteert in zuivere variogram schattingen. De fluctuatie in geschatte variogrammen is echter veel hoger voor het geoptimaliseerde schema. Wij concludeerden hieruit dat de puntenpaar verdeling op zich zelf geen zinnig optimalisatie criterium is voor variogram schattingen. In het tweede gedeelte worden extra observaties geselecteerd voor optimale puntenpaar verdeling, en vergeleken met random getrokken extra observaties. De random getrokken observaties resulteren in een veel hogere standaardafwijking op korte afstand. Hieruit concludeerden wij dat de puntenpaar verdeling een nuttig optimalisatie criterium is voor het selecteren van additionele observaties voor metingen op korte afstand. Het derde gedeelte richt zich op het optimale gebruik van variogrammen. Een bemonsteringsgrid van 81 observaties wordt, na voorlopige schatting van het variogram, aangevuld met 19 extra observaties voor minimale kriging variantie. Dit schema wordt vergeleken met een regelmatig grid van 100 observaties. Voor een exponentieel stochastisch veld zonder nugget wordt de gemiddelde gekwadraterde krigingfout teruggebracht van $0.39 [\text{unit}]^2$ tot $0.31 [\text{unit}]^2$. De maximum gekwadraterde krigingfout daalt van $6.05 [\text{unit}]^2$ naar $4.24 [\text{unit}]^2$. Voor een sferisch stochastisch veld met een nugget effect van 33 % blijft de gemiddelde gekwadraterde krigingfout gelijk, en daalt de maximum gekwadraterde kriging fout van $15.98 [\text{unit}]^2$ naar $11.52 [\text{unit}]^2$. Wij concludeerden dat minimalisatie van de gekwadraterde krigingfout vaak meer relevant is dan het nauwkeurig schatten van het variogram. Het nemen van observaties vlak buiten het eigenlijk onderzoeksgebied verbeterde de kwaliteit van de voorspellingen voor zowel kriging variantie als gekwadraterde krigingfout.

Curriculum Vitae

Jan-Willem van Groenigen was born in Den Helder on September 23th, 1970. He completed secondary school (Atheneum- β) in 1988 at 'Scholengemeenschap De Hoge Berg' on the island of Texel. In 1988, he started his study Soil Science at Wageningen Agricultural University, which he finished 'cum laude' in 1994, with specialisations in soil inventarisation, computer science and philosophy of science. In 1995 he received the C.T. de Witt thesis award for his thesis soil inventarisation. He started working on his Ph.D. research in October 1994 during community service at the Soil Science division of Wageningen Agricultural University. From January 1996 until December 1998 he continued his Ph.D. research at the Soil Sciences division of the International Institute for Aerospace Survey and Earth Sciences (ITC) in Enschede, The Netherlands. In 1998 he received the ITC Research Award. He is currently employed as a scientific staff member at ITC.