QUICS: Quantifying Uncertainty in Integrated Catchment Studies

D3.1 Sample design optimisation techniques and associated software

Lead Partner: Wageningen University

Revision: 23/11/2017

Report Details

Title: Sample design optimisation techniques and associated software

Deliverable Number (If applicable): 3.1

Author(s): Kasia Sawicka, Alexandre Wadoux, Gerard Heuvelink

Dissemination Level: Public

Document History

Version	Date	Status	Submitted by	Checked by	Comment
v 1.0	04/11/2016	Outline Draft	Gerard Heuvelink		Outline and tasks split out by person
v 1.1	07/11/2016	Draft	Alexandre Wadoux	Gerard Heuvelink	Chapter 3, Appendices
v 1.2	14/11/2016	Draft	Kasia Sawicka	Gerard Heuvelink	Chapter 2
v 1.3	28/11/2016	Complete draft	Gerard Heuvelink	Kasia Sawicka and Alexandre Wadoux	First complete draft
v 1.4	28/11/2016	Updated Draft	Kasia Sawicka and Alexandre Wadoux	Gerard Heuvelink	Corrections to draft
v 1.5	30/11/2016	Final Draft	Gerard Heuvelink	Simon Tait	Minor corrections
v 2.0	06/12/2016	Final	Gerard Heuvelink, Kasia Sawicka and Alexandre Wadoux	Simon Tait	Final typos corrected
V 2.1	23/11/2017	Revised Final	Alexandre Wadoux	Will Shepherd	Reproducible R codes provided, updated journal reference

Acronyms and Abbreviations

AKV	Average Kriging Variance
ANNs	Artificial Neural Networks
DEM	Digital Elevation Model
EA	Environmental Agency
EMO	Evolutionary Multi-objective Optimisation
ER	Experienced Researcher
ESR	Early Stage Researcher
GAs	Genetic Algorithms
GLS	Generalised Least Squares
GPP	Gross Primary Production
KED	Kriging with External Drift
LTM	Long-Term Monitoring
MCS	Monte Carlo Simulation
MKV	Mean Kriging prediction Variance

MOGA-ANN	Multi-Objective Genetic Algorithm and Adaptive Neural Networks
NSGAII	Non-Dominated Sorted Genetic Algorithm I
PDF	Probability Density Function or Probability Distribution Function
QUICS	Quantifying Uncertainty in Integrated Catchment Studies
REML	Restricted (or Residual) Maximum Likelihood
SPEA2	Strength Pareto Evolutionary Algorithm 2
SRTM	Shuttle Radar Topography Mission
SSA	Spatial simulated annealing
UCK	Universal Co-Kriging
WDS	Water Distribution System
WSN	Wireless Sensor Network
WU	Wageningen University
εMOEA	Epsilon-Dominance Multi-Objective Evolutionary Algorithm
ε-NSGAII	Epsilon-Dominance Non-Dominated Sorted Genetic Algorithm II

Acknowledgements

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 607000.

Executive Summary

This report:

1. Gives an overview of statistical methods for spatial sampling design optimisation, with a focus on applications in the environmental sciences, including a brief review of key recent publications.
2. Presents a specific application to sampling design optimisation of rain gauge locations for rainfall mapping in space and time, as developed within the QUICS project.
3. Provides a flowchart and computer code of an implementation of the rain gauge location optimisation algorithm in the R software for statistical computing.

CONTENTS

1 Introduction

1.1 Rationale and overview

Sampling is at the core of environmental assessment and involves collecting information about target variables in space and time. Data cannot be collected at all times and at all locations within the population of interest, and hence a subset from the population must be taken. This brings about many questions, such as: What to sample? When and where to sample? How many samples? How does the accuracy of the end result depend on sampling density and sample size? Can a statistical sampling design be optimised? If yes, how? All these questions are also important for the QUICS project, since natural and urban hydrological systems cannot be sampled exhaustively, sampling is expensive and essential to learn about the state of the system.

This deliverable starts with a theoretical chapter on statistical methods for sampling design optimisation. The chapter does not aim to be comprehensive and detailed, its main aim is to give an overview and flavour. It explains that sampling design optimisation basically follows three main steps: 1) description of the system of interest, the variables that characterise the system, the variables that are to be measured, and the type of sampling designs available; 2) definition of a cost criterion that allows judgement of the performance of a sampling design, usually a combination of monetary and lack-of-accuracy costs; 3) presentation of numerical search techniques that can optimise a sampling design by minimising the cost criterion. The chapter concludes with a review of recent applications of sampling design optimisation. Again, the review is not aimed to be comprehensive, but illustrates the concept of sampling design optimisation with real-world cases from environmental and hydrological catchment studies.

The next and final chapter works out one specific sampling design approach in more detail. This chapter is based on a journal manuscript prepared by QUICS fellow ESR2. It introduces the problem of optimisation of rain gauge locations used for mapping rainfall in space and time. In this specific case the optimisation problem is complicated since radar rainfall maps are also used for spatial prediction of rainfall. This implies that the locations of the radar stations influence the optimal rain gauge locations: it pays off to (slightly) increase the rain gauge sampling density in areas where the radar signal is weakened due to radar beam blocking and attenuation. This chapter shows how much the sampling density should increase and how much efficiency gain is achieved by optimising the rain gauge locations. The methodology is illustrated with a case study in the north of England.

The appendices present a flowchart of the sampling design software implementation and corresponding R scripts.

1.2 Partners involved in deliverable

Wageningen University (WU)

1.3 Deliverable objectives

The European project QUICS (Quantifying Uncertainty in Integrated Catchment Studies) collates 12 PhD Candidates (Early Stage Researchers, ESR) and four postdocs (Experienced Researchers, ER) to perform quality research and collaborate with each other for developing and implementing uncertainty analysis tools for Integrated Catchment Modelling.

The objectives of QUICS Deliverable 3.1 are:

1. Provide a review of statistical sampling design optimisation techniques.
2. Provide a description of a rule-base and associated software that generates the optimal sampling design for hydrological catchment studies.

2 Overview of statistical sampling design optimisation approaches

Sampling is at the core of environmental assessment and involves collecting information about a target variable in space, time, or in both space and time. Data cannot be collected at all times and at all locations within the domain of interest, however ancillary information related to the target variable may be exhaustively available (e.g. remotely sensed information, digital elevation model, outcomes of process models) and guide the sampling design. Sampling design is an important consideration to ensure that the data collected is as informative and accurate as possible, given the available sampling budget. In planning a sampling strategy, it is necessary to consider the intended use of the sampling data in the first place. Sampling design optimisation requires means to quantify the quality of measurements obtained and must take any constraints into account. Statistical and mathematical methods are most commonly applied for sampling design optimisation because these provide a more objective way to quantify errors in the result. In this chapter an overview of existing statistical sampling design approaches is presented.

2.1 Definition of a sampling design and sampling design considerations

In integrated catchment studies it is of importance to design networks set up to monitor environmental variables using a series of measuring stations, without reference to graphs or connecting linkages between monitoring stations. Hence, in QUICS we adapted the term ’sampling design optimisation’ rather than ’network optimisation’ to describe approaches to finding an optimal solution for sampling, given a target variable or quantity of interest. The term ’monitoring network’ or simply network will be used to indicate the current locations of monitoring stations.

Primarily, the sampling design should reflect the variation of the target variable in the study area (Heuvelink et al., 2006, Brungard and Boettinger, 2010). Suggested strategies infer sampling in the geographical space (Brus et al., 2006), in the variable (e.g. soil or water) related covariate space (Minasny and McBratney, 2006), or in a combination of both (Dobermann et al., 2006). Secondly, the sampling design should support field operability in terms of constrained accessibility, e.g. due to difficult terrain and restricted areas (Stumpf et al., 2016). Third, the sampling design should incorporate available legacy data and information to accommodate the demand on reducing high labour and monetary costs for sampling and laboratory analysis (Lagacherie, 2008). Yet, a spatial mismatch of statistically predefined sample sites, a lack in harmonisation with the target variable, and different spatial resolutions, formats and objectives remain problems when incorporating existing data into sampling designs (Carré et al., 2007, Krol, 2008, Sulaeman et al., 2013). A further possibility to increase the efficiency in data acquisition comprises an optimised sample set size. Few studies addressed this issue by comparing model results based on different calibration set sizes (Brungard and Boettinger, 2010, Ramirez-Lopez et al., 2014, Schmidt et al., 2014).

2.2 Aspects of the selection of a sampling design optimisation scenario

Table 1 presents the main aspects involved in the selection of a sampling design optimisation scenario, derived in part from work by de Gruijter et al. (2006). Many of the decisions shown in Table 1 have different theoretical and statistical implications as elaborated below.

Table 1 Main aspects to consider in selecting sampling design optimisation scenarios.

Aspects	**Description**	**Example(s)**
Target (universe, variable, quantity)	A precise definition of the spatial and temporal extent, unit of measure, i.e. quantitative or qualitative, variable to be determined in each sampling unit, or type of statistic needed.	Soil type, surface water quality, rainfall, air quality, number of households per km^2.
Constraints	Issues and conditions that prevent estimation or prediction of the target with minimal error; these are generally budgetary constraints (i.e. limiting the number of measurements stations).	Budget, technical limitations, operational or access limits, knowledge limits, non-response, etc.
Criterion	Objective function or quantity used to represent the statistical quality of the result. Criteria are used to compare different sampling designs.	Mean error, root mean squared error, average kriging prediction error variance, false discovery rate.
Approach	Method oriented means of representing or modelling reality with major consequences for sampling and inference. The criterion generally follows from the approach.	Design-based, geometrical, geostatistical model based.
Algorithm	Description of the mathematical solution used to optimise the criterion and compute the optimal sampling design.	Exhaustive search, evolutionary algorithms, (spatial) simulated annealing.

2.2.1 Target

A target is the variable or quantity of interest. The scope of the target will be limited by the extent of the sampling network or the domain under investigation. The act of measurement may also limit the scope or observational power of the target data due to technical limitations, sampling heterogeneities, and sampling errors. The target of a monitoring network may be a parameter in a statistical model, such as a trend parameter or the mean of the sampled distribution. Targets such as these that summarise the variable of interest over the entire spatial or temporal domain are global quantities, whereas targets that relate to point locations are referred to as local quantities. The decision to focus on global or local targets will determine the statistical methods that can be used. In addition, whether or not the target is quantitative or qualitative also determines the mode of statistical inference. With quantitative variables, both statistical estimation and prediction are possible. But with qualitative variables, such as whether or not a given toxin is detected above a threshold level or not, then the mode of statistical inference is generally limited to testing, classification and detection (de Gruijter et al., 2006).

Non-linear regression methods, such as logistic and Poisson regression, can be used with categorical, qualitative data (e.g. presence/absence and count data), permitting estimation or inference about the parameters of a deterministic regression relationship between the response and explanatory variables. Hence, the distinction between estimation and prediction is a subtle one and is perhaps most related to geostatistical or model-based methods that rely on the theory of spatial random fields (i.e. stochasticity in the spatial variation of the pattern, described further below). In geostatistical methods, estimation refers to inference about the parameters of a stochastic model: these parameters may be related to the regression relationship estimated using generalised least squares, or related to the parameters of a spatial model describing the covariance structure of the spatial random field (Diggle and Ribeiro, 2007). Geostatistical prediction, on the other hand, refers to inference about a realisation of the unobserved stochastic process at point or block support locations (Diggle and Ribeiro, 2007). Prediction using regression models refers to inference about the unknown, yet determinate pattern, which is predicted using parameter estimates of the underlying correlated relationships (i.e. between the response and predictor variables).

2.2.2 Constraints

The notion of constraints will often differ depending on the institutional perspective and approach to optimisation. Many existing monitoring networks have been designed on the basis of heuristic constraints and principles, i.e. prior experience, convention, prescribed protocols, expedience, and operational considerations. For instance, sampling locations may be appointed to represent larger areas, without explicitly defining an optimality criterion (Lophaven, 2002).

The most easily identified constraints are budgetary ones, because budgetary constraints limit sample size and measurement accuracy. Constraints are essentially connected to the definition of a statistical criterion because they place limits on where or when the criterion may be estimated. Indeed, the criterion or objective function itself may be defined in terms of 'cost minimisation' (de Gruijter et al., 2006, Heuvelink et al., 2010). However, constraints can also be related to different perspectives. Some institutions may place greater or lesser emphasis on operational issues (e.g. road access or the correct placement and maintenance of measuring instruments). Constraints can also be related to pre-emotive decisions about the maintenance of a minimum number of sampling stations in areas of political, ecological, or social interest, to the exclusion of other measurement areas.

2.2.3 Criterion

The criterion is a mathematical representation of a deciding factor, i.e. one needs to be able to decide why one design is better than the other (e.g. network quality in terms of prediction accuracy or the accuracy with which model parameters are estimated in a calibration procedure). The criterion - also referred to as the objective function - is quantitative and thus can be minimised or maximised (Ehrgott, 2005). The definition of a criterion, however, is not a panacea for the optimisation problem. Using a single statistical optimality criterion may not encompass all of the different aspects of a monitoring network due to conflicting objectives and differing definitions of

optimality. Moreover, the aim(s) of a monitoring network are often formulated in broad terms, making translation into a more stringent optimality formulation difficult (Lophaven, 2002).

2.2.4 Approaches

Sampling strategies can be divided into two main groups as noted for example by Brus and de Gruijter (1997). The first group, referred to as **design-based** methods, rely on probability-based sampling, wherein the units are selected randomly according to their known selection probabilities, and statistical inference is based on the probabilistic sampling design. The second group is known as **model-based** sampling, which relies on a pre-specified model of spatial variation. Here measurement values are considered as a realisation of a random field.

The distinction between design-based and model-based methods is based on the way in which the two approaches regard randomness in the environment. In design-based methods the target variable is unknown, but assumed to be deterministic or fixed rather than stochastic. Hence, uncertainty is evaluated by repeated sampling, i.e. of the same pattern with different sampling locations (de Gruijter et al., 2006), which is akin to probability-based sampling e.g. the 'coin tossing experiment' (Brus and de Gruijter, 1997). In model-based methods, uncertainty is estimated by repeatedly sampling a fixed set of sampling locations, assuming that the pattern of values in the area is stochastic or unfixed. The model-based approach bases inference on the underlying spatial or temporal statistical model, thus the selection of sampling locations is purposive and not random.

The main questions that must be addressed prior to selecting a design-based or model-based based approach are (de Gruijter et al., 2006):

1. Must the test on the target quantity be unbiased objectively, i.e. without the recourse of a specified model?
2. Should the accuracy of the test be quantified objectively?
3. Is random sampling feasible?
4. Is a reliable model available?
5. Do substantial spatial autocorrelations exist?

A third group of approaches is introduced in the following classification of perspectives: sampling based purely on geometrical concerns. These approaches are grouped under the label 'geometric designs'. They can be assimilated under the model-based approach because although they do not rely on a known spatial model (or covariance function), data collected using a geometric design can be used in a model-based context. Geometric approaches do not rely on probabilistic considerations, and the design of sampling schemes are not random in their construction. Geometric designs can be viewed as an alternative design option when utility measures or statistical measures are not feasible. These designs are the most simple in the sense that they are only based on geometrical constraints.

The approach used will depend mainly on the assumptions that can be made in calculating the result. Lophaven (2002) suggests that exploratory designs (e.g. geometric designs) should be used prior to optimisation with an assumed geostatistical model – quite independently from the objectives of the design. Design-based approaches may be favoured in cases where: estimation of the target variable must be assessed over large areas; it is important to obtain a more objective

assessment of uncertainty in the estimate; and where the estimate required is a statistic of the frequency distribution of the target variable. Model-based approaches are more appropriate for mapping or prediction of local point locations.

2.2.5 Algorithms

Mathematical algorithms are used to find optimal solutions for given optimisation problems. In this report, we can discuss only a fraction of the wide variety of optimisation algorithms available. Analytical solutions to optimisation problems are possible only in the simplest of cases, whereas 'exhaustive search' approaches only work when the search space can be efficiently explored and when the dimensionality of the search space is low. Empirical, or simulation-based approaches to finding optimal solutions trade confidence in finding the optimal solution for shorter run-times and the guarantee of a near optimal solution.

2.3 Classification and description of scenarios

The envisaged scenarios are classified in Table 2. For the following, we consider that an existing network is already installed in the region of interest and has to be optimised. The three main types of approaches are assigned to a binary category (+/–) depending on whether or not the target can be well-estimated with the approach. Classes of scenarios are given a code and are then described in the following section.

There are two main types of targets:

- *global*: linear global quantities (e.g. mean, variance over areas), non-linear global quantities (e.g. percentiles, proportion above threshold);
- *local*: linear (e.g. local estimation, mean, variance) and non-linear (e.g. proportion above threshold).

The classification of targets leads in the different approaches to an optimisation criterion. As was noted above about constraints, one can generally recall to the quality optimisation scenario.

Table 2 Classification of sampling design approaches (rows) and targets (columns) with corresponding envisaged scenarios.

Scenarios		Classification of target	
		Global (Lin/Non-lin)	Local (Lin/Non-lin)
Approach	Design based	A (++)	D (– –)
	Geometrical	B (+)	E (– –)
	Model based (geostatistical)	C (+)	F (++)

In Chapter 3 we present in detail a practical example using scenario F.

2.4 Review of numerical techniques that produce the optimal sampling design

Many different algorithms can be used to optimise spatial network designs. For spatial interpolation of hydrological and other environmental variables in routine and emergency situations, computation time and interpolation accuracy are important criteria. In this section we compare four different optimisation algorithms for both criteria.

2.4.1 Greedy algorithm based on entropy

Maximum entropy sampling is based on Shannon's concept of information. For a continuous random field with probability density $f(z)$, its information is defined as $E(\log(f(z)))$, and its entropy as $H(f) = -\int f(z)\log(f(z))\,dz$, i.e. information is negative entropy. In the spatial sampling context, the goal is to minimise the conditional entropy of the random field at the unobserved locations with respect to that at the observation locations. This is equivalent to maximising the entropy of the random field at the observation locations (Le and Zidek, 2006, Krause et al., 2008). For a Gaussian random field this leads, in turn, to maximising the determinant of the covariance matrix of the observations, see Caselton and Zidek (1984), Shewry and Wynn (1987), and Gebhardt (2003).

Baume et al. (2011) applied this criterion to the case of sampling from a grid of potential sites, which is split into two disjoint subsets: the design points at which the random field is observed and the complementary set. Shewry and Wynn (1987) proposed an exchange-type algorithm to find the optimal design. Their iterative procedure converges, but does not necessarily lead to an optimum. Ko et al. (1995), and Lee and Williams (2003) developed branch and bound methods, which, under certain conditions, lead to the global optimum. However, the computational complexity of these methods make their practical implementation computationally prohibitive when it comes to choosing several dozens of design points from a grid of several thousands of potential sites. Krause et al. (2008) developed a polynomial-time algorithm that is within $(1 - 1/e)$ of the optimum by exploiting the sub-modularity of mutual information, and design branch and bound procedures with efficient online bounds.

Greedy algorithms start with a non-feasible solution, either with too few (greedy algorithm) or too many (dual greedy algorithm) measurements. At each step, these algorithms select the design which leads to the maximum increase in entropy (when adding a new measurement, greedy algorithm) or to the minimum decrease in entropy (when deleting an existing measurement, dual greedy algorithm). This is known as D-optimality. This means that these only deliver approximate solutions. Exact optimisation can only be achieved by the branch and bound method. In many cases in Gebhardt (2003), it turns out that the initial greedy solutions were already optimal or at least near to optimal. But even greedy algorithms are limited by the data size, as with growing size of datasets also the covariance matrices get larger and their determinants may become numerically instable.

2.4.2 Spatial simulated annealing

The spatial simulated annealing algorithm has five main steps (Brus and Heuvelink, 2007):

1. start with an arbitrary initial design;
2. compute a candidate new design from the current design by random perturbation of the locations of one or several measurement sites;
3. evaluate the new candidate design with the chosen criterion (MKV);
4. accept the new design when the criterion has improved, or accept it with some probability when the criterion has deteriorated;
5. stop when a given (large) number of iterations have been done or when new candidate designs have not been accepted for a given number of times.

Simulated annealing requires several parameters to be defined. The probability of accepting worsening designs usually decreases as the iteration progresses (so-called 'cooling' schedule). This requires a choice of the initial probability and the manner in which it decreases. A stopping criterion of the optimisation procedure is also a key parameter to avoid a too long procedure. The selection of the best value for these parameters is largely dependent on the specificities of each case. In spatial simulated annealing, the perturbation of locations is controlled by specifying the maximum distance over which locations may be displaced. Typically, the maximum distance decreases as the iteration progresses. See Brus and Heuvelink (2007) and the references therein for more details. Heuvelink et al. (2013) generalised this method to a case of optimising the design of a space-time meteorological network.

2.4.3 Spatial coverage

The spatial coverage optimisation method targets at a geometrical criterion. Geometrical criteria are based only on the spatial configuration of the measurement locations and not on the measurement values or underlying geostatistical model. Spatial coverage algorithms are more often used in the context of design-based sampling strategies to estimate global quantities such, as the global mean (de Gruijter et al., 2006).

The spatial coverage approach can also be used to expand or thin an existing design. In a scenario where measurement locations are added, the spcosa method algorithm developed by Brus et al. (2006) may be used. The R package supporting the method is also called *spcosa*. Their method is based on the mean squared distance criterion, which allows optimisation with k-means clustering. In the case of deleting measurements from an initial dataset (thinning), it is advised to use the definition of coverage as in (Royle and Nychka, 1998). The heuristic search to select stations to be deleted from the network is a point swapping algorithm, similar to the one used in greedy algorithms.

2.5 Examples of sampling design optimisation publications

In this section we give an overview of recent applications of sampling design optimisation. We note that we do not aim for a comprehensive literature review, but would like to illustrate the concept of sampling design optimisation with different examples from environmental and catchment studies.

2.5.1 Applications in integrated catchment studies and urban hydrology

Since the degree of uncertainty in model parameters depends on the number and configuration of calibration data, sampling design optimisation can help reduce the variance of model parameters, for example complex integrated catchment models. Sampling design optimisation is also useful for validation of the results of an uncertainty propagation analysis. Here, we present five examples of application of the sampling design optimisations in integrated catchment modelling and urban hydrology studies:

1) *Optimised selection of river sampling sites* – in integrated catchment studies the managers of catchment water quality monitoring programs are responsible for considerable expenditure of funds and effort and the selection of river sampling sites ranks highly among their tasks. The optimum selection of sampling sites is related to the objective of the program, whether it is, for example, trend detection, regulatory enforcement, or estimation of pollutant loadings. Sampling programs are often required, however, to fulfil several roles or may have constraints, which make the manager's choice more difficult. An example study of Dixon et al. (1999) presents a methodology for optimising the selection of river sampling sites. The paper describes procedures using a geographical information system (GIS), graph theory and a simulated annealing algorithm. Three case studies were included which demonstrate the use of the methodology in (i) a simple regulatory monitoring situation, (ii) a situation where possible sampling sites are severely restricted and (iii) for monitoring an impounding catchment with problem inflows. Optimisation of sampling site location by simulated annealing is shown to be adaptable to a variety of practical situations and to perform better than the algorithmic method previously published by Sharp (1971).

2) *Evolutionary multi-objective algorithms for long-term groundwater monitoring design* – in a study by Kollat and Reed (2006) the performances of four state-of-the-art evolutionary multi-objective optimisation (EMO) algorithms are compared: the Non-Dominated Sorted Genetic Algorithm II (NSGAII), the Epsilon-Dominance Non-Dominated Sorted Genetic Algorithm II (ε-NSGAII), the Epsilon-Dominance Multi-Objective Evolutionary Algorithm (εMOEA), and the Strength Pareto Evolutionary Algorithm 2 (SPEA2), on a four-objective long-term groundwater monitoring (LTM) design test case. The performances of the four algorithms were assessed and compared using three runtime performance metrics (convergence, diversity, and ε-performance), two unary metrics (the hyper-volume indicator and unary ε-indicator) and the first-order empirical attainment function. Results of this analyses indicated that the ε-NSGAII greatly exceeds the performance of the NSGAII and the εMOEA. The ε-NSGAII also achieves superior performance relative to the SPEA2 in terms of search effectiveness and efficiency. In addition, the ε-NSGAII's simplified parameterisation and its ability to adaptively size its population and automatically terminate results in an algorithm which is efficient, reliable, and easy-to-use for water resources applications.

3) *Water distribution system optimisation using metamodels* - Genetic algorithms (GAs) have been shown to apply well to optimising the design and operations of water distribution systems (WDSs). The objective has usually been to minimise cost, subject to hydraulic constraints such as satisfying minimum pressure. More recently, the focus of optimisation

has expanded to include water quality concerns. This added complexity significantly increases the computational requirements of optimisation. Considerable savings in computer time can be achieved by using a technique known as metamodeling. A metamodel is a surrogate or substitute for a complex simulation model. Broad et al. (2005) used the metamodeling approach to optimise a water distribution design problem that includes water quality. The type of metamodels used were artificial neural networks (ANNs), as they are capable of approximating the nonlinear functions that govern flow and chlorine decay in a WDS. The ANNs were calibrated to provide a good approximation to the simulation model. In addition, two techniques are presented to improve the ability of metamodels to find the same optimal solution as the simulation model. Large savings in computer time occurred from training the ANNs to approximate chlorine concentrations (approximately 700 times faster than the simulation model) while still finding the optimal solution.

4) *Multi-objective design of water distribution systems under uncertainty* - The WDS design problem can be defined as a multi-objective optimisation problem under uncertainty. The two objectives are: (1) minimise the total WDS design cost, and (2) maximise WDS robustness. In Kapelan et al. (2005) the WDS robustness is defined as the probability of simultaneously satisfying minimum pressure head constraints at all nodes in the network. Decision variables are the alternative design options for each pipe in the network. They identified that the sources of uncertainty are future water consumption and pipe roughness coefficients. Uncertain variables were modelled using probability density functions (PDFs) assigned in the problem formulation phase. The optimal design problem is solved using the newly developed RNSGAII method based on the NSGAII algorithm. In RNSGAII a small number of samples are used for each fitness evaluation, leading to significant computational savings when compared to the full sampling approach. This methodology was tested on several cases, all based on the New York tunnels reinforcement problem. The results obtained demonstrated that the new methodology is capable of identifying robust Pareto optimal solutions despite significantly reduced computational effort.

5) *Stochastic sampling design for water distribution model calibration* - Behzadian (2008) showed an approach to determine optimal sampling locations under parameter uncertainty in a WDS for the purpose of its hydraulic model calibration. The problem was formulated as a multi-objective optimisation problem under calibration parameter uncertainty. The objectives were to maximise the calibrated model accuracy and to minimise the number of sampling devices as a surrogate of sampling design cost. Model accuracy was defined as the average of normalised traces of model prediction covariance matrices, each of which is constructed from a randomly generated sample of calibration parameter values. To resolve the computational time issue, the optimisation problem was solved using a multi-objective genetic algorithm and adaptive neural networks (MOGA-ANN). The verification of results was done by comparison of the optimal sampling locations obtained using the MOGA-ANN model to the ones obtained using the Monte Carlo Simulation (MCS) method. In the MCS method, an equivalent deterministic sampling design optimisation problem was solved for a number of randomly generated calibration model parameter samples. The results of that study showed that significant computational savings can be achieved by using MOGA-ANN

compared to the MCS model or the GA model based on all full fitness evaluations without significant decrease in the final solution accuracy.

2.5.2 Applications in other environmental studies

Sampling design optimisation is widely used in various environmental studies. Here we summarise five examples across different disciplines of environmental sciences:

1) *Developing an optimal sampling design for a coastal marine ecosystem study* - Kitsiou et al. (2001) presented the development of a sampling design for optimising sampling site locations collected from a coastal marine environment. They used a dataset that included data collected from 34 sampling sites spaced out in the Strait of Lesbos, Greece, arranged in a 1×1 NM grid. The coastal shallow ecosystem was subdivided into three zones, an inner one (7stations), a middle one (16 stations) and an offshore zone (11 stations). The standard error of the chlorophyll-a concentrations in each zone has been used as the criterion for the sampling design optimisation, resulting into reallocation of the sampling sites into the three zones. The positions of the reallocated stations have been assessed by estimation of the spatial heterogeneity and anisotropy of chlorophyll-a concentrations using variograms. Study of the variance of the initial dataset of the inner zone taking into account spatial heterogeneity, revealed two different sub-areas and therefore, the number of the inner stations has been reassessed. The proposed methodology eliminated the number of sampling sites and maximised the information of spatial data from marine ecosystems. The paper includes a step-by-step procedure that could be widely applied in sampling design concerning coastal pollution problems.

2) *Sampling design optimisation for multivariate soil mapping* - Sampling design optimisation is also used in a growing field of soil mapping. For example, Vašát et al. (2010) presented a method, implemented as R-code, that minimises the average kriging variance (AKV) for multiple soil variables simultaneously. The method was illustrated with real soil data from an experimental field in central Czech Republic. The goal of the method was to minimise the sample size while keeping the AKV values of all tested soil variables below given thresholds. They defined and tested two different objective functions, critical AKV optimisation and weighted sum of AKV optimisation, both based on the AKV minimisation with annealing algorithm. The crucial moment for such an optimisation was to define the mutual spatial relationship between all soil variables with the Linear Model of Coregionalisation and proper modelling of all (cross)variograms which are used in the optimisation process. In addition, a separate optimisation was made for each of the tested soil characteristics to evaluate a possible gain of the simultaneous approach. The results showed that the final design for multivariate sampling is “fully-optimal” for one soil variable - optimal number of observations and optimal structure of sampling pattern, and “sub-optimal” for the others, while no clear difference between the two optimisation criteria was found. The presented methods can be used therefore in situations where periodical soil surveys are planned and where multivariate soil characteristics are determined from the same soil samples at once (i.e. same point in time).

3) *Optimising the spatial pattern of networks for monitoring radioactive releases* - Melles et al. (2011) optimised the permanent network of radiation monitoring stations in the Netherlands and parts of Germany as an example. The optimisation method proposed combines minimisation of prediction error under routine conditions with maximising calamity detection capability in emergency cases. To calculate calamity detection capability, an atmospheric dispersion model was used to simulate potentially harmful radioactive releases. For each candidate monitoring network, it was determined if the releases were detected within one, two and three hours. Four types of accidents were simulated: small and large nuclear power plant accidents, deliberate radioactive releases using explosive devices, and accidents involving the transport of radioactive materials. Spatial simulated annealing (SSA) was used to search for the optimal monitoring design. SSA was implemented by iteratively moving stations around and accepting all designs that improved a weighted sum of average spatial prediction error and calamity detection capability. Designs that worsened the multi-objective criterion were accepted with a certain probability, which decreased to zero as iterations proceeded. This study presents a method to optimise the sampling design of environmental monitoring networks in a multi-objective setting. Results were promising and the method should prove useful for assessing the efficacy of environmental monitoring networks designed to monitor both routine and emergency conditions in other applications as well.

4) *Spatial sampling design for estimating regional gross primary production with spatial heterogeneities* - the estimation of regional gross primary production (GPP) is a crucial issue in carbon cycle studies. One commonly used way to estimate the characteristics of GPP is to infer the total amount of GPP by collecting field samples. In this process, the spatial sampling design will affect the error variance of GPP estimation. In one of the studies tackling this challenge, Wang et al. (2014) used geostatistical model-based sampling to optimise the sampling locations in a spatial heterogeneous area. The approach was illustrated with a real-world application of designing a sampling strategy for estimating the regional GPP in the Babao river basin, China. By considering the heterogeneities in the spatial distribution of the GPP, the sampling locations were optimised by minimising the spatially averaged interpolation error variance. To accelerate the optimisation process, a spatial simulated annealing search algorithm was employed. Compared with a sampling design without considering stratification and anisotropies, the proposed sampling method reduced the error variance of regional GPP estimation.

5) *Sampling design optimisation of a wireless sensor network for monitoring eco-hydrological processes* - optimal selection of observation locations is an essential task in designing an effective eco-hydrological process monitoring network, which provides information on eco-hydrological variables by capturing their spatial variation and distribution. Ge et al. (2015) presented a geostatistical method for multivariate sampling design optimisation, using a universal co-kriging (UCK) model. The approach was illustrated by the design of a wireless sensor network (WSN) for monitoring three eco-hydrological variables (land surface temperature, precipitation and soil moisture) in the Babao River basin of China. After removal of spatial trends in the target variables by multiple linear regression, variograms and cross-variograms of regression residuals were fit with the linear model of

coregionalisation. Using weighted mean UCK variance as the objective function, the optimal sampling design was obtained using a spatially simulated annealing algorithm. Their results demonstrated that the UCK model-based sampling method can consider the relationship of target variables and environmental covariates, and spatial auto- and cross-correlation of regression residuals, to obtain the optimal design in geographic space and attribute space simultaneously. Compared with a sampling design without consideration of the multivariate (cross-)correlation and spatial trend, the proposed sampling method reduces prediction error variance. The optimised WSN design has been shown to be efficient in capturing spatial variation of the target variables and for monitoring eco-hydrological processes in the Babao River basin.

3 Application to sampling design optimisation of rain gauges for space-time rainfall interpolation

Accurate information about the space-time distribution of rainfall is essential for hydrological modelling, both in natural and urban environments. Rain-gauge rainfall measurements are accurate and have high temporal resolution, but they typically have a low spatial density and are therefore unable to account for the high spatial variability of rainfall. In contrast, weather radar imagery provides a fuller spatial coverage of the rainfall field in combination with sufficiently high temporal resolution. However, radar-derived rainfall predictions experience complex spatio-temporal disturbances and can be inaccurate. Over the past years, several attempts have been made to combine the strengths of the two measurement devices with numerous geostatistical approaches, but the accuracy of the predicted rainfall maps are dependant of the rain gauge network location, which have to be optimally defined.

In this chapter we propose a method to optimise the sampling design for space-time mapping of rainfall. The model parameters (regression coefficients and correlogram parameters) are estimated from the rain-gauge data. The correlogram parameters and regression coefficients for the residual standard deviation are estimated by Restricted Maximum Likelihood. Finally, we optimise the rain-gauge locations through minimising the prediction error variance averaged over space and time with SSA. The model is tested in a case study in the North of England in the United Kingdom and the rain-gauge pattern is optimised for the year 2010 for daily rainfall mapping.

This chapter presents the underlying methodology and application. Chapter 4 shows a flowchart of the software implementation, while the computer code itself is presented in Appendix A.

This chapter is derived from a recently submitted journal manuscript: Wadoux, A.M.J-C., Brus, D.J., Rico-Ramirez, M.A., Heuvelink, G.B.M., 2017. Sampling design optimisation for rainfall prediction using a non-stationary geostatistical model. Adv. Water Resour. 107, 126–138. doi:10.1016/j.advwatres.2017.06.005

3.1 Study area and data

The study area is located in the United Kingdom, North-East of the city of Manchester. The area is 27 874 km^2 in size and contains several hydrological catchments of different sizes and shapes. Two rainfall datasets are used in this study, rain gauges and radar-derived rainfall maps. The area is covered by a network of 229 tipping bucket rain gauges operated by the Environment Agency (EA). The data originally provided by the EA are at 15-min resolution and were aggregated to daily values. We checked the quality of the data from the available gauges and reduced the number of gauges to 185, excluding the ones having anomalies (e.g. long negative time series or excessive missing data).

The radar composite imagery is obtained from the MetOffice NIMROD system. The system makes use of three radars (Hameldon Hill, Ingham and High Moorsley) shown in Figure 3.1 and has 1 km^2 spatial resolution. The pre-processing of the weather radar data includes removal of non-meteorological echoes (e.g. ground clutter, ground echoes due to anomalous propagation), correction for antenna pointing, correction for beam blockage, rain attenuation correction, vertical reflectivity profile correction and rain-gauge adjustment. The radar rainfall product is available with

a spatial resolution of 1 km and a temporal resolution of 5 min. The 5 min resolution images were aggregated to daily values.

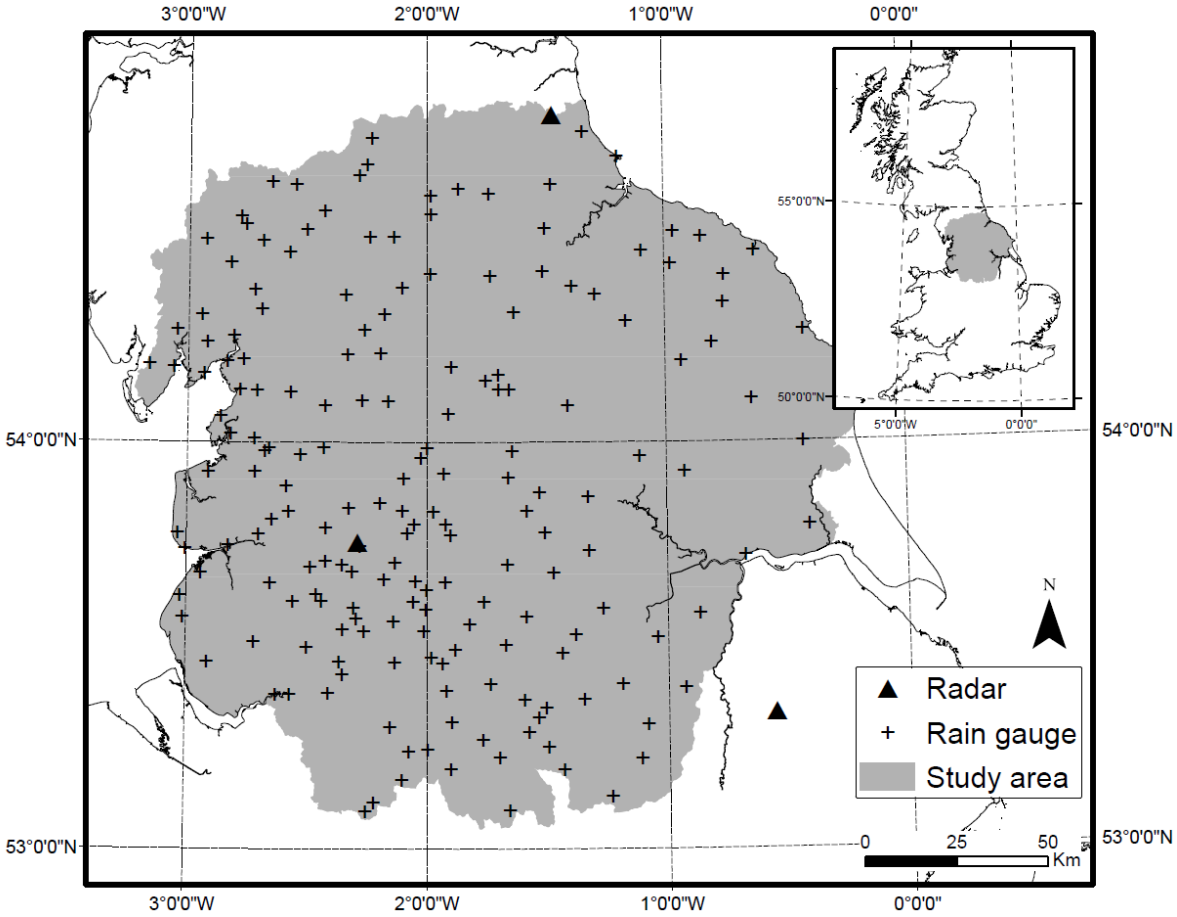


Figure 3.1: Map of the study area with locations of the rain gauges and radars.

Besides these two datasets on rainfall, the following datasets of covariates were used:

- Digital Elevation Model (DEM) at 50 m resolution from the SRTM (Shuttle Radar Topography Mission). The elevation ranges from 6 m to 926 m above sea level.
- Radar beam blockage map at 1 km resolution. The radar beam blockage maps were generated for each radar using the DEM at 50 m resolution and the ground clutter model. The individual beam blockage maps were combined to produce a single map with 1 km^2 resolution for the 0.5-degree radar scan inclination. The blockage maps represent the degree of systematic error of the 0.5 degree inclination of the radar due to topographic obstacles. The values are expressed in percentage from 0 to 100. Their mean is 4.8 %.
- Distance from the radar map at 1 km resolution. The values are expressed in meters and vary from 0 (radar location) to 102554 m. The mean is 51300 m.

3.2 Model Definition

Daily rainfall as measured by rain gauges Z at any location s in the study area A is modelled at a time t by:

$$Z_t(s) = m_t(s) + \sigma_t(s) \cdot \varepsilon_t(s) \tag{3.1}$$

where $m_t(s)$ is the spatial trend, $\sigma_t(s)$ the spatial standard deviation and $\varepsilon_t(s)$ the zero-mean, unit variance, normally distributed, second order stationary and spatially correlated standardised residuals at location s. Note that both the trend and the standard deviation vary in space. They are modelled as a linear combination of covariates:

$$Z_t(s) = \sum_{k=0}^{K} \beta_{t,k} f_{t,k}(s) + \sum_{l=0}^{L} \alpha_{t,l} g_{t,l}(s) \cdot \varepsilon_t(s) \tag{3.2}$$

where the $\beta_{t,k}(s)$ are regression coefficients and the $f_{t,k}(s)$ are covariates for the mean, ($f_{t,0}(s)$ equals 1, so that $\beta_{t,o}$ is the intercept), $\alpha_{t,l}$ (s) are regression coefficients and $g_{t,l}(s)$ covariates for the standard deviation (again $g_{t,0}$equals 1). The covariates $f_{t,k}$ and $g_{t,l}$are time variant dependant of time t.

For more details about the method of non-stationary variance, we refer to Wadoux et al. (2017). Two subsets of model parameters must be estimated, β_k with the regression coefficients for the spatial trend, and the covariance structure parameters, i.e. all parameters for the stochastic part of the model. Given the covariance structure parameters, the estimation of β_k is straightforward and can be done analytically by Generalised Least Squares (GLS). The solution is to make use of Restricted (or Residual) Maximum Likelihood (REML). Similar to Maximum Likelihood, REML aims to find the vector of parameters for which the observed data have the highest probability density (likelihood).

3.3 Sampling design optimisation

We suppose that, due to budget constraints, the number of rain-gauges is fixed. We therefore only have influence on the locations of the rain gauges. The kriging variance only depends on the sampling locations, the correlogram, the trend and the standard deviation covariates. The sampling locations for spatial prediction by kriging can be optimised if the covariance structure is known. With the spatially averaged KED (Kriging with External Drift) variance we achieve a proper balance between optimisation in geographic and feature space. Using space-time data we propose to minimise the KED variance averaged over space and over the times as criterion, on which observations were taken to find the optimal spatial sampling design. Note that using this criterion implies that the gauge locations are static, i.e. they do not move through the area over time. The alternative would have been to optimise for each day separately using the spatially averaged KED variance, which would lead to a dynamic spatial design. We consider this an impractical space--time design.

In theory, with a finite number of possible gauge locations, we could try all possible sampling design combinations, and choose the one that minimises the criterion. However, finding the optimal gauge network in this way is practically impossible given the exorbitant number of possible

combinations, even with a coarse discretisation of the study area. We used a spatial numerical search algorithm instead, in our case SSA.

Spatial simulated annealing is an iterative optimisation algorithm in which a sequence of samples is generated (see Chapter 2). A new sample is derived by selecting randomly one location and shifting this location to a new location across a random distance and in a random direction. Each time a new sample is generated, the criterion is evaluated and compared with the value of the previous sample. The new sample is always accepted if the average KED-variance is smaller. If the criterion is larger the new sample is sometimes accepted, with probability.

3.4 Results

Having built the model for each day of the year 2010, we optimise the static configuration of rain-gauge locations for the year 2010. Figure 3.2 shows the decrease of the prediction error variance as the sampling is re-organised. The graph shows that several worse designs are accepted at the beginning but the prediction error variance finally decreases and stabilises. No substantial reduction is made after 20,000 iterations, suggesting that the algorithm reached a nearly optimum design. Surprisingly a slight decrease at the end is observed. The criterion (i.e., the space-time averaged kriging variance) diminishes from 5.7 to 5.3, which represents a decrease of 4.6 %.

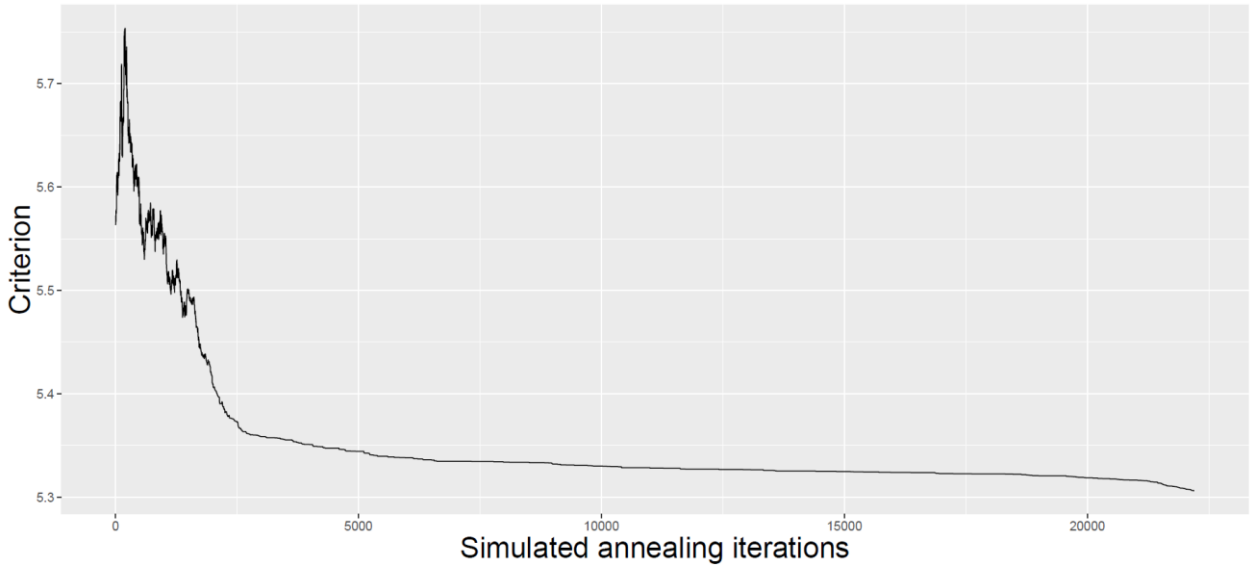


Figure 3.2: Convergence of the criterion with increasing SSA iterations (total number of iterations equals 22 200)

Figure 3.3 presents the initial and the optimised sampling location of the rain gauges for comparison. The optimised pattern shows a dense spatial coverage pattern in the East and in a North-South band. The study area reflects some almost empty space where only a few rain-gauges are present. The optimised design has a fairly uniform spatial coverage of rain gauges. However, more rain gauges are present where the radar is inaccurate. We refer to Wadoux et al. (2017) for more detailed description and interpretation of results and maps of the covariates used to model the trend and standard deviation.

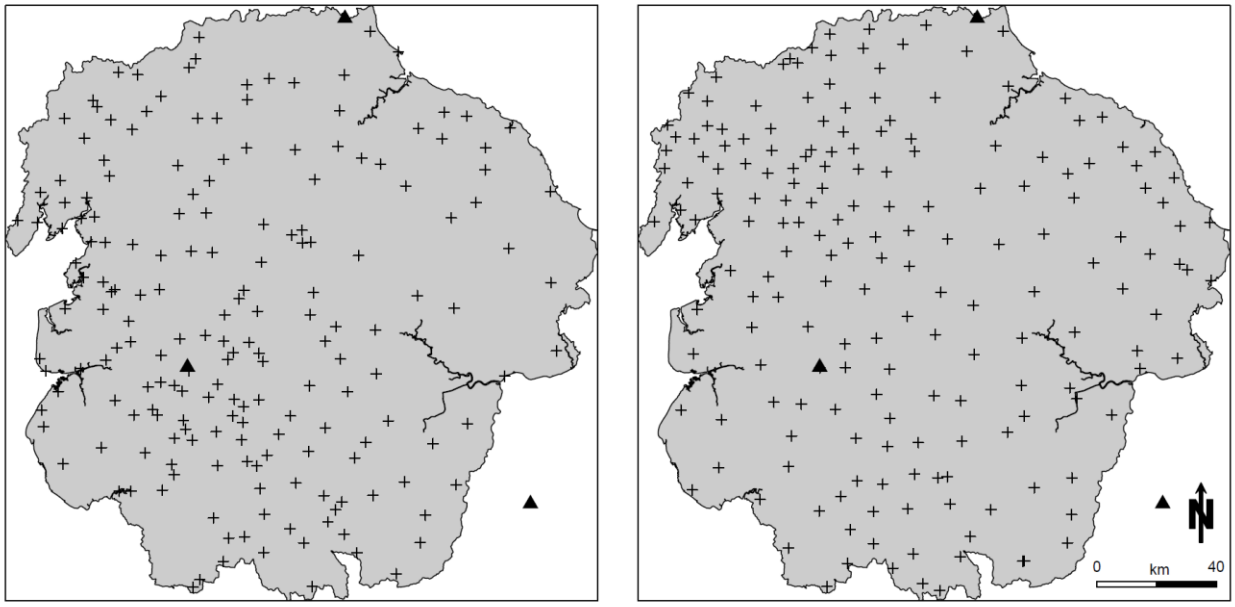


Figure 3.3: Pattern for the initial (left) and optimised (right) rain-gauge network.

3.5 Conclusions

This work presented sampling design optimisation with a simple parametric and non-stationary variance model. The variance is modelled as a function of covariates and the sampling design is optimised on this basis. Optimising the network reduced the space-time averaged kriging variance, which is a summary measure of the interpolation accuracy, by about 5%. The rain-gauge design proposed is optimal for the prediction of rainfall at daily scale.

The method offers the advantage of being relatively easy and accounts for the complex spatial variance of this case study. Non-stationarity in the covariance structure is often neglected, even if a simple exploratory analysis of the data could give a diagnostic on the choice of the geostatistical model.

The decrease of error prediction variance is relatively low compared to other similar studies. This is due to the difficulty to find a better positioning of the sample locations that decreases on average the criterion over space but also time, while previous work deal with optimisation in space only. The optimal spatial sampling design varies between days and for a full year a compromise must be made.

The SSA algorithm is slow, which is not new, but it becomes prohibitively computationally intensive when the calculated criterion is averaged over a long period, on a high resolution prediction grid. One solution proposed in this work is to use parallel computing for calculating the criterion and on a coarser grid.

Sampling design optimisation for rainfall mapping can lead to reduction of sampling costs or more accurate maps without increasing sampling costs.

4 References

BAUME, O. P., GEBHARDT, A., GEBHARDT, C., HEUVELINK, G. B. M. & PILZ, J. 2011. Network optimization algorithms and scenarios in the context of automatic mapping. *Computers & Geosciences,* 37**,** 289-294.

BEHZADIAN, K. A. A., ABDOLLAH%A KAPELAN, ZORAN%A SAVIC, DRAGAN 2008. Stochastic sampling design for water distribution model calibration. *International Journal of Civil Engineering,* 6**,** 48-57.

BROAD, D. R., DANDY, G. C. & MAIER, H. R. 2005. Water Distribution System Optimization Using Metamodels. *Journal of Water Resources Planning and Management,* 131**,** 172-180.

BRUNGARD, C. W. & BOETTINGER, J. L. 2010. Conditioned Latin Hypercube Sampling: Optimal Sample Size for Digital Soil Mapping of Arid Rangelands in Utah, USA. *In:* BOETTINGER, J. L., HOWELL, D. W., MOORE, A. C., HARTEMINK, A. E. & KIENAST-BROWN, S. (eds.) *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation.* Dordrecht: Springer Netherlands.

BRUS, D. J. & 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**,** 1-44.

BRUS, D. J., DE GRUIJTER, J. J., VAN GROENIGEN, J. W., P. LAGACHERIE, A. B. M. & VOLTZ, M. 2006. Chapter 14 Designing Spatial Coverage Samples Using the k-means Clustering Algorithm. *Developments in Soil Science.* Elsevier.

BRUS, D. J. & HEUVELINK, G. B. M. 2007. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma,* 138**,** 86-95.

CARRÉ, F., MCBRATNEY, A. B. & MINASNY, B. 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. *Geoderma,* 141**,** 1-14.

CASELTON, W. F. & ZIDEK, J. V. 1984. Optimal monitoring network designs. *Statistics & Probability Letters,* 2**,** 223-227.

DE GRUIJTER, J., BRUS, D. J., BIERKENS, M. F. P. & KNOTTERS, M. 2006. *Sampling for Natural Resource Monitoring*, Springer.

DIGGLE, P. & RIBEIRO, P. 2007. *Model-based Geostatistics*, Springer.

DIXON, W., SMYTH, G. K. & CHISWELL, B. 1999. Optimized selection of river sampling sites. *Water Research,* 33**,** 971-978.

DOBERMANN, A., SIMBAHAN, G. C., P. LAGACHERIE, A. B. M. & VOLTZ, M. 2006. Chapter 13 Methodology for Using Secondary Information in Sampling Optimisation for Making Fine-resolution Maps of Soil Organic Carbon. *Developments in Soil Science.* Elsevier.

EHRGOTT, M. 2005. *Multicriteria Optimization*, Springer.

GE, Y., WANG, J. H., HEUVELINK, G. B. M., JIN, R., LI, X. & WANG, J. F. 2015. Sampling design optimization of a wireless sensor network for monitoring ecohydrological processes in the Babao River basin, China. *International Journal of Geographical Information Science,* 29**,** 92-110.

GEBHARDT, C. 2003. Bayesian Methods for Geostatistical Design. Ph.D. Thesis. Austria: University of Klagenfurt.

HEUVELINK, G. B. M., BRUS, D. J., DE GRUIJTER, J. J., P. LAGACHERIE, A. B. M. & VOLTZ, M. 2006. Chapter 11 Optimization of Sample Configurations for Digital Mapping of Soil Properties with Universal Kriging. *Developments in Soil Science.* Elsevier.

HEUVELINK, G. B. M., GRIFFITH, D. A., HENGL, T. & MELLES, S. J. 2013. *Sampling design optimization for space-time kriging. In: Spatio-temporal design: Advances in efficient data acquisition.*, Wiley.

HEUVELINK, G. B. M., JIANG, Z., BRUIN, S. D. & TWENHOFEL, C. J. W. 2010. Optimization of mobile radioactivity monitoring networks. *Int. J. Geogr. Inf. Sci.,* 24**,** 365-382.

KAPELAN, Z. S., SAVIC, D. A. & WALTERS, G. A. C. W. 2005. Multiobjective design of water distribution systems under uncertainty. *Water Resources Research,* 41**,** n/a-n/a.

KITSIOU, D., TSIRTSIS, G. & KARYDIS, M. 2001. Developing an Optimal Sampling Design. A Case Study in a Coastal Marine Ecosystem. *Environmental Monitoring and Assessment,* 71**,** 1-12.

KO, C.-W., LEE, J. & QUEYRANNE, M. 1995. An Exact Algorithm for Maximum Entropy Sampling. *Operations Research,* 43**,** 684-691.

KOLLAT, J. B. & REED, P. M. 2006. Comparing state-of-the-art evolutionary multi-objective algorithms for long-term groundwater monitoring design. *Advances in Water Resources,* 29**,** 792-807.

KRAUSE, A., SINGH, A. & GUESTRIN, C. 2008. Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies. *J. Mach. Learn. Res.,* 9**,** 235-284.

KROL, B. G. C. M. 2008. Towards a Data Quality Management Framework for Digital Soil Mapping with Limited Data. *In:* HARTEMINK, A. E., MCBRATNEY, A. & MENDONÇA-SANTOS, M. D. L. (eds.) *Digital Soil Mapping with Limited Data.* Dordrecht: Springer Netherlands.

LAGACHERIE, P. 2008. Digital Soil Mapping: A State of the Art. *In:* HARTEMINK, A. E., MCBRATNEY, A. & MENDONÇA-SANTOS, M. D. L. (eds.) *Digital Soil Mapping with Limited Data.* Dordrecht: Springer Netherlands.

LE, N. & ZIDEK, J. 2006. *Statistical Analysis of Environmental Space-Time Processes,* New York, Springer.

LEE, J. & WILLIAMS, J. 2003. A linear integer programming bound for maximum-entropy sampling. *Mathematical Programming,* 94**,** 247-256.

LOPHAVEN, S. 2002. Design and analysis of environmental monitoring programs. PhD Thesis. Technical University of Denmark.

MELLES, S. J., HEUVELINK, G. B. M., TWENHÖFEL, C. J. W., VAN DIJK, A., HIEMSTRA, P. H., BAUME, O. & STÖHLKER, U. 2011. Optimizing the spatial pattern of networks for monitoring radioactive releases. *Computers & Geosciences,* 37**,** 280-288.

MINASNY, B. & MCBRATNEY, A. B. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences,* 32**,** 1378-1388.

RAMIREZ-LOPEZ, L., SCHMIDT, K., BEHRENS, T., VAN WESEMAEL, B., DEMATTÊ, J. A. M. & SCHOLTEN, T. 2014. Sampling optimal calibration sets in soil infrared spectroscopy. *Geoderma,* 226–227**,** 140-150.

ROYLE, J. A. & NYCHKA, D. 1998. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. *Computers & Geosciences,* 24**,** 479-488.

SCHMIDT, K., BEHRENS, T., DAUMANN, J., RAMIREZ-LOPEZ, L., WERBAN, U., DIETRICH, P. & SCHOLTEN, T. 2014. A comparison of calibration sampling schemes at the field scale. *Geoderma,* 232–234**,** 243-256.

SHARP, W. E. 1971. A Topologically Optimum Water-Sampling Plan for Rivers and Streams. *Water Resources Research,* 7**,** 1641-1646.

SHEWRY, M. C. & WYNN, H. P. 1987. Maximum entropy sampling. *Journal of Applied Statistics,* 14**,** 165-170.

STUMPF, F., SCHMIDT, K., BEHRENS, T., SCHONBRODT-STITT, S., BUZZO, G., DUMPERTH, C., WADOUX, A., XIANG, W. & Scholten, T. (2016). Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. *Journal of Plant Nutrition and Soil Science*, *179*, 499-509.

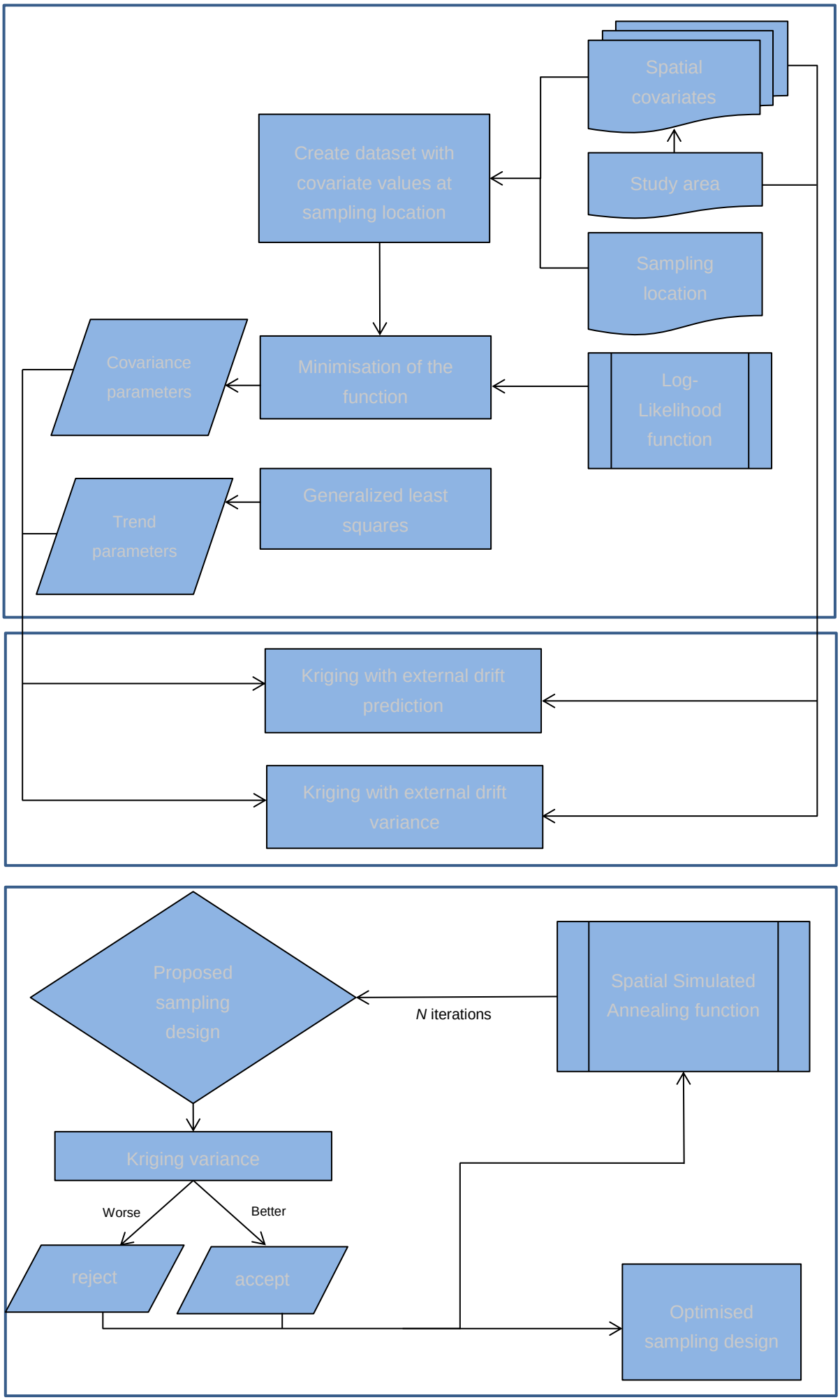
SULAEMAN, Y., MINASNY, B., MCBRATNEY, A. B., SARWANI, M. & SUTANDI, A. 2013. Harmonizing legacy soil data for digital soil mapping in Indonesia. *Geoderma,* 192**,** 77-85.

VAŠÁT, R., HEUVELINK, G. B. M. & BORŮVKA, L. 2010. Sampling design optimization for multivariate soil mapping. *Geoderma,* 155**,** 147-153.

WADOUX, A.M.J-C., BRUS, D.J., RICO-RAMIREZ, M.A., HEUVELINK, G.B.M., 2017. Sampling design optimisation for rainfall prediction using a non-stationary geostatistical model. Adv. Water Resour. 107, 126–138. doi:10.1016/j.advwatres.2017.06.005

WANG, J., GE, Y., HEUVELINK, G. B. M. & ZHOU, C. 2014. Spatial Sampling Design for Estimating Regional GPP With Spatial Heterogeneities. *IEEE GEOSCIENCE AND REMOTE SENSING LETTERS,* 11.

Appendix A - Flowchart sampling design software implementation



Appendix B – R script of sampling design optimisation of a non-stationary variance model

Codes can be downloaded in GitHub through:
https://github.com/AlexandreWadoux/non_stationary_variance_kriging

With reference to the original manuscript:

Wadoux, A.M..-C., Brus, D.J., Rico-Ramirez, M.A., Heuvelink, G.B.M., 2017. Sampling design optimisation for rainfall prediction using a non-stationary geostatistical model. Adv. Water Resour. 107, 126–138. doi:10.1016/j.advwatres.2017.06.005

Simple example showing REML estimation of non-stationary variance parameters, and kriging with an external drift with non-stationary variance paarmeters

Load the necessary packages

```
library(variography) # upon request, not available on CRAN
library(DEoptim)
library(gstat)
library(sp)
library(fields)
```

Simulate field Define discretization grid

```
x1<-seq(1:50)-0.5
x2<-x1
grid<-expand.grid(x1,x2)
names(grid)<-c("x1","x2")
```

Compute spatial trend; x1 is used as covariate z

```
grid$z <- grid$x1
b1 <- 2
grid$mu<-b1*grid$z
```

Define variogram for simulation of residuals

```
sill<-1
range<-10
vgm<-sill*Exp(range)
```

Compute matrix with distances between simulation nodes

```
distx <- outer(grid$x1,grid$x1,FUN="-")
disty <- outer(grid$x2,grid$x2,FUN="-")
dist <- sqrt(distx^2+disty^2)
```

Compute matrix with mean covariances

```
cvm <- as(vgm,"CovariogramStructure") #coerce variogram to covariance function
f <- as(cvm, "function")
C <- f(h=dist)
```

Simulate values for residuals by Cholesky decomposition

```
set.seed(31415)
Upper <- chol(C)
G <- rnorm(n=nrow(grid),0,1) #simulate random numbers from standard normal dist
ribution
grid$residuals <- crossprod(Upper,G)
```

Multiply residuals in upper half by a constant so that residual variance becomes non-stationary

```
ids <- which(grid$x2>25)
grid$residuals[ids] <- grid$residuals[ids]*3
```

Add residuals to trend

```
grid$y <- grid$mu+grid$residuals
```

Select simple random sample from grid

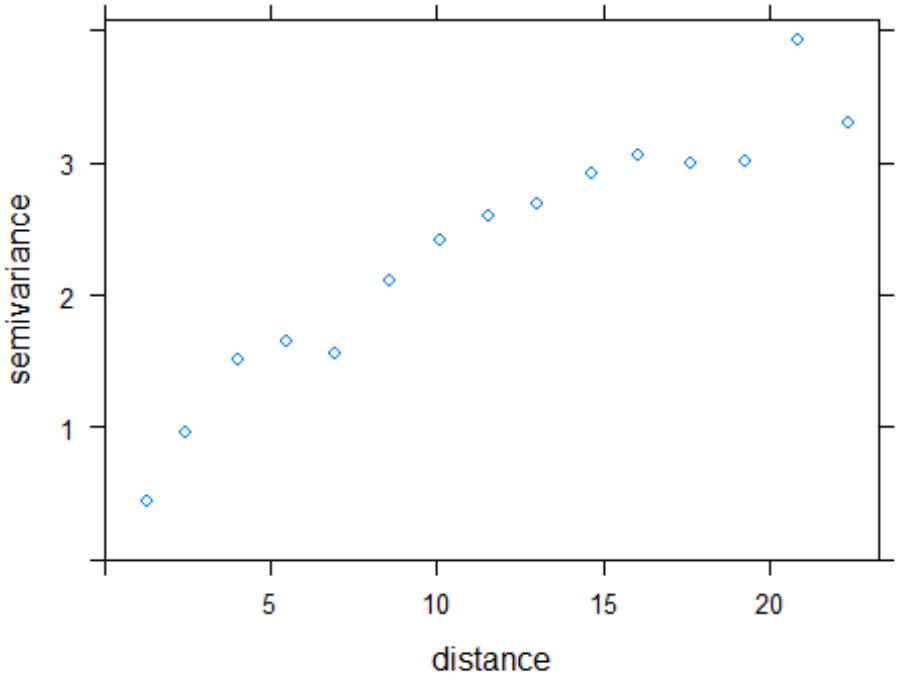
```
idssam <- sample.int(nrow(grid),size=100,replace=F)
dat <- grid[idssam,]
```

Compute matrix with distances between sampling points

```
D <- as.matrix(dist(cbind(dat$x1,dat$x2)))
```

Define the variogram

```
olsmodel <- lm(y~z,data=dat)
X <- model.matrix(olsmodel)
y <- dat$y
coordinates(dat) <- ~x1+x2
vg <- variogram(y~z,data=dat)
plot(vg)
```



Open empty matrices

```
dat <- as.data.frame(dat)
S <- diag(nrow=nrow(dat))
```

Load the negative log-likelihood

```
neglogLikelihood <-
  function(theta) {
    c1 <- theta[1]
    a1 <- theta[2]
    sigma1 <- theta[3]
    sigma2 <- theta[4]
    R<- c1*exp(-D/a1)
    diag(R) <- 1
    diag(S) <- (dat$x2<25)*sigma1+(dat$x2>25)*sigma2
    V <- t(S) %*% R %*% S
    V_inv <- chol2inv(chol(V))
    XV <- crossprod(X, V_inv)
    XVX <- XV %*% X
    XVX_inv <- chol2inv(chol(XVX))
    I <- diag(nrow(D))
    logDetV <- determinant(x = V, logarithm = TRUE)$modulus
    logDetXVX <- determinant(x = XVX, logarithm = TRUE)$modulus
    P= I - X %*% chol2inv(chol(crossprod(X,X)))%*%t(X)
    y.t <-t(P)%*%y
    Q <- X %*% XVX_inv %*% XV
    logLikelihood <- -0.5 * (logDetV + logDetXVX + crossprod(y.t, V_inv) %*%  (
I - Q) %*% y)
    neglogLikelihood <- -1 * logLikelihood
```

```
    return(neglogLikelihood)
  }
```

Parameter range in the parameter search c1 is partial sill parameter of correlogram, so must be < 1.
Optimisation of the parameters

```
lbound <- c(c1 = 0, a1 = 0.1, sigma1=0.1, sigma2=0.1)
ubound <- c(c1 = 1, a1 = 20, sigma1=10, sigma2=10)
optPars <- DEoptim(
  fn = neglogLikelihood,
  lower = lbound,
  upper = ubound,
  control = DEoptim.control(strategy =2, bs=F, NP=40, itermax=200, CR=0.5, F=0.
8, trace=FALSE)
)

result<-optPars$optim$bestmem
c1 <- result[1]
a1 <- result[2]
sigma1 <- result[3]
sigma2 <- result[4]
```

Now estimate regression coefficients by GLS using REML estimates of variogram parameters

```
R<- c1*exp(-D/a1)
diag(R) <- 1
diag(S) <- (dat$x2<25)*sigma1+(dat$x2>25)*sigma2
V <- t(S) %*% R %*% S
V_inv <- chol2inv(chol(V))
XV <- crossprod(X, V_inv)
XVX <- XV %*% X
XVX_inv <- chol2inv(chol(XVX))
Vy <- crossprod(y, V_inv)
XVy<-crossprod(X,t(Vy))
(beta_GLS<-XVX_inv%*%XVy)

##        [,1]
## [1,] -0.107
## [2,]  2.013
```

Compute matrix with correlation between sampling points and prediction node

```
D0 <- rdist(cbind(grid$x1,grid$x2),cbind(dat$x1,dat$x2))
R0<- c1*exp(-D0/a1)
```

Compute residuals at sampling points

```
resid <- dat$y-X%*%beta_GLS
```

Compute sigma at the sampling points

```
sigma0 <- (grid$x2<25)*sigma1+(grid$x2>25)*sigma2
```

Compute sigma at the prediction nodes

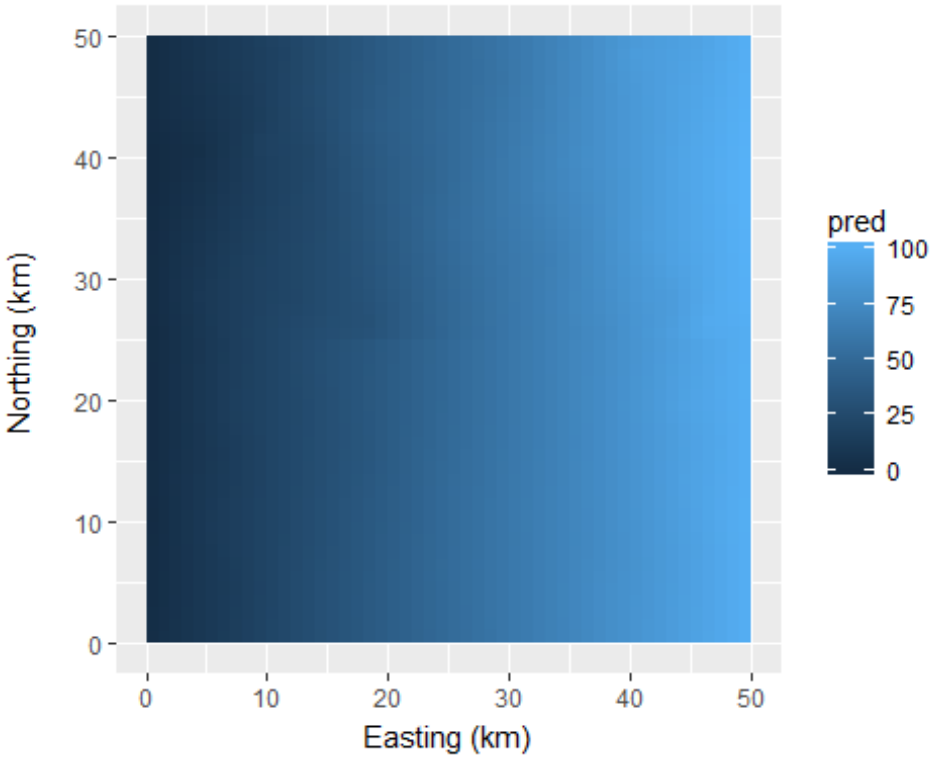
```
sigmadata <- (dat$x2<25)*sigma1+(dat$x2>25)*sigma2
```

Make the kriging predictions

```
pred <- varpred <- numeric(length=nrow(grid))

for (i in 1:nrow(grid)) {
  x0 <-c(1,grid$z[i])
  mu0 <- x0%*%beta_GLS
  r0 <- R0[i,]
  c0 <- r0*sigmadata*sigma0[i]

  Cc0 <- solve(V,c0)
  pred[i]<- mu0 + crossprod(Cc0, as.vector(resid))

  x_a <- x0 - crossprod(X,Cc0)
  varpred[i] <- c1*sigma0[i]^2 - crossprod(c0, Cc0) + crossprod(x_a,solve(XVX,
x_a))
}

grid$pred<-pred
grid$krigvar <- varpred

hist(varpred)
```

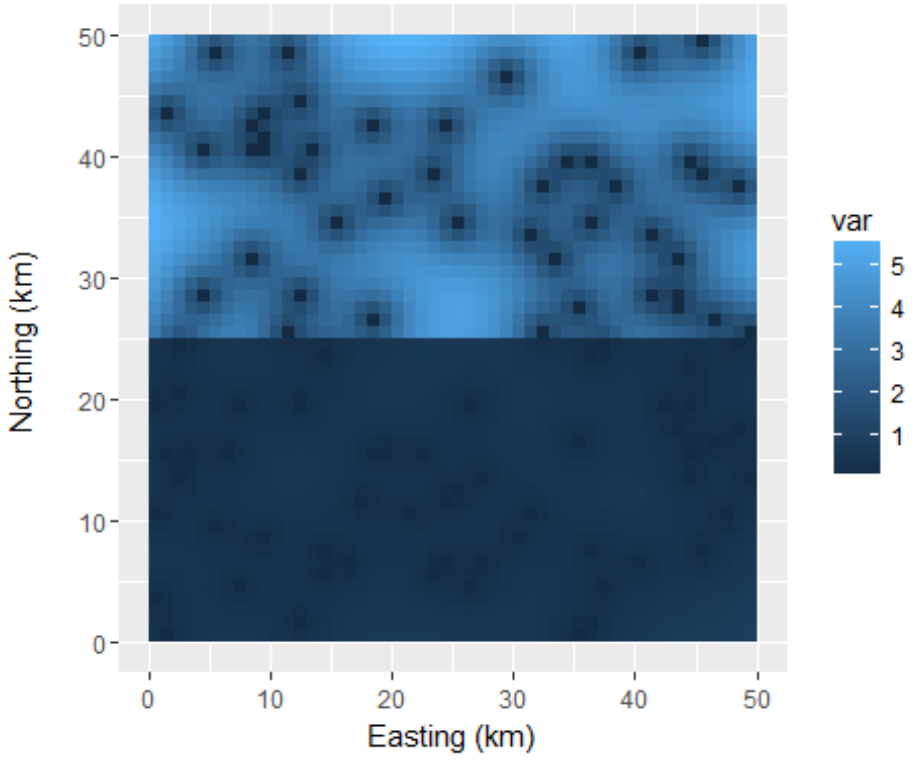
Plot the prediction

```
library(ggplot2)
ggplot(data = grid) +
  geom_tile(mapping = aes(x = x1, y = x2, fill = pred)) +
  scale_x_continuous(name = "Easting (km)") +
  scale_y_continuous(name = "Northing (km) \n") +
  scale_fill_continuous(name = "pred")+
  coord_equal(ratio = 1)
```



Plot the prediction error variance

```
ggplot(data = grid) +
  geom_tile(mapping = aes(x = x1,y = x2,fill = krigvar)) +
  scale_x_continuous(name = "Easting (km)") +
  scale_y_continuous(name = "Northing (km) \n") +
  scale_fill_continuous(name = "var")+
  coord_equal(ratio = 1)
```



Save the parameters for the optimization

```
save(grid, dat, c1, a1, sigma1, sigma2, file='temp_res.RData' )
```

Simple example showing spatial optimisation of the observations for kriging with a non-stationary variance

Load the necessary packages

```
library(spsann)
library(sp)
library(raster)
library(ggplot2)
```

Load the parameters saved from the main.R document

```
load(temp_res.RData')
```

Define discretisation grid

```
x1<-seq(1:50)-0.5
x2<-x1
candi<-expand.grid(x1,x2)
names(candi)<-c("x","y")
```

Define a polygon of the study boundary: first create a raster from candi

```
r <- candi
coordinates(r) <- ~ x + y
gridded(r) <- TRUE
```

take the extent of the raster

```
e <- extent(r)
```

coerce to a SpatialPolygons object

```
p <- as(e, 'SpatialPolygons')
```

Create the objective function for minimize

```
FUN <- function(points, a1, c1, sigma1, sigma2, grid,...){
  dat <- points

  S <- diag(nrow=nrow(dat))


  #Compute design matrix of the propozed design
  X <- matrix(data=1, nrow=nrow(dat), ncol=2)
  rownames(X) <- dat[,1]
  t <- cbind(rownames(grid), grid$z)
  X[,2] <- as.numeric(t[,2][as.numeric(rownames(X))])

  D <- as.matrix(dist(cbind(dat[,2],dat[,3])))
  R<- c1*exp(-D/a1)
  diag(R) <- 1
  diag(S) <- (dat[,3]<25)*sigma1+(dat[,3]>25)*sigma2
  V <- t(S) %*% R %*% S
  V_inv <- chol2inv(chol(V))
```

```
  XV <- crossprod(X, V_inv)
  XVX <- XV %*% X

  #Compute matrix with correlation between sampling points and prediction node
  library(fields)
  D0 <- rdist(cbind(grid[,1],grid[,2]),cbind(dat[,2],dat[,3]))
  R0<- c1*exp(-D0/a1)

  #compute sigma at the sampling points
  sigmadata <- (dat[,3]<25)*sigma1+(dat[,3]>25)*sigma2

  #compute sigma at the prediction nodes
  sigma0 <- (grid$x2<25)*sigma1+(grid$x2>25)*sigma2

  varpred <- numeric(length=nrow(grid))
  for (i in 1:nrow(grid)) {
    x0 <-c(1,grid$z[i])

    r0 <- R0[i,]
    c0 <- r0*sigmadata*sigma0[i]
    Cc0 <- solve(V,c0)
    x_a <- x0 - crossprod(X,Cc0)
    varpred[i] <- c1*sigma0[i]^2 - crossprod(c0, Cc0) + crossprod(x_a,solve(XVX
, x_a))
  }

  return(mean(varpred))
}
```

Set the control parameters of the simulated annealing function

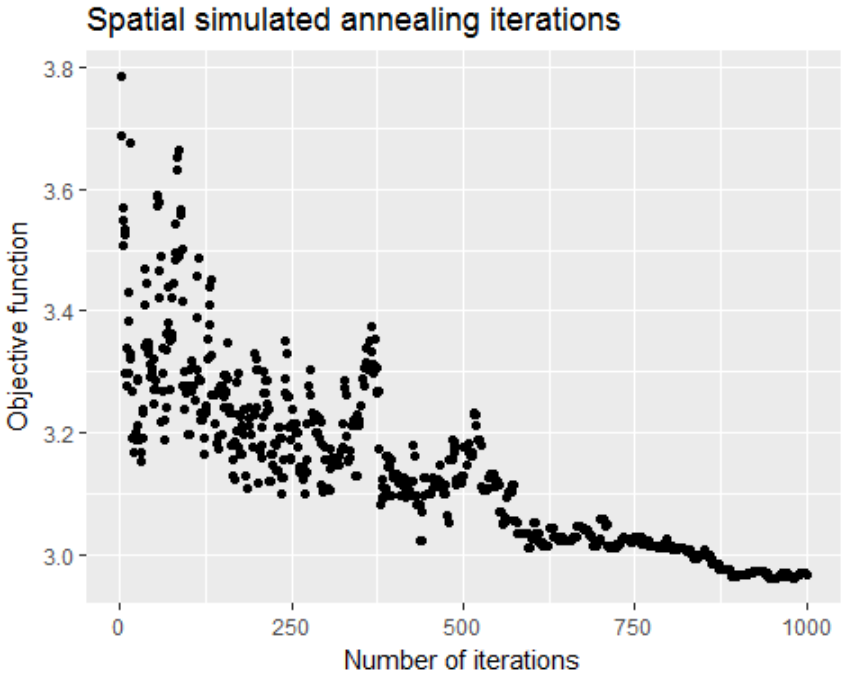
```
schedule <- scheduleSPSANN(chains = 100, initial.temperature = 0.3,
                           initial.acceptance = 0.7, x.max = max(grid$x1),
                           y.max = max(grid$x2), x.min = 0, y.min = 0, cellsize
= 1)
```

Optimisation of the sampling scheme (change to plotit=TRUE if you wish to see the process) - takes several minutes. Optimization for 10 observations only to speed up computation.

```
optimized <- optimUSER(points = 10, candi, fun = FUN, schedule = schedule,
                       plotit = F, track = T, progress = "txt",boundary = p,
                       verbose = T, a1 = a1, c1, sigma1, sigma2, grid)
```

Plot the decrease of the objective function

```
qplot(seq(1, length(optimized$objective$energy$obj), by = 1), optimized$objecti
ve$energy$obj,
      xlab='Number of iterations', ylab = 'Objective function', main = 'Spatial
simulated annealing iterations' )
```



Plot the optimized design

```
qplot(optimized$points$x, optimized$points$y,
      xlab='x', ylab='y', main='Optimized design for 10 observations' )
```