Sampling design optimization for geostatistical modelling and prediction

Alexandre M. J.-C. Wadoux

Thesis committee

Promotor
Prof. Dr Gerard B.M. Heuvelink
Special Professor Pedometrics and Digital Soil Mapping
Wageningen University & Research

Co-promotor
Dr Dick J. Brus
Researcher, Biometris, Wageningen

Other members
Prof. Dr Arnold K. Bregt, Wageningen University & Research
Dr Philippe Lagacherie, INRA Montpellier
Prof. Dr Marc F.P. Bierkens, Utrecht University
Prof. Dr Alfred Stein, University of Twente

This research was conducted under the auspices of the C.T. de Wit Graduate School for Production Ecology and Resource Conservation.

Sampling design optimization for geostatistical modelling and prediction

Alexandre M.J.-C. Wadoux

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 30 August 2019
at 4:00 p.m. in the Aula.

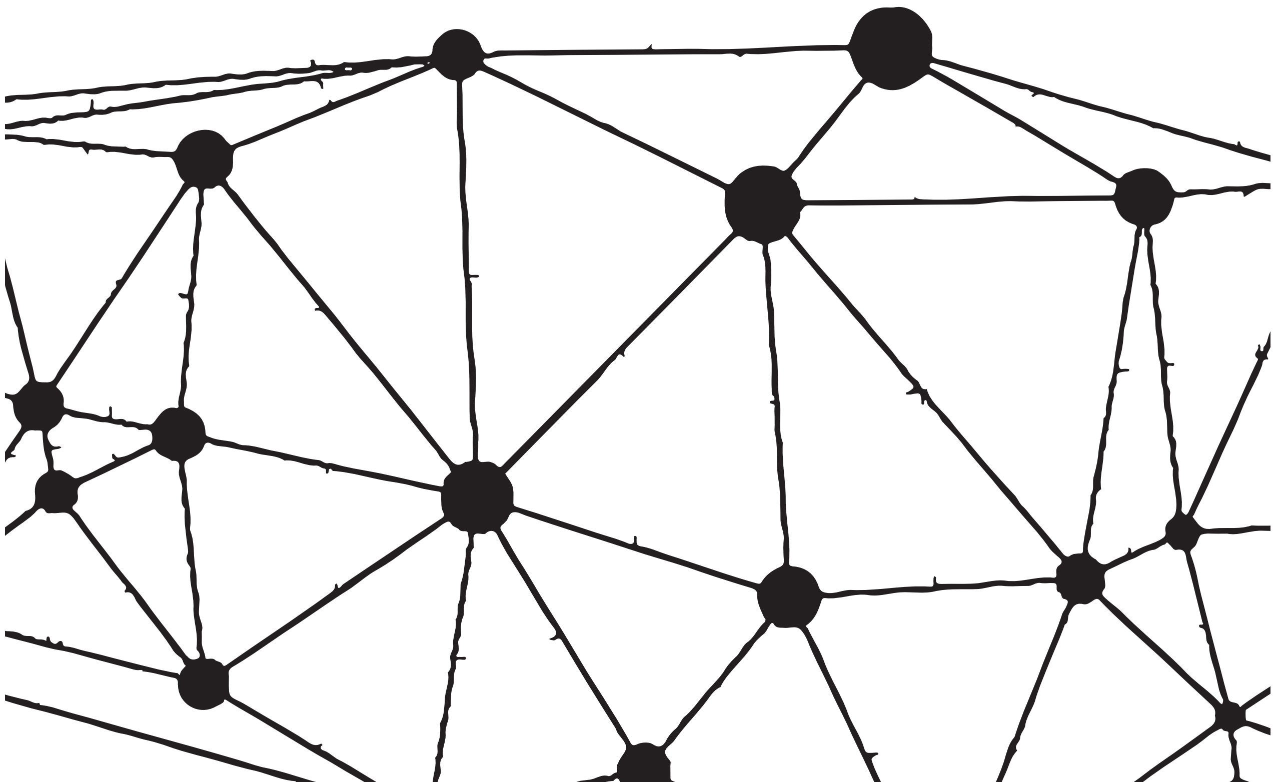
Alexandre M.J.-C. Wadoux
Sampling design optimization for geostatistical modelling and prediction
192 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2019)
With references, with summaries in English and French

ISBN 978-94-6343-930-5
DOI https://doi.org/10.18174/474486

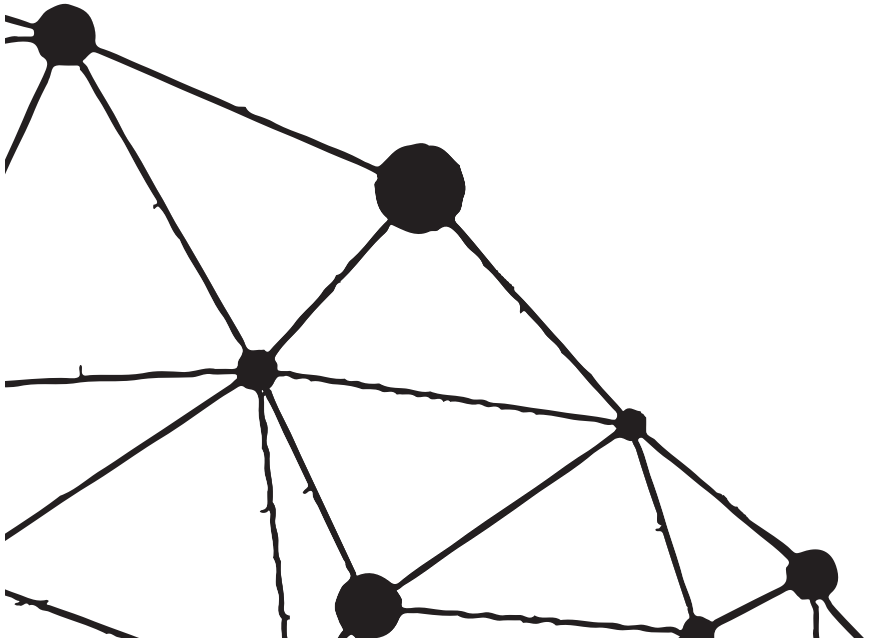
D'or, à une Notre-Dame de carnation vêtue d'azur et de gueules, tenant dans ses bras l'enfant Jésus de carnation

Index



Chapter 1

General introduction



1.1 Background

We constantly require information on our surrounding environment. We wonder how much rain we had during the day, we want to know whether our soil is fertile so as to plant a vegetable garden. Most of our questions can be answered via simple observing. One may look at the sky and identify dark clouds or analyse whether the current garden vegetation flourishes. In many cases, this will give sufficient information to answer our questions.

In other circumstances, such as for space-time monitoring and prediction of environmental variables, quantitative environmental information is needed. Whether we are regulators who want to assess over time the increase of the river discharge after a rainfall event or policy makers who want to know the daily rate of nitrogen dioxide emission in the city's air and its consequences in terms of public health, real-world measurements are needed to obtain the required information. Scientists can provide such information by inspecting and measuring the environment. But environmental variables cannot be measured everywhere. We cannot measure the rainfall at every single point in space and time and we cannot continuously measure the nitrogen dioxide concentration everywhere in a city. Instead, scientists can collect a fragment, one or several units of the environment, with the purpose of using these units as reference values for the whole area to be studied.

Difficulties arise when collecting a single unit from a spatially varying variable (Webster and Lark, 2012). First, one may collect the unit in a location where the variable of interest exhibits abnormally large or small values. For example, monitoring the carbon dioxide concentration of the air of a city may give extremely high pollution concentration when measured near a ring road. This may persuade policy makers to limit severely the traffic within the city, which may in turn restrict unreasonably the local economy. In contrast, a monitoring station placed on the top of a building may provide too small values of the city air pollution, and not exhort actions against dangerous carbon dioxide concentration for the population. Second, even if the collected unit does not show an atypical value, one may want to map the environmental variable. The air is likely to be more polluted close to the city roads and less within the city largest parks. There might be a few local conditions boosting the carbon dioxide concentration. Collecting one single unit is not a realistic option in practice as it enables neither mapping nor estimating statistics related to the spatial variation of the variable of interest. We need replication of the sampling at multiple locations for mapping a target variable (Lark, 2003), i.e. collecting a sample of multiple units. In this thesis I am interested in mapping environmental variables and so it is evident that the sample must contain multiple units.

A sample is a set of units collected from the whole population. In this thesis a sampling unit is a point-location in the two-dimensional geographical space. Replication involves measuring the target variable at multiple locations in space. Collecting a sampling unit might be very costly and time consuming. For instance, this is the case in the mining industry, where an individual mechanized borehole costs several thousands of euros. In fact, the ground might need to be drilled for several days to retrieve a single unit at great depth. One can avoid unnecessary spending by collecting the smallest number of units needed for a precision requirement or the largest number that can be afforded for a given sampling budget. These units must then be optimally selected in space so as to optimize a criterion related to the target variable and its intended use.

1.2 Problem definition

1.2.1 What is optimal?

A design is optimal with respect to a criterion. In this thesis, I define a criterion as a mathematical quantification of the quality of a sampling configuration. Its evaluation gives information on whether one sampling configuration is better than another. For example, a farmer who wishes to assess the topsoil carbon content of his field may wish to sample evenly in space. In this case, a bad design is a design in which the units of the sample are taken at close locations one to another, while a good design is found when sampling uniformly over the field. In this case, a criterion that evaluates the quality of a design should quantify the overall dispersion of the units in space. As the criterion is quantitative, it can be either maximized or minimized using optimization algorithms (Ehrgott, 2005). An optimized value of the criterion serves for finding the optimal spatial sampling design. In fact, a single criterion may not cover all different expected qualities of a sampling design, due to contrasting and sometimes conflicting definitions of what makes a suitable design (Sawicka et al., 2017). In some cases, the formulation of the criterion itself contributes to the understanding of the problem (Van Groenigen et al., 1999). Another difficulty arises when the formulation of the design quality is qualitative and subjective, making the translation of the objectives into a mathematical, quantitative criterion difficult (Lophaven, 2004).

From the above, I define a sample of size n with sampling configuration $\xi = \{s_1, s_2, \ldots, s_n\}$ where s_i are geographic locations in the area of interest $\mathcal{A}$. The optimal sampling configuration (i.e. optimal design) is found through evaluation of the criterion for all possible sampling configurations. The optimal sampling configuration ξ^* is the one achieving the smallest value of the criterion. Thus, the choice of the criterion reflects the purpose of the optimization and determines the resulting optimized design (Mateu and Müller, 2012). I acknowledge that this assumes that there is one unique optimized sampling configuration, because of the model-based setting. This is explained later in this chapter.

1.2.2 Constraints on the design

Candidate locations for sampling can be selected anywhere in $\mathcal{A}$. In practice, however, one may want to constrain this choice by avoiding to select sampling units at unreachable locations or that are very expensive to collect. Typical constraints are budgetary ones, where the sample size cannot exceed a certain limit. Cost can be accounted for by penalizing the sampling locations that are expensive to reach (Roudier et al., 2012), or by accounting for fieldwork, field equipments and laboratory costs (Brus et al., 1999). Operational constraints may lead decision making, by placing emphasis on accessibility or easy maintenance of the monitoring stations (Changnon et al., 1980). Constraints can also be related to subjective decisions in order to favour areas of specific ecological interest (Asadollahfardi, 2015). The above constraints commonly set limits on the spatial design for which the mathematical criterion is evaluated.

1.2.3 Criteria for mapping cost and accuracy

The quality of a design is evaluated based on two main types of criterion assessing either mapping cost or mapping accuracy. Designs optimal in term of mapping costs are often found by minimizing a function of the sample size or sampling unit accessibility as a surrogate of the total sampling costs, with respect to a map accuracy measure (Yang et al., 2018; Brus et al., 2019). Examples of such designs can for example be found in ecology (Lugg et al., 2018), soil science (Brungard and Boettinger, 2010) and hydrology (Nunes et al., 2004). Mapping costs have been minimized jointly with mapping accuracy in a multi-criteria optimization (Pardo-Igúzquiza, 1998). The authors used a trade-off parameter weighing the separate contribution due to costs of installing new meteorological stations and to variance reduction achieved by adding more stations. During the last decade, much focus has been put on criteria to comply with high mapping accuracy. The most used is the spatially averaged prediction error variance, first used in the 80's (Delhomme, 1978; Bastin et al., 1984) and common nowadays (Brus and Heuvelink, 2007; Barca et al., 2015; Ge et al., 2015). This criterion is eventually combined with the error associated to computing a trend (Hengl et al., 2003), or a covariance structure (Russo, 1984; Müller and Zimmerman, 1999) in a weighted multi-criteria optimization (Müller and Stehlík, 2010). Other possible criterion are the false discovery rate for remediation of contaminated areas (Marchant et al., 2013), the calamity detection capability of a network (Melles et al., 2011), the false classification into safe and unsafe zones due to radioactive plume (Heuvelink et al., 2010), or the mean squared shortest distance between sampling locations. The latter leads to a geometric design (Brus et al., 2007).

1.2.4 Sampling strategies

Some of the mentioned criteria need a model of spatial variation to be computed. This relies on the model-based statistical inference strategy decided upon. De Gruijter et al. (2006) distinguish two sampling approaches, namely the model- and the design-based approaches. Design-based sampling is based on classical sampling theory, wherein the sampling units are selected randomly in such a way that every element of the population has a given probability of being selected (Cochran, 1977). In this thesis I use solely non-probability sampling designs, as they are generally more suitable for mapping (De Gruijter et al., 2006). Non-probability sampling designs comprise among others regular grid sampling, spatial coverage sampling, feature space coverage sampling using k-means, conditioned Latin Hypercube sampling, response surface sampling, Kennard-Stone sampling and model-based sampling (Brus, 2019). This thesis focuses mainly on model-based sampling, in combination with a model-based inference strategy for mapping, even if the former does not necessarily imply the latter (De Gruijter et al., 2006). In reality, environmental variables are the outcome of a deterministic rather than a random process (Webster, 2000). In practice, however, we proceed by assuming the variable to be the outcome of a random process, as assumed in geostatistical inference. The spatial variation of the variable is described by a stochastic model which opens the possibility for optimized, model-based (purposive) sampling so as to make inference on the assumed geostatistical model. This is the model-based geostatistical mapping mentioned in the previous paragraphs of this Introduction.

1.2.5 Geostatistical mapping

Model-based inference using geostatistics is a sub-branch of spatial statistics introduced in the early 1960s by Matheron (1963) to study the spatial distribution of regionalized variables. I first define the sampling location of the units $s_1, s_2, \ldots, s_n$ and a random field Z modeled by $Z(\mathbf{s}) = \mu + \varepsilon(\mathbf{s})$ where μ is the mean and ε is a zero mean random process with covariance $\mathrm{cov}(\varepsilon(s_i), \varepsilon(s_j)) = C(|s_i - s_j|)$. This statistical representation of the reality is based on a set of assumptions, the most important ones being first-order stationarity, i.e. the unknown mean μ is constant over the area $\mathcal{A}$, and second-order stationarity, i.e. the covariance of the random process at two locations is independent of their spatial locations and depends only on the geographic separation distance and direction between them.

The next step relies on the definition of the structure of the spatially correlated residual ε with correlogram $\rho(\mathbf{h};\theta)$, where $\mathbf{h}$ is the separation distance between units. Since μ is constant, $\rho(\mathbf{h})$ is specified by the covariance function C of Z so that the correlation function $\rho(\mathbf{h}) = C(\mathbf{h})/C(0)$. The parameter vector θ of the model $\rho(\mathbf{h};\theta)$ can be either assumed to be known, taken from a correlogram whose parameters are already estimated from the same variable in similar conditions, or fitted using sample data by parameter estimation methods such as methods-of-moments, maximum likelihood (ML) or restricted maximum likelihood (REML).

In the third and final step, values at non-measured locations can be predicted with their associated prediction error variance. The core geostatistical prediction method is known as *kriging* from the work of the mining engineer Danie Krige (Krige, 1951). Kriging is a model-based method used in many different fields, such as in soil science, meteorology, epidemiology and mineral resources evaluation. A large number of kriging variants have been adopted to face the different natural processes. Kriging with external drift (KED) relaxes the first-order stationarity assumption by replacing μ by a trend model, so that $\mu(\mathbf{s}) = \sum_{k=0}^{K} \beta_k g_k(\mathbf{s})$ (Goovaerts, 1997). Indicator kriging models binary responses $I(\mathbf{s})$ (Solow, 1986) while disjunctive kriging deals with non-linear estimation based on $f(Z(\mathbf{s}))$, an arbitrary function of $Z(\mathbf{s})$ (Rendu, 1980; Webster and Oliver, 1989). Provision to address non-normality of the variable of interest also exists, such as in lognormal kriging (Dowd, 1982).

1.3 Sampling design optimization

1.3.1 Conventional model-based case

The conventional approach for model-based sampling design optimization relies on the underlying model of spatial variation that we assume. The random variables $Z(\mathbf{s}_i)$ and $Z(\mathbf{s}_j)$ are not independent, they are correlated as characterized by the parametrized correlogram $\rho(\mathbf{h};\theta)$. In the kriging system, prediction is made using the sample values and the covariance function by minimization of the prediction error variance under the constraint of unbiasedness (Goovaerts, 1997). This allows one to obtain not only the prediction of the variable at unvisited location, but also its associated error variance. The latter might be used to formulate a criterion, which depends on ξ and θ, quantifying the overall quality of the map:

$$\text{criterion} = \frac{1}{|\mathcal{A}|} \int_{\mathbf{s} \in \mathcal{A}} \operatorname{var}\big(Z(\mathbf{s}) - \hat{Z}(\mathbf{s})\big) \mathrm{d}s. \tag{1.1}$$

Several authors (e.g. Delhomme, 1978; Van Groenigen and Stein, 1998) noted that estimation of the kriging variance depends on the variogram $\gamma(\mathbf{h};\theta)$ and the sampling design ξ, not on the actual sample values (McBratney et al., 1981; Brus and Heuvelink, 2007). By knowing the variogram parameters θ, one may then compute the kriging variance from a given sampling design before the actual data collection in the field. Finding the optimal design is therefore simply the minimization of the criterion with respect to ξ, for given parameters θ.

1.3.2 Optimization algorithms

A straightforward solution to the optimization problem is to evaluate all potential sampling configurations exhaustively and select the one achieving the lowest criterion value. This solution is tractable when the search space is small. In practice this is often not the case and one may rather use a numerical search optimization algorithm instead. In this thesis optimization of a criterion is always achieved by minimizing that criterion. This is without loss of generality because maximizing a criterion is the same as minimizing its opposite (i.e. the criterion multiplied with -1). Several numerical optimizations can be adopted, such as greedy algorithms (Baume et al., 2011), genetic algorithms (Behzadian et al., 2009), particle swarm optimization (Jarboui et al., 2007), metaheuristic search (e.g. NSGAII by Deb et al., 2003) and simulated annealing (Kirkpatrick et al., 1983). In this thesis I essentially use the latter, which was extended for spatial optimization by Van Groenigen and Stein (1998).

1.3.3 Spatial Simulated Annealing

Spatial simulated annealing (SSA) works by proposing new sampling configurations based on a random perturbation of one unit of the sample. For each new design, the criterion is evaluated and compared to that of the previous design. Sampling configurations that reduce the criterion are always accepted, configurations that increase the criterion are accepted with a probability that is initially fairly large and decreasing exponentially with the number of iterations. Thus, a worse sampling configuration might be accepted, particularly at the beginning of the iterative procedure. The process is repeated several thousands of times, which leads to a minimum (or maximum) value of the criterion, associated to an optimized sampling design. In the ordinary kriging case, this leads to a design in which the units are distributed fairly evenly over the geographic space (Van Groenigen et al., 1999; Marchant, 2018), with a few units at the boundary of the study area. Extension of the optimization for the KED variance case leads to a distribution of the units according to both geographic and predictors (i.e. covariates) space (Brus and Heuvelink, 2007).

1.3.4 Recent developments

Up to now, most sample optimization studies considered the standard kriging cases reviewed above. But model-based geostatistics developed over time and new advances have emerged. The kriging prediction is based on the variogram, which is characterized by a set of parameters. The latter are commonly estimated. This introduces an additional source of uncertainty which can be added to the kriging prediction error variance. This uncertainty has been quantified in several studies in the last decades, for example by Pardo-Igúzquiza and Dowd (2001), Ortiz and Deutsch (2002) and Marchant and Lark (2004). Clearly, some sampling designs yield more accurate variogram parameter estimates than others. Another key aspect in kriging is the assumption of second-order stationarity. There is a recent trend towards relaxation of the assumption of stationarity in the variance. Some contributions have been made by Pintore and Holmes (2004), Lark (2009) and Fouedjio (2017). In some cases, the assumptions pertaining to the kriging system may simply be avoided by using machine learning models. They emerged as a valuable tool to make "assumption-free" spatial prediction from large sets of environmental covariates (Grimm et al., 2008; Hengl et al., 2015). Another development in recent years is that interpolated maps can be used as input in dynamic environmental models and their associated uncertainty propagated to the model output. This has been done efficiently using Bayesian uncertainty analysis (Kavetski et al., 2006; Huard and Mailhot, 2008; Renard et al., 2011).

However, the implications of recent developments in geostatistical modelling for sampling design optimization have not been thoroughly studied, although some solutions have been found. Lark (2002), Zhu and Stein (2006) and Marchant and Lark (2007a) optimized a sampling scheme for estimation of the variogram parameters. More recently, Lark and Marchant (2018) showed that a sample optimized for minimization of the covariance function parameter uncertainty and kriging variance can be approximated by simply adding close-pair units to a spatial coverage design. Optimal designs for non-stationary variance models have also been documented, such as by Atkinson and LLoyd (2007) and Marchant et al. (2009). These publications show how a non-stationary variance model can be fitted, and how the latter can be used for sampling design optimization. But they did not consider the case where multiple environmental covariates are available to model the variance. Sampling designs for machine learning methods have been barely investigated. A contribution has been made by Tuia et al. (2013), where a monitoring network is optimized using a neural network and active learning. A thorough search of the relevant literature yielded no further results concerning neither the optimization of sample patterns for machine learning nor optimization of the input maps for Bayesian calibration for a dynamic environmental model.

1.3.5 Objectives

I define four topics, each aiming at addressing the optimal sampling design associated with a recent advance in geostatistical modelling and mapping. The objectives comprise a set of research questions, which are addressed in this thesis.

1. Sampling optimization for a non-stationary variance model.
   - How can a geostatistical model account for non-stationarity in the mean and the variance?
   - Does using a non-stationary variance model improve mapping accuracy?
   - How can a sampling design of a non-stationary variance model be optimized?
   - What are the characteristics of the optimized design?
2. Optimal spatial coverage design while accounting for covariance parameter uncertainty and kriging variance.
   - What is a suitable criterion that accounts for both prediction error variance and covariance function parameter uncertainty?
   - How accurate is a spatial coverage design compared to an optimized design?
   - Does allocating 10% of the samples at short distance improve the mapping accuracy?
   - Can a spatial coverage design with close-pair units be recommended as a robust strategy for mapping?
3. Sampling design optimization for mapping with random forest.
   - Can a sampling design be optimized for mapping using random forest?
   - What are the characteristics of the optimized design?
   - How does the optimized design compare to commonly used sampling designs?

   - Can recommendations be made on sampling strategies for mapping using random forest and other machine learning techniques?

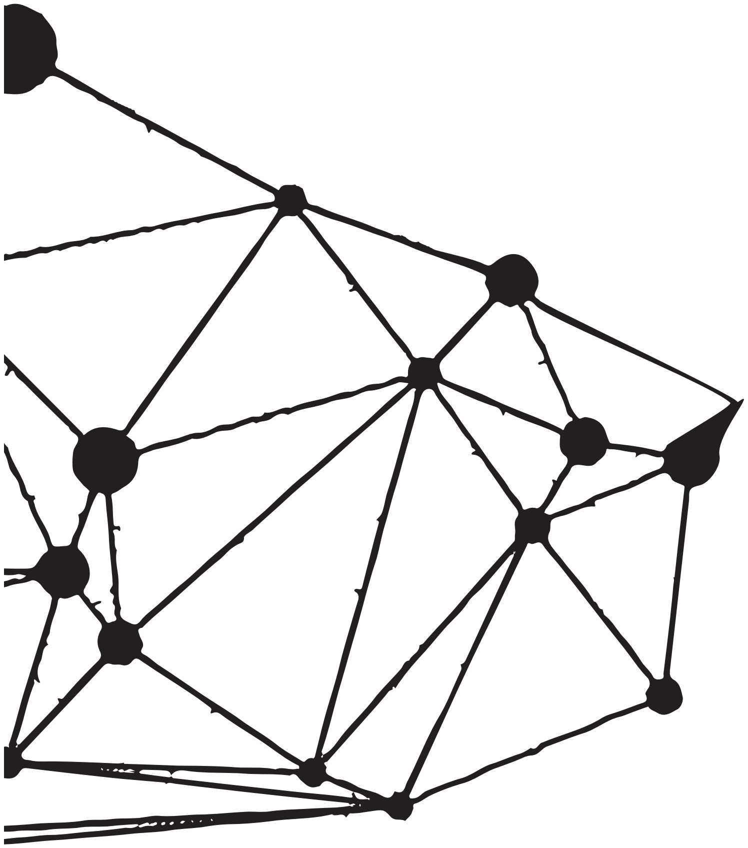
4. Sampling density optimization for Bayesian uncertainty analysis of a rainfall runoff model.
   - Is it possible to include uncertainty about model input, model structure, model parameters and output observations in a simple dynamic environmental model?
   - Does the input measurement density impact the model prediction uncertainty?
   - How does the input uncertainty interact with other sources of uncertainty?
   - Which recommendation can I provide on input measurement density, while accounting for all sources of error in the calibration of a dynamic environmental model?

I will test and illustrate the developed methods for case studies in soil science and hydrology. By chronological order of the thesis chapters, the case studies are: (1) a small agricultural area in the Hunter Valley, Australia, (2) a regional scale area in the North of England, UK, (3) a synthetic case study, (4) 23 European countries covered by the LUCAS dataset and (5) the Thur catchment in North-East Switzerland. While in this thesis I focus on two fields of natural sciences, the methods are applicable to a wider range of application domains in the Earth and environmental sciences.

1.4 Thesis structure

The thesis is organized in seven chapters, including this introduction chapter. In Chapter 2 the geostatistical framework to include non-stationarity in the mean and variance is presented and applied to a soil science case study. The results are evaluated and compared to that of a stationary variance model. In Chapter 3 the non-stationary variance model is used as a basis for sampling design optimization. The design is optimized for predicting daily rainfall for a case study in England. Chapter 4 presents a study where a spatial coverage design with close-pair units are compared to that of a design optimized for the kriging variance and variogram uncertainty. A synthetic study is presented as well as a case study using an average variogram of soil clay. In Chapter 5 a European sampling design is optimally thinned for mapping a soil variable using the random forest machine learning algorithm. The spatial distribution of optimized designs is compared to that of commonly used model-based sampling designs. Chapter 6 addresses optimization of the rain gauge sampling density for a dynamic rainfall-runoff model, while accounting for model structural, initial state, model parameters and model output measurements uncertainty. Chapter 7 gives the conclusion of this thesis and recommendations for future research.

Chapters 2 to 6 can be read separately. Chapters 2 to 4 are published or accepted peer-reviewed publications while Chapters 5 and 6 are submitted to refereed journals. Literature references for all chapters have been combined at the end of this thesis.



Chapter 2

Accounting for non-stationary variance in geostatistical mapping of soil properties

Simple and ordinary kriging assume a constant mean and variance of the soil variable of interest. This assumption is often implausible because the mean and/or variance are linked to terrain attributes, parent material or other soil forming factors. In kriging with external drift (KED) non-stationarity in the mean is accounted for by modelling it as a linear combination of covariates. In this study, we applied an extension of KED that also accounts for non-stationary variance. Similar to the mean, the variance is modelled as a linear combination of covariates. The set of covariates for the mean may differ from the set for the variance. The best combinations of covariates for the mean and variance are selected using Akaike's information criterion. Model parameters of the selected model are then estimated by differential evolution using the restricted maximum likelihood (REML) in the objective function. The methodology was tested in a small area of the Hunter Valley, NSW Australia, where samples from a fine grid with gamma K measurements were treated as measurements of the variable of interest. Terrain attributes were used as covariates. Both a non-stationary variance and a stationary variance model were calibrated. The mean squared prediction errors of the two models were somewhat comparable. However, the uncertainty about the predictions was much better quantified by the non-stationary variance model, as indicated by the mean and median of the standardized squared prediction error and by accuracy plots. We conclude that the non-stationary variance model is more flexible and better suited for uncertainty quantification of a mapped soil property. However, parameter estimation of the non-stationary variance model requires more attention due to possible singularity of the covariance matrix.

Based on:
Wadoux, A. M. J.-C., Brus, D. J. and Heuvelink, G. B. M. (2018). Accounting for non-stationary variance in geostatistical mapping of soil properties. *Geoderma*, **324**, pp.138-147.

2.1 Introduction

Standard geostatistical mapping approaches predict a soil variable of interest at the unsampled nodes of a fine grid using measurements of this variable at sampling locations. In many cases predictions can be improved by exploiting a relation between the soil variable and one or more environmental covariates of which maps are available, such as terrain attributes derived from a digital elevation model and remote sensing images. This is usually done by modelling the soil variable as the sum of a linear combination of covariates and a spatially autocorrelated residual. This leads to kriging with external drift (KED) (Goovaerts, 1997). In situations where the covariates explain a considerable part of the variation of the soil variable, KED is superior to simple or ordinary kriging that both assume that the mean of the soil variable is constant within a global or local neighbourhood and not dependent on covariates.

In KED we allow for a non-stationary mean, but the variance is assumed stationary (i.e. constant). More specifically, it is assumed that the covariance between the soil variable Z at two locations $\mathbf{s}$ and $\mathbf{s} + \mathbf{h}$ only depends on the separation vector $\mathbf{h}$: $\mathrm{cov}(Z(\mathbf{s}), Z(\mathbf{s} + \mathbf{h})) = C(\mathbf{h})$. Taking $\mathbf{h} = \mathbf{0}$ shows that the variance is assumed constant: $\mathrm{var}(Z(\mathbf{s})) = C(\mathbf{0})$ for all $\mathbf{s}$. However, in many cases the assumption of a stationary variance may be implausible, i.e. when the residual spatial variation is substantially different in different parts of the study area. For instance, McBratney and Webster (1981) identified several discontinuities in the variograms of soil colour and pH along a transect in north-east Scotland. The authors attributed the changes to boundaries between soil types. Similarly, Voltz and Webster (1990) found important differences between topsoil clay content variograms of contrasting Jurassic sediments.

In some cases, non-stationarity in the variance can be solved by transforming the data prior to geostatistical modelling, e.g. by a square-root or log-transformation (e.g. Jacques et al., 1999). Several solutions have been proposed in case a transformation does not solve the problem. Pintore and Holmes (2004) and later Haskard and Lark (2009) proposed to account for a non-stationary variance by spectral tempering. The method tempers a spectrum based on a stationary correlation matrix, but the modelled covariance structure can vary spatially while maintaining positive-definiteness. The authors showed that modifying the spectrum of the data according to a covariate on a transect gave a more realistic variance model for a case study on rates of emission of nitrous oxide from soils. Alternatively, McBratney and Minasny (2013) proposed to equalize variogram parameters by deformation of the geographic space. This method renders a stationary covariance function in the transformed space. Spatial predictions made in the transformed space are then back-transformed to the original geographic space. However, while this approach addresses differences in spatial correlation, it does not solve the non-stationary variance problem.

The work presented here builds on the work of Lark (2009) and Marchant et al. (2009). They demonstrated how a model in which the variance is a function of the spatial coordinate or covariates can be fitted by REML, and how such model can be used in geostatistical prediction of soil properties. The same approach is applied by Brus et al. (2016) in three-dimensional soil property mapping. They assumed that the residual variance is a stepwise or continuous function of depth, while in the horizontal plane, at a given depth, the residual variance was assumed constant.

The objective of this study is to test the approach proposed by Lark (2009) in a case study where several covariates are available for modelling the non-stationarity of the mean and variance. The best stationary variance model is compared with the best non-stationary variance model, using evaluation criteria that measure both the quality of the predictions as well as the quality of the estimated prediction uncertainty.

2.2 Statistical methodology

2.2.1 Model definition

A soil variable of interest Z at any location $\mathbf{s}$ in the study area $\mathcal{A}$ is modelled by:

$$Z(\mathbf{s}) = m(\mathbf{s}) + \sigma(\mathbf{s})\varepsilon(\mathbf{s}) \tag{2.1}$$

where $m(\mathbf{s})$ is the mean at location $\mathbf{s}$, $\sigma(\mathbf{s})$ the standard deviation at location $\mathbf{s}$ and ε a stationary, spatially correlated Gaussian random field with zero mean and unit variance. The mean m and standard deviation σ are deterministic functions that are modelled as linear combinations of covariates, unconditional on the observations:

$$Z(\mathbf{s}) = \sum_{k=0}^{K} \beta_k w_k(\mathbf{s}) + \sum_{l=0}^{L} \kappa_l g_l(\mathbf{s})\varepsilon(\mathbf{s}) \tag{2.2}$$

where the β_k and κ_l are regression coefficients (the latter are used for modelling the standard deviation), and the w_k and g_l spatially distributed covariates. We take $w_0(\mathbf{s}) = g_0(\mathbf{s}) = 1$ for all $\mathbf{s}$, so that β_0 and κ_0 are space-invariant constant contributions to the mean and standard deviation, respectively.

Let Z be measured at n locations $\mathbf{s}_i$ $(i = 1, \ldots, n; \mathbf{s}_i \in \mathcal{A})$. The measurements $z(\mathbf{s}_i)$ are treated as realizations of the Gaussian field Z and prediction is done for Z at a new, unobserved location $\mathbf{s}_0$. Stacking the $z(\mathbf{s}_i)$ in a (column) vector $\mathbf{z}$ and changing to matrix notation yields:

$$\mathbf{z} = \mathbf{W}\boldsymbol{\beta} + \mathbf{H}\boldsymbol{\varepsilon} \tag{2.3}$$

where $\mathbf{W}$ is the $n \times (K+1)$ design matrix of covariates for the mean at the observation locations, $\boldsymbol{\beta}$ is the $(K+1)$ vector of regression coefficients for the mean and $\boldsymbol{\varepsilon}$ is the n-vector of (standardized) residuals, which has variance-covariance matrix $\mathbf{R}$. $\mathbf{H}$ is an $n \times n$ diagonal matrix defined by:

$$\mathbf{H} = \mathrm{diag}\{\mathbf{G}\boldsymbol{\kappa}\} \tag{2.4}$$

where $\mathbf{G}$ is the $n \times (L+1)$ matrix of standard deviation covariates at observation locations and $\boldsymbol{\kappa}$ is an $(L+1)$ vector of standard deviation regression coefficients. Note that while $\boldsymbol{\varepsilon}$ has variance-covariance matrix $\mathbf{R}$, the stochastic component $\mathbf{H}\boldsymbol{\varepsilon}$ of Eq. 2.3 has variance-covariance matrix $\mathbf{C} = \mathbf{HRH}'$. The parameters of the model defined by Eq. 2.3 are $\boldsymbol{\beta}$, $\boldsymbol{\kappa}$ and the parameters of a model for the spatial autocorrelation of the standardized residual. In this work we will parametrize the spatial autocorrelation by an isotropic exponential correlogram $r(h) = r_0\{\exp(-\frac{h}{a})\}$ (where $h > 0$ is the Euclidean distance between two locations, by definition $r(0) = 1$), thus introducing two more parameters, namely r_0 and a. Parameter r_0 equals one minus the nugget-to-sill ratio, while a refers to the spatial correlation length (or range, $3a$ being the effective range). Note that the stationary variance model is a special case of the non-stationary variance model. It is obtained by setting parameters $\kappa_l, l = 1 \ldots L$ to zero, so that $\sigma(s) = \kappa_0$ for all $\mathbf{s}$.

2.2.2 Parameter estimation and model selection

Parameter estimation - In estimation the parameters are subdivided in two subsets, the regression coefficients $\boldsymbol{\beta}$ for the mean, and all parameters of the stochastic part of the model, $\Phi = [\boldsymbol{\kappa}, r_0, a]$. For a stationary variance model the second subset reduces to $\Phi = [\kappa_0, r_0, a]$. The standard maximum likelihood estimates of Φ depend non-linearly on the regression coefficients for the mean $\boldsymbol{\beta}$, which introduces a bias in the estimates of Φ if both parameter subsets are estimated jointly (Lark and Webster, 2006). This problem can be avoided by restricted (or residual) maximum likelihood (REML) parameter estimation. REML first estimates Φ and next $\boldsymbol{\beta}$. Similar to standard maximum likelihood estimation, REML aims to find the vector Φ for which the observed data yield the highest probability density (i.e. likelihood, if treated as a function of the parameters instead of the data). The problem is that the likelihood of Φ depends on the regression coefficients for the mean, which are unknown and must also be estimated. Patterson and Thompson (1971) solved this problem by detrending the data by multiplying the data vector by a projection matrix. The new variable is a function of the original variable but independent of the regression coefficients for the mean. The associated restricted log-likelihood function is given by (Webster and Oliver, 2007):

$$L_r(\Phi|\mathbf{z}) = \text{constant} - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\log|\mathbf{W}'\mathbf{C}^{-1}\mathbf{W}| - \frac{1}{2}\mathbf{z}'\mathbf{P}'\mathbf{C}^{-1}(\mathbf{I}-\mathbf{Q})\mathbf{z} \tag{2.5}$$

where **I** is an identity matrix and **P** and **Q** are defined as:

$$\mathbf{P} = \mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}' \tag{2.6}$$

$$\mathbf{Q} = \mathbf{W}(\mathbf{W}'\mathbf{C}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{C}^{-1} \tag{2.7}$$

After estimating Φ by maximizing the restricted log-likelihood given in Eq. 2.5 above, the regression coefficients $\boldsymbol{\beta}$ for the mean can be estimated by generalized least squares (GLS):

$$\boldsymbol{\beta} = (\mathbf{W}'\mathbf{C}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{C}^{-1}\mathbf{z} \tag{2.8}$$

Here, matrix **C** is computed from the optimized values for Φ. Note that the regression coefficients κ_l in Eq. 2.2 can be positive or negative, as long as the covariance matrix **C** is not singular.

Model selection - Two subsets of covariates must be chosen, one for the mean and one for the standard deviation. Suppose we have in total K candidate covariates for modelling the mean. For a subset of covariates of size k, there are $\binom{K}{k}$ possible combinations. Since the size is not fixed, we have $\sum_{k=0}^{K}\binom{K}{k}$ possible models in total for the stationary variance model. Using the same set of candidate covariates for the standard deviation, for the non-stationary variance model the total number of models equals $\left(\sum_{k=0}^{K}\binom{K}{k}\right)^2$. Ideally, all model combinations are fitted and compared. Models with different sets of covariates for the mean cannot be compared on the basis of the restricted log-likelihood, because this is a function of just the covariance parameters Φ. Models can better be compared on the basis of the standard log-likelihood, using a likelihood ratio test or by comparing quality measures that are functions of the log-likelihood and the number of model parameters. Common quality measures are the Akaike information criterion (AIC, Akaike, 2011) and Bayesian information criterion (BIC, Kass and Wasserman, 1995). AIC, used in the case study hereafter, is defined as:

$$\mathrm{AIC} = 2p - 2\log L \tag{2.9}$$

where p is the number of estimated parameters and $L = p(\mathrm{z}|\boldsymbol{\beta}, \Phi)$ is the ordinary log-likelihood, given by (Diggle and Ribeiro (2007b), Eq. 5.13):

$$L(\boldsymbol{\beta}, \Phi|\mathrm{z}) = -\frac{1}{2}n\log(2\pi) - \frac{1}{2}\log|\mathrm{C}| - \frac{1}{2}(\mathrm{z} - \mathbf{W}\boldsymbol{\beta})'\mathrm{C}^{-1}(\mathrm{z} - \mathbf{W}\boldsymbol{\beta}) \tag{2.10}$$

The number of covariate combinations for the non-stationary variance model will be very large, unless the number of covariates K is small. Often, one will therefore resort to numerical search algorithms (e.g. greedy algorithms) to overcome large computation times. In the case study presented in Section 2.3, the small number of covariates K makes exhaustive search possible. Note that models are compared based on the AIC using the ordinary log-likelihood, while prediction is made with parameters estimated by restricted log-likelihood (Hoeting et al., 2006).

2.2.3 Kriging

Ignoring estimation errors in $\boldsymbol{\Phi}$ allows us to use a standard result from universal kriging (Webster and Oliver, 2007) to predict the soil variable of interest at a new, unobserved locations $\mathbf{s}_0$:

$$\hat{z}(\mathbf{s}_0) = (\boldsymbol{c}_0 + \mathbf{W}(\mathbf{W}'\mathrm{C}^{-1}\mathbf{W})^{-1}(\boldsymbol{w}_0 - \mathbf{W}'\mathrm{C}^{-1}\boldsymbol{c}_0))'\mathrm{C}^{-1}\mathrm{z} \tag{2.11}$$

where $\boldsymbol{w}_0$ is a vector of covariates for the mean at the prediction location and $\boldsymbol{c}_0$ is an n vector of covariances between the residuals at the observation and prediction location. Note that these are covariances of the (unstandardized) residuals $\sigma\cdot\varepsilon$ and thus depend on the standard deviation covariates $g_l(\mathbf{s}_0)$, their associated (estimated) regression coefficients κ_l and the correlogram of ε.

The associated prediction error is given by (Webster and Oliver, 2007):

$$\begin{aligned}\mathrm{var}(Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0)) =& (\mathrm{g}(\mathbf{s}_0)'\boldsymbol{\kappa})^2 - \boldsymbol{c}_0'\mathrm{C}^{-1}\boldsymbol{c}_0 \\ &+ (\boldsymbol{w}_0 - \mathbf{W}'\mathrm{C}^{-1}\boldsymbol{c}_0)'(\mathbf{W}'\mathrm{C}^{-1}\mathbf{W})^{-1}(\boldsymbol{w}_0 - \mathbf{W}'\mathrm{C}^{-1}\boldsymbol{c}_0)\end{aligned} \tag{2.12}$$

where $(\mathrm{g}(\mathbf{s}_0)'\boldsymbol{\kappa})^2$ is the variance of $Z(\mathbf{s}_0)$. The first two terms on the right-hand side of Eq. 2.12 quantify the prediction error variance of the residuals, while the last term is the variance of the estimation error of the mean. Note that here we take uncertainty about the β_k into account, whereas uncertainty about the κ_l and variogram parameters r_0 and a is ignored. Taking the latter uncertainties into account is beyond the scope of this work.

2.2.4 Quality of prediction and estimated uncertainty

Let there be $(N - n)$ validation locations $\mathbf{s}_i, i = (n + 1) \dots N$. In the case study discussed in Section 2.3, N is the number of nodes of a fine grid covering the study area, of which n are used for model calibration and $(N - n)$ for model evaluation. To quantify the quality of the predictions we computed the mean prediction error (ME), root mean squared prediction error (RMSE) and modelling efficiency coefficient (MEC). The latter is derived as follows (Janssen and Heuberger, 1995):

$$\text{MEC} = 1 - \frac{\sum_{i=n+1}^{N} (z(\mathbf{s}_i) - \hat{z}(\mathbf{s}_i))^2}{\sum_{i=n+1}^{N} (z(\mathbf{s}_i) - \bar{z})^2} \tag{2.13}$$

where $\bar{z}$ denotes the mean of the observations. MEC quantifies the improvement of the model made over using the mean of the observations as a predictor. MEC can become negative, while its optimal value is one.

To evaluate the quality of the prediction error variance, we used the standardized squared prediction error θ (Lark, 2000; Marchant et al., 2009):

$$\theta(\mathbf{s}_i) = \frac{(z(\mathbf{s}_i) - \hat{z}(\mathbf{s}_i))^2}{\text{var}(Z(\mathbf{s}_i) - \hat{Z}(\mathbf{s}_i))}, \quad i = (n + 1) \dots N \tag{2.14}$$

For a model giving unbiased predictions and correct estimates of the prediction error variance, θ has a χ^2-distribution with one degree of freedom, so that the average value of θ over all validation locations should be close to 1, while its median should be close to 0.455.

Accuracy plots

Deutsch (1997) and Goovaerts (2001) proposed a visual assessment of the quality of the estimated prediction uncertainty through a so-called accuracy plot. Since the prediction error at each validation location s_i is normally distributed with zero mean and known variance (i.e. the kriging variance given by Eq. 2.12), its cumulative distribution F_i is known too. From this one can easily compute for a given probability p a symmetric interval around the predicted value through computing the (1-p)/2 and (1+p)/2 quantiles, and use these as the lower and upper bounds of a prediction interval. This is done for a series of values for p. The proportion of validation locations at which the p prediction interval includes the observed value is then obtained by:

$$\bar{\xi}(p) = \frac{1}{(N-n)} \sum_{i=n+1}^{N} \xi(\mathbf{s}_i; p) \quad \forall p \in [0,1] \tag{2.15}$$

where $\xi(\mathbf{s}_i; p)$ is given by:

$$\xi(\mathbf{s}_i; p) = \begin{cases} 1 & \text{if } F_i^{-1}(\mathbf{s}_i; (1-p)/2) < z(\mathbf{s}_i) \leq F_i^{-1}(\mathbf{s}_i; (1+p)/2) \\ 0 & \text{otherwise} \end{cases} \tag{2.16}$$

A correct modelling of the uncertainty would entail that the proportion of validation locations where the p prediction interval covers the observed value approximately equals the nominal value p, for all values of p. These proportions are plotted in a scattergram against p; such scattergram is referred to as an accuracy plot. Ideally, the points in the accuracy plot are on the 1:1 line. The absolute deviation from this line can be summarized by an integral, calculated as:

$$A = \int_0^1 |\bar{\xi}(p) - p| \mathrm{d}p \tag{2.17}$$

Ideally, $A = 0$ for a model that perfectly describes the uncertainty. Note that the measure A does not separate over- and underestimation of the uncertainty. Therefore we also derive P_O, the proportion of A that is above the 1:1 line (overestimation of uncertainty), given by:

$$P_O = \frac{1}{A} \int_0^1 \max\{0, \bar{\xi}(p) - p\} \mathrm{d}p \tag{2.18}$$

and P_U, the proportion of A below the 1:1 line (underestimation of uncertainty):

$$P_U = \frac{1}{A} \int_0^1 \max\{0, p - \bar{\xi}(p)\} \mathrm{d}p \tag{2.19}$$

Note that P_O and P_U sum to 1.

2.3 Case study

2.3.1 Study area and data

We tested the methodology in the 140 ha Scarborough area (Fig. 2.1), located in the Hunter Valley, Australia. Elevation ranges from 86 to 144 m above sea level with an average of 113 m. Radiometric gamma was surveyed using a vehicle-born

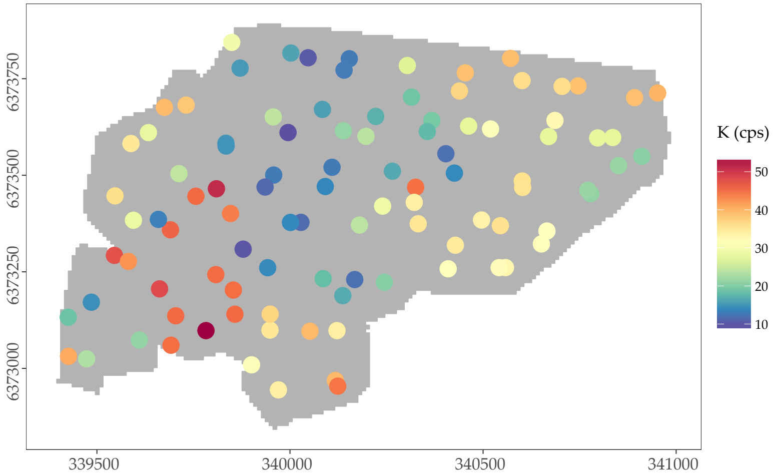


***Figure 2.1** – Scarborough study area in the Hunter Valley, Australia, with location and values of gamma-radiomatric potassium at 100 locations.*

passive gamma spectrometer. This yielded a raster file of 10 m × 10 m resolution of gamma-radiometric potassium (K) expressed in cps. The data collection procedure is detailed in Stockmann et al. (2012).

We used the R package spcosa (Walvoort et al., 2010) to divide the area into 50 compact geostrata of equal size, from which we selected two locations per stratum: one in the centre and one randomly. This resulted in a total of 100 sampling locations for the whole area (out of 10,473 grid locations), shown in Fig. 2.1.

In addition to gamma K observations at the 100 sampling locations, three covariates were derived from the 10 m resolution digital elevation model (DEM):

- Topographic wetness index (TWI) (Fig. 2.2b), which is the steady state wetness index, based on Moore et al. (1993).
- Slope (Fig. 2.2c), which is the angle of inclination of the soil surface from the horizontal, derived using a 3 × 3 window and the method of Zevenbergen and Thorne (1987).
- Combined curvature (Fig. 2.2d), which is a combination of profile and planform curvature, based on Moore et al. (1993).

2.3.2 Practical implementation

Parameter estimation and model selection - All covariates were centred on 0 and scaled to a standard deviation of 1, to enable direct comparison of the associated

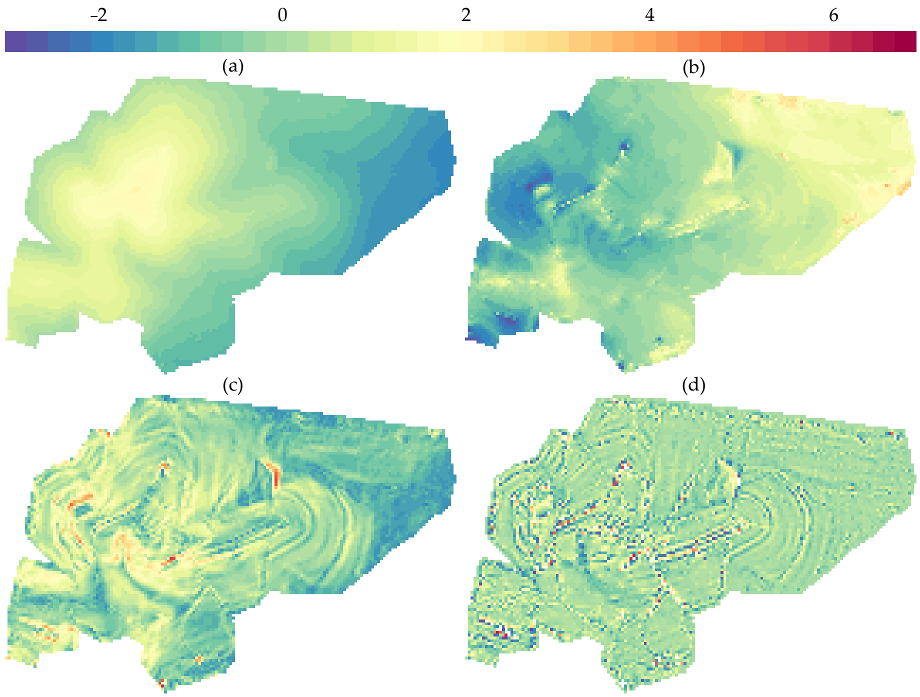


***Figure 2.2** – Standardized covariates used in model selection: (a) DEM, (b) topographic wetness index (TWI), (c) slope and (d) combined curvature.*

regression coefficients. The four candidate covariates chosen in Section 2.2.2 were used to compute the best stationary variance model and the best non-stationary variance model using the AIC criterion. Since there were only four candidate covariates the total number of models to be compared was only 16 for the stationary variance model and 256 for the non-stationary variance model, allowing exhaustive search. The global optimum of the restricted log-likelihood function was found using differential evolution (Storn and Price, 1997), implemented in the R package DEoptim (Ardia et al., 2015). The convergence threshold was fixed at 10^{-10}. Calculations were done in parallel on a standard desktop with eight cores. REML estimation of the model parameters for all combinations of covariates (272 models in total) took approximately 35 hours. The standard deviation parameters were bounded to large positive and negative values to speed up computation (intercept between -50 and 50 and coefficients between -100 and 100). Likewise, variogram parameters were constrained within plausible ranges (the short distance correlation parameter r_0 was forced between 0 and 1 and the correlation length parameter a between 0 and one-third of the extent of the study area). In addition, any proposal combination of parameters in DEoptim that resulted in a near-singular or singular C matrix was

***Table 2.1** – Estimated coefficients for the mean and standard deviation and variogram parameters for the stationary variance and non-stationary variance models.*

Parameter	Associated with	Estimated value
Stationary variance model		
β_0 (cps)	Intercept for the mean	29.948
β_1 (m)	Elevation	−1.688
κ_0 (-)	Standard deviation	11.98
r_0 (-)	Short-distance correlation parameter	1.000
a (m)	Range parameter	149.2
Non-stationary variance model		
β_0 (cps)	Intercept for the mean	27.274
β_1 (%)	TWI	4.583
κ_0 (-)	Intercept for the standard deviation	18.34
κ_1 (m)	Elevation	7.115
κ_2 (%)	Slope	−2.161
r_0 (-)	Short-distance correlation parameter	0.999
a (m)	Range parameter	544.0

rejected. This problem is discussed more extensively in the Discussion.

Kriging - Predictions were made with global point kriging on a 10 m × 10 m resolution grid, excluding the 100 observation locations. As far as we know there are no existing R packages for kriging with non-stationary variance, so we implemented this in our own R script. The R script and a test case are available in Sawicka et al. (2017). The prediction to the 10,373 remaining grid cell centres ($N - n$) took less than a minute on a standard desktop computer.

2.3.3 Results

Based on the procedure detailed in Section 2.2.2, elevation was chosen as a covariate for the mean in the stationary variance model. For the non-stationary variance model, TWI was chosen for the mean while elevation and slope were chosen as covariates for the standard deviation. Table 2.1 presents the estimated coefficients for the mean, standard deviation and variogram parameters for the two models. Recall that the covariates are standardized to allow comparison between regression coefficients. Elevation has a negative effect on the mean gamma K for the stationary variance model. The stationary variance model also indicates strong residual spatial correlation at short distances (r_0 equals one), but the correlation decreases rapidly with distance (range parameter of 149.2 m for an area with an extent of about 1,500 m in the East-West direction). The coefficient for the mean of the non-stationary variance model exhibits a strong positive relationship of gamma K with the wetness index (β_1 of about 4.6). The sign of the standard deviation coefficients κ_1 and κ_2 show that their associated covariates are positively and negatively correlated with the unconditional standard deviation, respectively. Elevation has a larger impact on the standard deviation (κ_1 = 7.115) but slope also has an important (negative) contribution (κ_2 = -2.161). In contrast with the stationary variance model, the range parameter of the non-stationary variance model is much larger (a = 544 m), while having a similar short-distance correlation parameter.

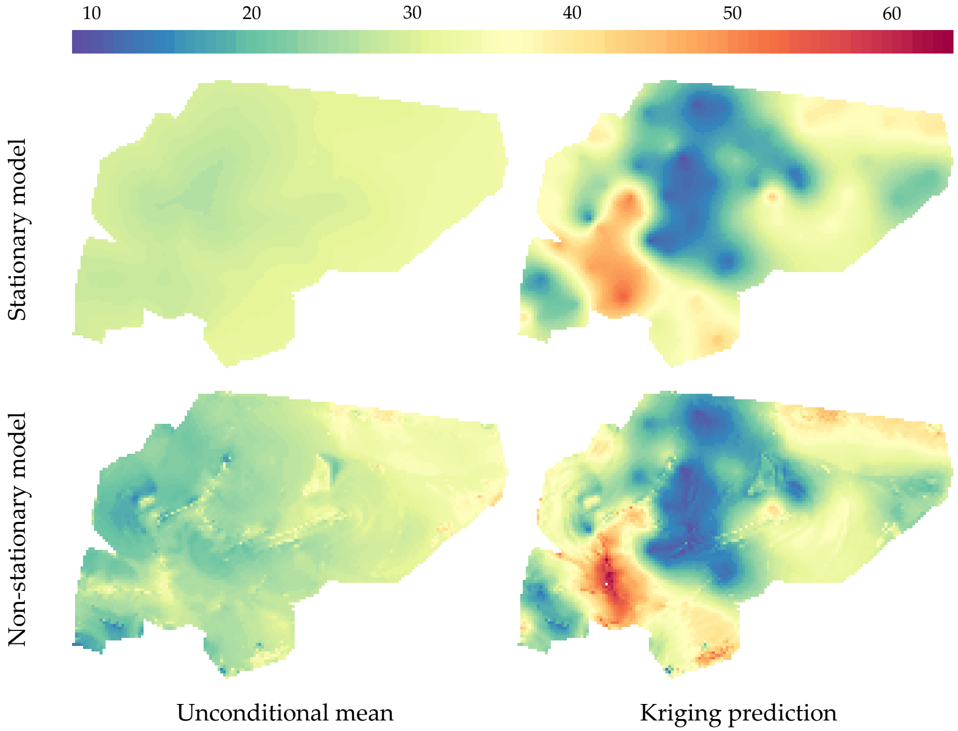


***Figure 2.3** – Maps of the unconditional mean and kriging prediction.*

The maps of the (unconditional) means show large differences between the two models, both in terms of the magnitude of spatial variation and the spatial pattern (Fig. 2.3). There is more spatial variation on the map obtained with the non-stationary variance model. The difference in spatial pattern is due to the use of different covariates: elevation was used for the stationary variance model and wetness index for the non-stationary variance model. Despite the large differences in unconditional mean, the kriging prediction maps obtained with the two models are quite similar (Fig. 2.3). They have comparable ranges of predicted values and a similar spatial pattern. The non-stationary variance model exhibits greater fine-scale detail than the stationary variance model.

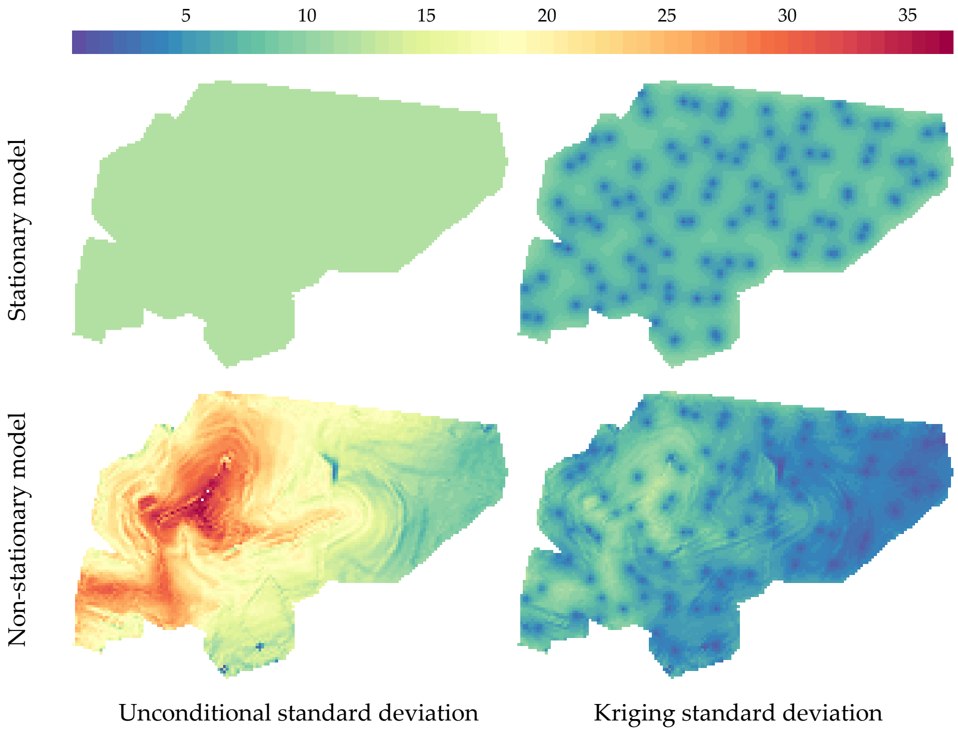


Figure 2.4 *– Maps of the unconditional standard deviation and kriging standard deviations.*

The spatial patterns of the unconditional standard deviation maps of the two models are clearly different (Fig. 2.4). The unconditional standard deviation of the stationary variance model is constant (equal to 11.98), while for the non-stationary variance model it varies smoothly through the area with high values in the East and low values in the West. The higher values are about five times greater than the low values, which clearly indicates that gamma K has non-stationary variance. The standard deviation shows a spatial trend due to the prominent effect of elevation, see Fig. 2.2a. This is refined at the local scale by the slope, where a large slope value leads to a smaller standard deviation (such as the two patches in the south of the area in Fig. 2.2c and Fig. 2.4). The kriging standard deviation map of the stationary variance model has a familiar pattern with circular areas around observation locations with relatively low values. Uncertainty quickly increases with distance from the sampling locations because of the small range parameter (Table 2.1). The kriging standard deviation map of the non-stationary variance model shows the combined effect of the East-West standard deviation trend and the circular areas with low values near observation locations. On average, the kriging standard deviation of the non-stationary variance model is considerably smaller than that of the stationary variance model (see the mean kriging variance (MKV) in Table. 2.2). This is confirmed by the θ statistics, which suggest that the stationary variance model severely overestimates the true uncertainty (see Section 2.3.4 below). For comparison, Fig. 2.5 shows a map of the local variance of the observation residuals of the non-stationary variance model.

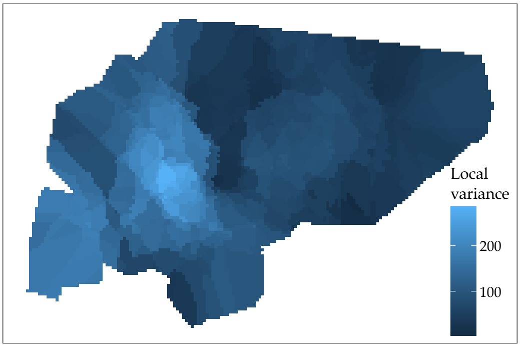


***Figure 2.5** – Local variance of the observation residuals of the non-stationary variance model, calculated using twelve nearest neighbour observations as implemented in the R package RANN (Arya et al., 2017).*

2.3.4 Quality of predictions and prediction error variance

Table 2.2 shows that the log-likelihood (Eq. 2.10) of the non-stationary variance model is larger than that of the stationary variance model, while the AIC (Eq. 2.9) is smaller despite the larger number of model parameters. The non-stationary variance model provides a lower model efficiency measure (MEC of 0.692 against 0.714 for the stationary variance model) and a slightly larger accuracy measure (RMSE of 6.18 against 5.96 for the stationary variance model) but with a ME almost 20% smaller. This implies that the variance of the non-stationary variance model error is larger, because the mean squared error is the sum of the variance and the squared mean error.

The mean of θ (standardized squared prediction error) is closer to 1 for the non-stationary variance model. Figure 2.6 shows the spatial pattern of the mean of θ for both models, computed using a local window. The mean θ of the stationary variance model has a clear spatial pattern, with an strong underestimation of the variance in the eastern part and a overestimation in the western part of the area. The pattern is much more smoothed for the non-stationary variance model, even though there remains a slight underestimation of the variance in the mid-eastern part of the area.

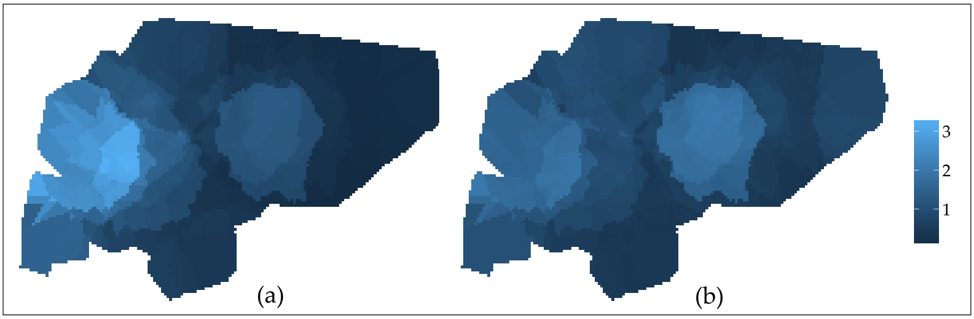


Figure 2.6 *– Maps of the local mean θ for the (a) stationary variance model and (b) non-stationary variance model, calculated using twelve nearest neighbour observations and implemented with the R package RANN (Arya et al., 2017).*

Table 2.2 *– Performance indicators for the stationary and non-stationary variance models.*

	Stationary variance model	Non-stationary variance model
AIC	1667	756.2
Log-likelihood	−828.4	−371.2
MKV	49.4	37.4
ME	−0.84	−0.68
RMSE	5.96	6.18
MEC	0.714	0.692
mean(θ)	0.718	1.126
median(θ)	0.218	0.395
A	0.094	0.016
P_O	1.000	0.914
P_U	0.000	0.086

The median of θ indicates a large improvement of the non-stationary variance model over the stationary variance model. The 0.2184 value of the median θ for the stationary variance model indicates that this model seriously over-estimates the prediction error variance. This is confirmed by the accuracy plots (Fig. 2.7): prediction intervals as obtained with the stationary variance model are too wide, leading to larger absolute deviation from the nominal p values (the absolute deviation parameter A of the stationary variance model is six times larger than that of the non-stationary variance model). The over- and underestimation measures P_O and P_U show that for both models the main problem is overestimation of the uncertainty, although the mean θ and Fig. 2.7 suggest that for the non-stationary variance model, this is likely a chance effect and not significant. This problem is more severe for the stationary variance model, where absolute deviation from the nominal p is almost entirely due to overestimation of the uncertainty.

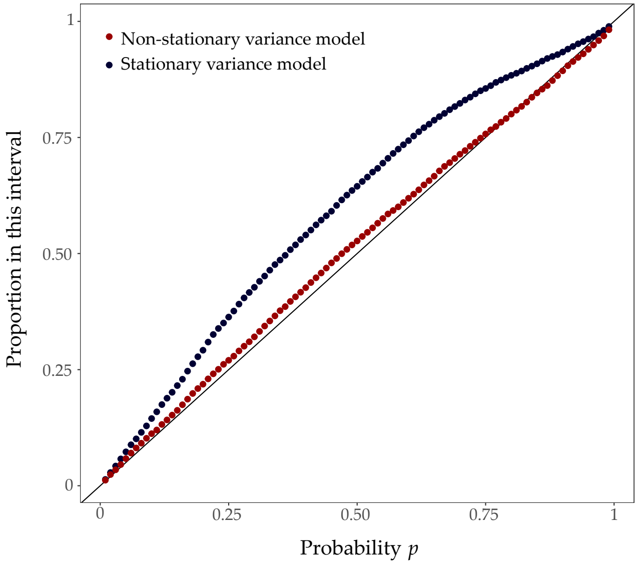


***Figure 2.7** – Accuracy plot for the stationary variance and non-stationary variance models.*

2.4 Discussion

For gamma K prediction, stationary and non-stationary variance models have very different unconditional mean maps. This is because different covariates were selected. Gamma K is linked to elevation for the stationary variance model and to topographic wetness for the non-stationary variance model. This difference may be explained by the large number of soil forming factors influencing spatial variation of gamma K. For example, Viscarra Rossel et al. (2007) showed that gamma K is mostly determined by soil minerals and soil particle size, with some effects of soil moisture and bulk density. In our case study, the area is covered by a wide range of parent materials, such as lithic sandstone, siltstone, mudstone, shale, limestone and volcanic rocks; as well as a large number of soil types such as Red Dermosols, followed by Brown, Black and Grey Dermosols (Kovac and Lawrie, 1991; Stockmann et al., 2012). These contrasting soil types and parent materials lead to large spatial variation of soil mineralogy and soil particle size, as well as soil moisture characteristics. None of these factors were directly included as a covariate, although each may have a (complex) relationship with the four covariates included in this study. It is therefore difficult to draw conclusions regarding a possible causal effect of the selected mean covariates on the gamma K distribution.

In spite of the apparent differences in the unconditional means between the two models, the final prediction maps were nearly the same and quite different from the unconditional means. This shows that in this study the kriging step is important. It not only improves prediction accuracy but also compensates for the differences made in modelling the unconditional mean, thus producing more robust predictions that are less sensitive to choices made during the model selection process.

Elevation and slope were selected to model the standard deviation of the non-stationary variance model. Elevation had a positive and slope a negative effect. Apparently, residual variation is greater at high elevation and shallow slopes. Stockmann et al. (2012) notes that the top-of-hill vineyards were irrigated during the survey. This might have led to an increase of the local variance with elevation, since TWI does not account for this unexpected artificial process (reflected in Fig. 2.5). The role of slope on the standard deviation is more difficult to explain, and we do not wish to speculate. In fact, many authors have noted that interpretation of empirical digital soil mapping models is difficult (e.g. Bishop and McBratney, 2001). Recent work, such as by Angelini et al. (2017), have made a step towards “conscious” digital soil mapping, where the selection of covariates and their role within the model are primarily based on a soil-landscape conceptual model. While this is challenging for the mean, it is even more difficult to explain how local soil spatial variation (i.e. the standard deviation) is influenced by covariates. If non-stationary variance models gain popularity in future digital soil mapping research, then pedological interpretation of the selected models, including the structure of the standard deviation, requires attention.

For the case study, both models provide good predictive ability as shown by the ME, RMSE and MEC. The spatial patterns of the prediction maps also closely resembles those produced by others (e.g. McBratney and Minasny, 2013). In spite of the good performance, the stationary variance model did not provide satisfactory results regarding uncertainty quantification. The low median of θ and the accuracy plots in Fig. 2.7 show that the stationary variance model systematically over-estimates the local standard deviation for most prediction intervals, except at the tails (0.1 and 0.9 predictive intervals). This is not surprising as the study area reveals strong non-stationarity (Fig. 2.5), which the stationary variance model cannot capture. The non-stationary variance model is more flexible and can address local differences in standard deviation. The median θ and accuracy plot statistics show that uncertainty quantification is significantly improved using this model. Given the substantial differences between the two kriging standard deviation maps shown in Fig. 2.4, this has important implications, such as for uncertainty propagation (Heuvelink, 1998) and sampling design optimization studies (Wadoux et al., 2017).

The non-stationary variance model did not improve the accuracy of the predictions.

The RMSE of the non-stationary variance model is slightly greater than that of the stationary variance model and the MEC is slightly smaller. The mean kriging variance is about 25% smaller, suggesting that the non-stationary variance model is more accurate than the stationary variance model, but this merely reflects that the stationary variance model systematically overestimated the uncertainty. The smaller RMSE of the stationary variance model suggests that having a more flexible variance leads to slightly worse predictions. We investigated this by comparing results of the non-stationary model with those of its sub-models (including the stationary version). The results (not shown) confirm that the non-stationary model provides slightly worse predictions than its stationary sub-model. There is no obvious explanation and this effect may be investigated more closely in future work.

While it was shown above that the assumption of a stationary variance was too restrictive for the case study and produced an unrealistic model of the true spatial variation of gamma K, the extension to a non-stationary variance model poses additional problems. We used a model in which the standard deviation is a linear combination of covariates. Parameter estimation and kriging require the inverse of the covariance matrix C, which depends on the covariates through matrix H defined in Eq. 2.4. Thus, C may become near-singular or even singular for specific combinations of the standard deviation covariates, which leads to numerical instability. Different approaches may be used to avoid this problem. Inspired by Marchant et al. (2009), Wadoux et al. (2017) propose to reject combinations of parameters suggested by differential evolution if these lead to a reciprocal condition number smaller than a given threshold. This seems to work fine but it affects the search for optimal parameters in parameter space, which might lead to sub-optimal parameter combinations. Alternatively, singularity might be tackled by making use of the generalized inverse (Sen and Srivastava, 2012), since kriging is about solving a set of linear equations which can also be accomplished using a generalized inverse. We have not investigated this and used the method proposed by Wadoux et al. (2017) instead. Another solution might be to model the log-transformed standard deviation as a linear combination of covariates (as in Pintore and Holmes, 2004). This would assure that the standard deviation is positive regardless of the parameter values. We explored this approach by taking $\sigma(s) = e^{\left(\sum_{l=0}^{L} \kappa_l g_l(\mathbf{s})\right)}$ and re-estimating the parameters on the log-scale. We observed that a slight change in a parameter value may lead to a large change in the standard deviation. The estimated standard deviation therefore became very unstable and near-singularity still occurred. Thus, we did not pursue this any further.

In this work, uncertainty about the mean regression coefficients was accounted for in the second term of Eq. 2.12, while uncertainty about the standard deviation coefficients and correlogram parameters was ignored. This may lead to under-estimation of the actual “true” uncertainty. Taking uncertainty about the correlogram parameters and standard deviation regression coefficients into account is possible, although it complicates the analysis. One possible approach is described in Lark (2002) and Marchant and Lark (2007a). The authors use the derivative of the kriging prediction error variance and kriging weights with respect to the variogram parameters to infer a map of the covariance uncertainty, which can then be added to the prediction error variance map. We anticipate that this can be easily extended with respect to the standard deviation regression coefficients. Another technique to account for uncertainty in all parameters is to take a Bayesian approach, such as in Diggle and Ribeiro (2007b, Chapt. 8). We judged uncertainty in the standard deviation coefficients and correlogram parameters to play a minor role, given that the total number of model parameters was much smaller than the number of observations, but taking these additional sources of uncertainty into account would certainly make a valuable extension.

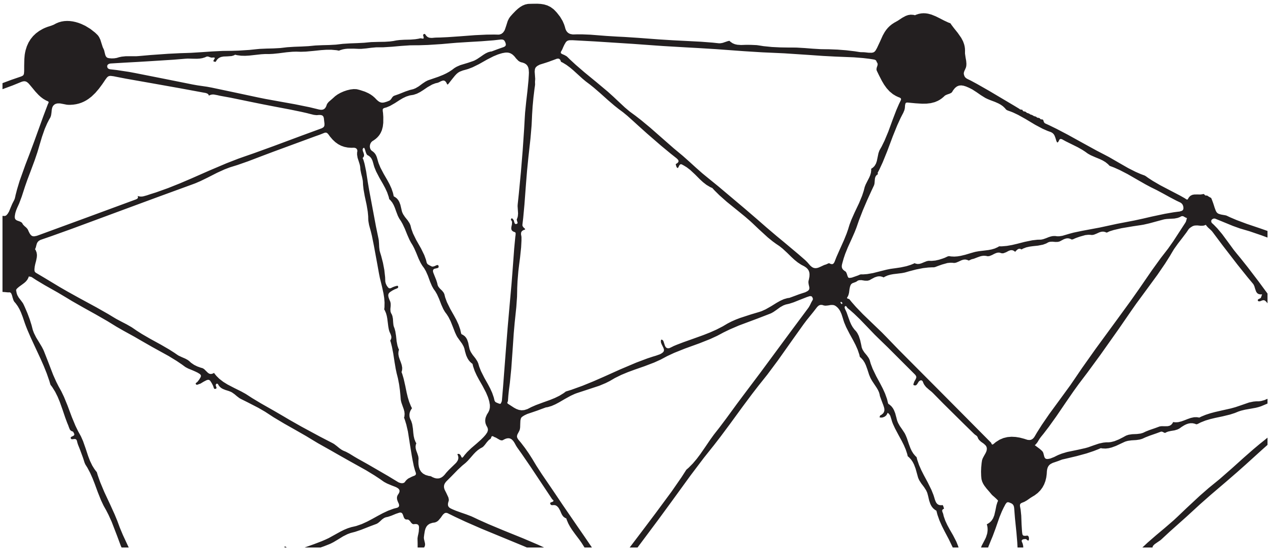
There is also a need to further investigate the data requirements (number of observations and their spatial locations) for adequate fitting of non-stationary variance models. In this work, we fitted a total of seven parameters (including two additional standard deviation parameters) for the non-stationary variance model from 100 observation locations, which we considered adequate. However, the number of standard deviation parameters when using a non-stationary variance model might grow manifold, such as when using a more complex standard deviation function (e.g. splines). It has been demonstrated that covariance uncertainty is minimal when observations are clustered (Marchant and Lark, 2007a), while interpolation error is reduced by spreading the observations in geographic and feature (i.e. covariate) space (Brus and Heuvelink, 2007). How large the sample size should be and which sampling design is best for estimation of the standard deviation coefficients has not been thoroughly explored, although Wadoux et al. (2017) indicates that the non-stationary variance model benefits from spreading the observations in the standard deviation covariate space, while keeping the sampling density fairly constant over the area.

Finally, we emphasize the value of improved uncertainty quantification, as obtained through the use of a non-stationary variance model. Map users often take the prediction map as their first interest, but visualization of the prediction alone can give a wrong impression about the quality of the map (Hengl and Toomanian, 2006) and bias the subsequent decision-making process (Goovaerts, 2001). Uncertainty quantification of the prediction is as important as the prediction itself to obtain a full impression about the quality of the maps. If the uncertainty is too large, users may decide to invest in obtaining a more accurate map (Heuvelink, 2014), but they can only make such decision if they have reliable information about the map quality. We also emphasize the importance of validating the uncertainty estimate, with measures such as θ statistics and accuracy plots. Our study showed that assuming stationarity in the variance can lead to erroneous quantification of uncertainty. In such cases we advocate the use of the non-stationary variance model because it is more flexible and leads to improved estimation of the prediction uncertainty.

2.5 Conclusion

We tested the non-stationary variance model developed in Lark (2009) for spatial interpolation of a soil property. We used multiple covariates to model the spatial standard deviation. Covariates were chosen based on the Akaike information criterion and model parameters fitted using a maximum likelihood approach. We compared the non-stationary variance model to the stationary variance model in a case study. The main conclusions are:

- When modelling a soil property that exhibits local differences in spatial variation, using a non-stationary variance model is recommended over a stationary variance model because it yields a more realistic quantification of prediction uncertainty. Using a constant standard deviation is often not realistic and may lead to local over- or underestimation of the uncertainty.
- Estimation of the parameters of a non-stationary variance model is hampered by near-singularity of the covariance matrix, for which several solutions are proposed but that need further investigation.
- In a case study mapping gamma K in the Hunter Valley, Australia, the kriging standard deviation maps of the stationary and non-stationary variance models were very different. Evaluation using independent validation data showed that the non-stationary model captured the uncertainty much better.
- In the case study different covariates were chosen to model the unconditional mean of the stationary and non-stationary variance models. However, the kriging prediction maps were nearly the same. This suggests that these are insensitive to choices made in the model selection process. Future research may show whether this is a consistent finding or case-dependent.



Chapter 3

Sampling design optimization for rainfall prediction using a non-stationary geostatistical model

The accuracy of spatial predictions of rainfall by merging rain gauge and radar data is partly determined by the sampling design of the rain gauge network. Optimizing the locations of the rain gauges may increase the accuracy of the predictions. Existing spatial sampling design optimization methods are based on minimization of the spatially averaged prediction error variance under the assumption of intrinsic stationarity. Over the past years, substantial progress has been made to deal with non-stationary spatial processes in kriging. Various well-documented geostatistical models relax the assumption of stationarity in the mean, while recent studies show the importance of considering non-stationarity in the variance for environmental processes occurring in complex landscapes. We optimized the sampling locations of rain gauges using an extension of the kriging with external drift (KED) model for prediction of rainfall fields. The model incorporates both non-stationarity in the mean and in the variance, which are modelled as functions of external covariates such as radar imagery, distance to radar station and radar beam blockage. Spatial predictions are made repeatedly over time, each time recalibrating the model. The space-time averaged KED variance was minimized by spatial simulated annealing (SSA). The methodology was tested using a case study predicting daily rainfall in the north of England for a one-year period. Results show that (i) the proposed non-stationary variance model outperforms the stationary variance model, and (ii) a small but significant decrease of the rainfall prediction error variance is obtained with the optimized rain gauge network. In particular, it pays off to place rain gauges at locations where the radar imagery is inaccurate, while keeping the distribution over the study area sufficiently uniform.

Based on:
Wadoux, A. M. J.-C., Brus, D. J., Rico-Ramirez, M. A. and Heuvelink, G. B. M. (2017).
Sampling design optimisation for rainfall prediction using a non-stationary geostatistical model. *Advances in Water Resources*, **107**, pp.126-138.

3.1 Introduction

Accurate information about the space-time distribution of rainfall is essential for hydrological modelling. Rain gauge rainfall measurements are generally accurate and have high temporal resolution, but they typically have a low spatial density, which may cause large errors in interpolated maps given the high spatial variability of rainfall. In contrast, weather radar imagery provide a full spatial coverage of the rainfall field in combination with high temporal resolution. However, radar-derived rainfall predictions experience complex spatio-temporal disturbances and can be inaccurate, especially in mountainous regions.

Over the past years, many statistical techniques have been used to combine the strengths of the two measurement devices, such as Bayesian techniques (Todini, 2001), spatial logistic regression (Fuentes et al., 2008), radar bias correction (Seo and Breidenbach, 2002; Sinclair and Pegram, 2005) and copulas (Vogl et al., 2012). There is also a wide range of geostatistical prediction methods that combine rain gauge measurements with radar imagery, such as kriging with external drift (KED) (Velasco-Forero et al., 2005) and co-kriging (Sideris et al., 2014). Provisions to address non-normality have also been employed, e.g. Box-Cox, square root and normal-score transformation. Besides, various techniques for parameter estimation are available, such as least squares and (restricted) maximum likelihood estimation. Velasco-Forero et al. (2009) and Schiemann et al. (2011) make use of a non-parametric correlogram to derive a rainfall field from radar imagery, dealing with anisotropy and temporal variation of the rainfall structure. Goudenhoofdt and Delobbe (2009) showed that geostatistical merging methods gave the best results for rainfall prediction in the Walloon region in Belgium, although the performance was dependent on the network configuration. For a more detailed review of radar-gauges merging techniques, we refer to Goudenhoofdt and Delobbe (2009), Nanding et al. (2015) and Jewell and Gaussiat (2015).

Few studies focus on the sampling design of the rain gauge network. For example, Pardo-Igúzquiza (1998) derives the optimal network design by minimizing an objective function based on prediction accuracy combined with monetary costs, Barca et al. (2008) explore the optimal location of new monitoring stations by minimizing the mean shortest distances. Spatial optimization of the gauge network in radar-gauge merging studies remains largely unexplored. In sampling design for spatial prediction of rainfall by ordinary kriging (OK), using the average OK variance as a minimization criterion leads to spreading of the locations in geographic space. However, for mapping with the help of covariates as in KED, we also need to spread the locations in feature (i.e. covariate) space. By selecting locations such that the covariate space is fully covered, uncertainty about the regression coefficients is minimized. Brus and Heuvelink (2007) showed that minimizing the spatially averaged KED variance achieves a proper balance between optimization in geographic and feature space. Heuvelink et al. (2012) extended this to a space-time kriging case and minimized the space-time averaged KED variance to optimize static as well as dynamic sampling designs.

In this study we only consider static designs, i.e. we assume that the rain gauge locations do not change over time. This is because it is impractical to move rain gauges in an operational context. Our objective is to optimize the static rain gauge sampling design such that it minimizes the space-time averaged prediction error variance. We use a geostatistical model in which both the mean and the standard deviation are assumed to be a linear combination of covariates. The model parameters (regression coefficients and correlogram parameters) are estimated from the rain gauge data using Restricted Maximum Likelihood. We optimize the rain gauge locations with spatial simulated annealing (SSA). The model is tested in a case study in the north of England for daily rainfall mapping in the year 2010.

3.2 Materials and methods

3.2.1 Case study 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 from the Environment Agency (EA). The data originally provided by the EA are at 15-min resolution and were aggregated to daily sums. We checked the quality of the rain gauge data and reduced the number of gauges to 185, by excluding gauges with anomalies, such as an excessive number of missing values. The locations of the remaining 185 gauges are shown in Fig. 3.1.

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 Fig. 3.1. 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

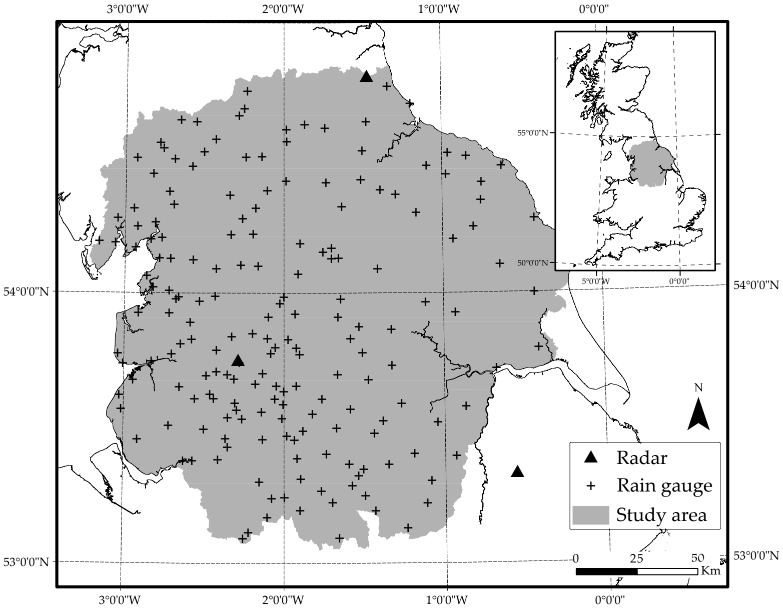


Figure 3.1 *– Map of the study area with locations of rain gauges and radar stations.*

attenuation correction, vertical reflectivity profile correction and rain gauge adjustment (Harrison et al., 2009). The radar rainfall product is available with a spatial and temporal resolution of 1 km and 5 min, respectively (Met Office, 2003). The radar data set contains several missing 5 min periods and therefore a nowcasting model was used to interpolate missing periods for a maximum of 3 h. Next the 5-min resolution images were aggregated to daily sums.

Besides these two rainfall datasets, the following covariate maps were used:

- Digital elevation model (DEM)(Fig. 3.3a) at 50 m resolution from the SRTM (shuttle radar topography mission), see Farr et al. (2007). The elevation ranges from 6 m to 926 m above sea level (a.s.l.) with an average of 159 m a.s.l.
- Radar beam blockage map at 1 km resolution (Fig. 3.3b). The radar beam blockage maps were generated for each radar station using the DEM and a ground clutter model described in Rico-Ramirez et al. (2009). The individual beam blockage maps were combined to produce a single map with 1 km resolution for the 0.5 degree radar scan inclination. When merging overlapping areas, priority was given to the lowest beam blockage value. The blockage maps represent the degree of deviation from the 0.5 degree radar inclination due to topographic obstacles. Values are expressed in percentages from 0 to 100. Their mean is 4.8%.
- Distance from nearest radar stations map at 1 km resolution (Fig. 3.3c). Values

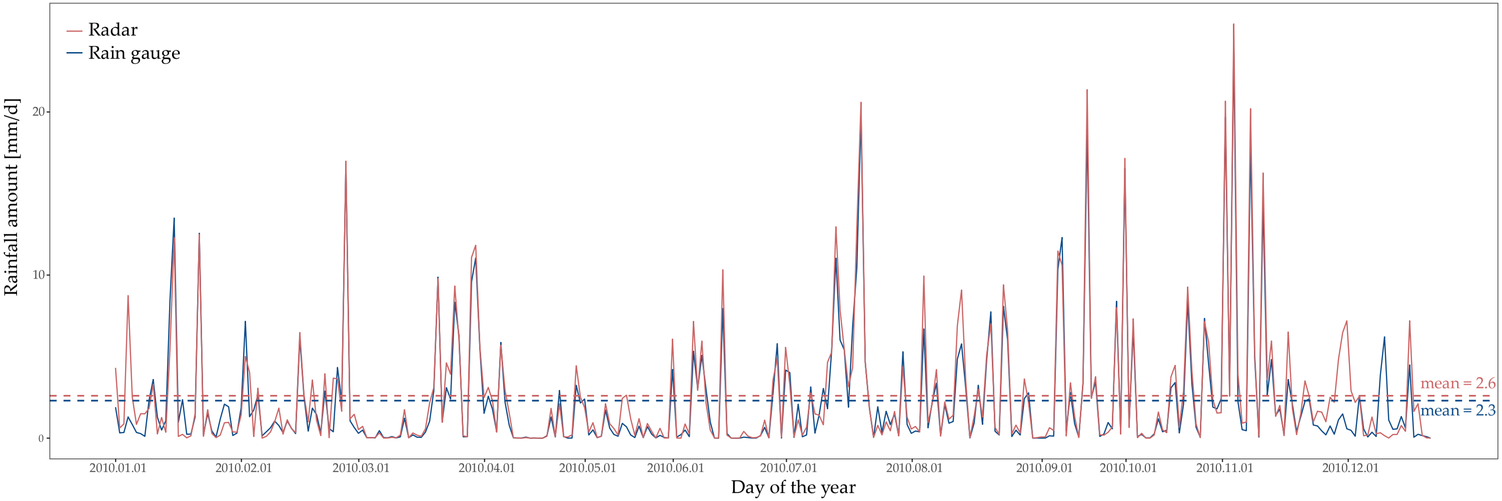
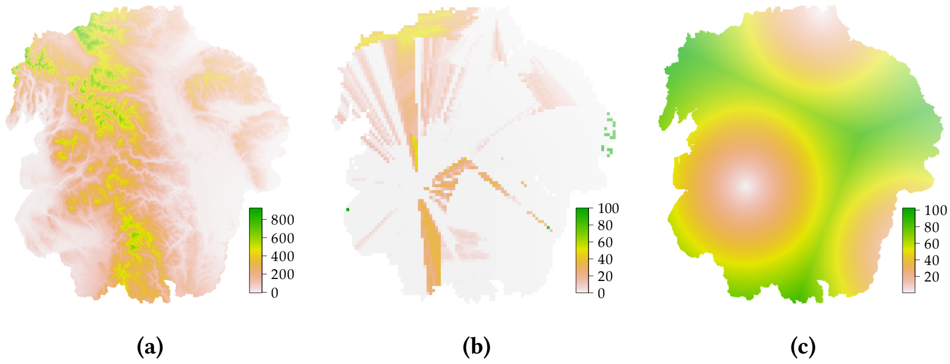


***Figure 3.2** – Time series of spatially averaged daily rainfall derived from rain gauges and radar imagery for the year 2010.*



(a) **(b)** **(c)**

Figure 3.3 – *Covariates used in model calibration: (a) Terrain elevation (m a.s.l.), (b) radar beam blockage (%), (c) Distance from nearest radar station (km).*

are expressed in km and vary from 0 (radar station location) to 102.6 km. The mean is 51.3 km.

3.2.2 Model definition

Daily rainfall as measured by rain gauges $Z_t(s)$ at any location s in the study area $\mathcal{A}$ and time (day) $t \in T$ is modelled by:

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

where $m_t = \{m_t(s) | s \in \mathcal{A}\}$ is a spatial trend, σ_t the spatial standard deviation and ε_t a zero-mean, unit variance, normally distributed, second-order stationary and spatially correlated residual at time t. Note that ε_t may be correlated in space, whereas we assume that ε_t and $\varepsilon_{t'}$ are uncorrelated if $t \neq t'$. Both the trend and the standard deviation are modelled as linear combinations of covariates:

$$m_t(s) = \sum_{k=0}^{K} \beta_{tk} f_{tk}(s) \tag{3.2}$$

$$\sigma_t(s) = \sum_{l=0}^{L} \kappa_{tl} g_{tl}(s) \tag{3.3}$$

where the β_{tk} and κ_{tl} are regression coefficients and the f_{tk} and g_{tl} are covariates. We assume that $f_{t0}(s) = g_{t0}(s) = 1$ for all t and s, so that β_{t0} is an intercept and κ_{t0} is a space-invariant constant contribution to the standard deviation. Note that the covariates may vary in space and time. Note also that the space-time model effectively consists of a set of separate spatial models, one for each day of the year. Temporal correlation is not modelled in this case study.

We consider the situation that Z_t has been measured at n locations s_i ($i = 1, \ldots, n; s_i \in \mathcal{A}$). The measurements $z_t(s_i)$ are treated as realizations of $Z_t(s_i)$ and prediction is done for Z_t at new, unobserved locations s_0. Stacking the $z_t(s_i)$ in a (column) vector $\mathbf{z}_t$ and changing to matrix notation yields:

$$\mathbf{z}_t = \mathbf{F}_t \boldsymbol{\beta}_t + \mathbf{H}_t \boldsymbol{\varepsilon}_t, \tag{3.4}$$

where $\mathbf{F}_t$ is the $n \times (K+1)$ matrix of spatial trend covariates at the observation locations, $\boldsymbol{\beta}_t$ is the $(K+1)$ vector of trend coefficients, $\boldsymbol{\varepsilon}_t$ is the n-vector of standardised residuals with correlation matrix $\mathbf{R}_t$ and $\mathbf{H}_t$ is an $n \times n$ diagonal matrix defined by:

$$\mathbf{H}_t = \text{diag}\{\mathbf{G}_t \cdot \boldsymbol{\kappa}_t\}, \tag{3.5}$$

where $\mathbf{G}_t$ is the $n \times (L+1)$ matrix of standard deviation covariates at the observation locations and $\boldsymbol{\kappa}_t$ is an $(L+1)$ vector of standard deviation regression coefficients. Note that while $\boldsymbol{\varepsilon}_t$ has correlation matrix $\mathbf{R}_t$, the stochastic component $\mathbf{H}_t\boldsymbol{\varepsilon}_t$ of Eq. 3.4 has variance-covariance $\mathbf{C}_t = \mathbf{H}_t\mathbf{R}_t\mathbf{H}_t'$. The parameters of the model defined by Eq. 3.4 are the $\boldsymbol{\beta}_t$, $\boldsymbol{\kappa}_t$ and the parameters of a model for the spatial autocorrelation of the standardized residuals. We assume an isotropic exponential correlogram $r_t(h) = r_{t0}\{\exp(-\frac{h}{a_t})\}$ (where $h > 0$ is the Euclidean distance between two points, by definition $r_t(0) = 1$), thus introducing two more parameters, namely the micro-scale correlation r_{t0} and the spatial correlation length parameter a_t. Note that parameter r_{t0} equals one minus the nugget-to-sill ratio. For notational convenience from here on we drop the subscript t.

3.2.3 Parameter estimation

For each day, two subsets of model parameters must be estimated, the spatial trend regression coefficients $\boldsymbol{\beta}$ and all parameters of the stochastic part of the model, $\Phi = [\boldsymbol{\kappa}, r_0, a]$. Given Φ the estimation of $\boldsymbol{\beta}$ is straightforward and can be done analytically by generalized least squares (GLS) (Lark and Webster, 2006). However, estimation of Φ is more difficult. We used a restricted (or residual) maximum likelihood (REML) approach for this. Similar to maximum likelihood, REML aims to find the vector of parameters Φ for which the observed data yield the highest probability density (i.e. likelihood). In our case the model contains a spatial trend (fixed effect), and so the likelihood must be computed from the probability distribution of the model residuals, which can be computed from the observations if the spatial trend is known. This implies that the likelihood depends on the regression coefficients, which are unknown and also must be estimated. The solution to this problem, proposed by Patterson and Thompson (1971) is to detrend the data by multiplying the data by a projection matrix (see also Lark and Cullis, 2004). After detrending the data and estimating Φ by minimizing the negative restricted log-likelihood given in Eq. 3.6 below, the estimate of $\boldsymbol{\beta}$ is obtained by substituting the REML estimates of Φ in the GLS equations. The negative restricted log-likelihood function is given by (Webster and Oliver, 2007):

$$\ell(\Phi|\mathbf{z}) = \text{constant} + \frac{1}{2}\log|\mathbf{C}| + \frac{1}{2}\log|\mathbf{F}'\mathbf{C}^{-1}\mathbf{F}| + \frac{1}{2}\mathbf{z}'\mathbf{P}'\mathbf{C}^{-1}(\mathbf{I}-\mathbf{Q})\mathbf{z} \tag{3.6}$$

where $\mathbf{I}$ is an identity matrix and $\mathbf{Q}$ is defined as:

$$\mathbf{Q} = \mathbf{F}(\mathbf{F}'\mathbf{C}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{C}^{-1} \tag{3.7}$$

and:

$$\mathbf{P} = \mathbf{I} - \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}' \tag{3.8}$$

Next matrix $\mathbf{C}$ is obtained by substituting the optimized Φ and used to estimate $\boldsymbol{\beta}$ using GLS (Marchant et al., 2009):

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}'\mathbf{C}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{C}^{-1}\mathbf{z} \tag{3.9}$$

3.2.4 Kriging

In KED, predictions at new locations are made by:

$$\hat{Z}(s_0) = \mathbf{f}(s_0)'\hat{\boldsymbol{\beta}} + \mathbf{g}(s_0)'\hat{\boldsymbol{\kappa}}\hat{\boldsymbol{\varepsilon}}(s_0) \tag{3.10}$$

where $\hat{\boldsymbol{\varepsilon}}(s_0)$ is the kriged standardized residual. Ignoring estimation errors in $\hat{\boldsymbol{\kappa}}$ allows us to use a standard result from universal kriging (Webster and Oliver, 2007) which yields:

$$\hat{Z}(s_0) = (\boldsymbol{c}_0 + \mathbf{F}(\mathbf{F}'\mathbf{C}^{-1}\mathbf{F})^{-1}(\boldsymbol{f}_0 - \mathbf{F}'\mathbf{C}^{-1}\boldsymbol{c}_0))\mathbf{C}^{-1}\mathbf{z} \tag{3.11}$$

where $\boldsymbol{f}_0$ is a $(K+1)$ vector of trend covariates at the prediction location and $\boldsymbol{c}_0$ is an n vector of covariances between the residuals at the observation and prediction locations. Note that these are covariances of the (unstandardized) residuals $\sigma_t \cdot \varepsilon_t$ and thus depend on the standard deviation covariates g_l, their associated (estimated) regression coefficients κ_l and the correlogram of ε. Recall that since we use separate, independent models for each day, only rainfall observations of that day are used to predict $Z(s_0)$.

The variance of the prediction error is given by (Cressie, 2015):

$$\begin{aligned}\operatorname{var}(Z(s_0) - \hat{Z}(s_0)) =& \sigma^2(s_0) = \boldsymbol{c}(0) - \boldsymbol{c}_0'\mathrm{C}^{-1}\boldsymbol{c}_0 \\ &+ (\boldsymbol{f}_0 - \mathrm{F}'\mathrm{C}^{-1}\boldsymbol{c}_0)'(\mathrm{F}'\mathrm{C}^{-1}\mathrm{F})^{-1}(\boldsymbol{f}_0 - \mathrm{F}'\mathrm{C}^{-1}\boldsymbol{c}_0)\end{aligned} \tag{3.12}$$

where $\sigma^2(s_0)$ is the variance of $Z(s_0)$. The first two terms on the right-hand side of Eq. 3.12 quantify the prediction error variance of the residuals, while the last term is the estimated spatial trend error variance. Note that here we take uncertainty about the β_k into account, whereas uncertainty about the κ_l and correlogram parameters r_0 and a is ignored. Taking the latter uncertainties into account is not an easy task and beyond the scope of this work.

3.2.5 Optimizing the rain gauge locations

We suppose that, due to budget constraints, the number of rain gauges n is fixed. The aim is to find the optimal locations of the rain gauges for predicting daily rainfall for a given time period. In order to do this, a criterion is needed that defines the performance of a given sampling design and that allows to compare designs. It makes sense to use the spatially averaged kriging variance as a criterion, because this provides an appropriate summary measure of the prediction accuracy (Brus and Heuvelink, 2007). In our case, where a static rain gauge network must be optimized for a longer period of time, in addition we should also average the criterion over time. This results in the following minimization criterion:

$$u = \frac{1}{T}\frac{1}{|\mathcal{A}|}\sum_{t=1}^{T}\int_{s\in\mathcal{A}} \operatorname{var}\left(Z_t(s) - \hat{Z}_t(s)\right) \mathrm{d}s \tag{3.13}$$

A closer look at the kriging variance Eq. 3.12 and hence the criterion Eq. 3.13 shows that it only depends on the sampling locations s_i, the correlogram r and the spatial trend and standard deviation covariates. This implies that the configuration can be optimized before the observations are taken, provided that the model and covariance structure are known. Recall that in this study we only consider static designs, i.e. we assume that the rain gauge locations do not change over time. In theory, with a finite number of possible rain gauge locations N derived from discretizing the study area $\mathcal{A}$, we could try all $\binom{N}{n}$ combinations, and choose the one that minimizes 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 discretization of the study area. A solution to this problem is to use a spatial numerical search algorithm. We used spatial simulated annealing (SSA), as proposed in Van Groenigen and Stein (1998).

Spatial simulated annealing is an iterative optimization algorithm in which a sequence of new possible sampling locations is generated. A new sampling location is derived by selecting randomly one sampling location and shifting it in a random direction over a random distance. Each time a new possible location is generated, the criterion (Eq. 3.13) is calculated for the new candidate design and compared with the criterion value of the current design. The new location is always accepted if the criterion becomes smaller. If the criterion becomes larger the new location is sometimes accepted, namely with probability:

$$P(\text{accept}) = e^{\frac{u(\text{old})-u(\text{new})}{temp}} \tag{3.14}$$

where *temp* is a control parameter accounting for the number of remaining iterations, called the temperature. It decreases from a positive starting value to zero as the number of iterations increases. Eq. 3.14 shows that, given *temp*, the larger the increase of the criterion, the smaller the probability of accepting a worse design. Also, the smaller *temp*, i.e. the larger the number of iterations already done, the smaller the acceptance probability of a worse sample. The *temp* parameter is kept constant during a set of m iterations, called a chain, after which it is decreased to a value $\alpha * temp$, with $\alpha < 1$. This process repeats itself until the total number of planned iterations have been completed. Parameter α should be chosen such that the acceptance probability is close to one in the first chain and approximating zero during the final stage of iterations. At first worsening designs are accepted to be able to escape from local minima, but towards the end only designs that improve the criterion are accepted. We refer to Heuvelink et al. (2010) for a more detailed explanation of the numerical optimization algorithm used in this study.

3.2.6 Application to the case study

Parameter estimation - We chose one covariate for the spatial trend (radar image) and three for the standard deviation (DEM, distance from nearest radar station and radar beam blockage). These were chosen after consulting experts on rainfall mean and radar uncertainty. The standard deviation covariates were multiplied by the radar image of each day to obtain a standard deviation that is proportional to the amount of rainfall and to avoid having a positive standard deviation where no rainfall is detected by the radar. All covariates were projected to the British National Grid system and resampled to a spatial resolution of 500 m × 500 m. The spatial trend covariate was standardized for each individual day, using time-specific means and variances. The standard deviation covariates were not standardized to avoid negative values. Negative values might lead to singularity of the covariance matrix, as explained below. Because of extreme values in the radar imagery (likely anomalies) the upper 0.1% of the radar image values were bounded to the 99.9% quantile and inspected visually to ensure that inconsistent values were detected and corrected. Rainfall measurements were not transformed prior to modelling. Averaging to daily values and including trend covariates removed much of the skewness, as confirmed by a post-hoc analysis of the residuals.

The global optimum for the log-likelihood function was obtained using differential evolution (Storn and Price, 1997) as implemented in the R package DEoptim (Ardia et al., 2015). We fixed the convergence threshold at 10^{-10}. Calculations were done using parallel computing on an eight cores computer and estimation of the parameters took approximately 15 hours for the whole year.

The correlogram and standard deviation parameters were bounded prior to estimation. The corresponding upper and lower limits are given in Table 3.1. The limits were chosen based on physical reasoning and theoretical restrictions, e.g. the correlation length parameter a was not allowed to be greater than one-third of the extent of the study area and the micro-scale correlation parameter r_0 was forced between 0 and 1. The intercept for the standard deviation κ_0 was bounded with a lower bound set to a small positive value to avoid singularity problems that would occur if the standard deviation were too close to zero. For the same reason all other standard deviation coefficients were restricted to non-negative values. In addition, whenever a proposal combination of model parameters generated by a DEoptim iteration produced a near-singular C matrix, such combination was rejected. We will discuss the singularity issue more extensively in the Discussion. The calibration was performed for 315 days of the year 2010. Days with no rainfall or excessive missing data were excluded.

***Table 3.1** – Model parameters with lower and upper estimation bounds.*

Parameter	Lower bound	Upper bound	Associated to
r_0 (-)	0	1	Micro-scale correlation
a (km)	0.1	50	Correlogram length parameter
β_0 (mm)	-	-	Intercept for the mean
β_1 (-)	-	-	Radar image
κ_0 (-)	0.0001	50	Intercept for the standard deviation
κ_1 (m^{-1})	0	0.01	Elevation model × Radar image
κ_2 (km^{-1})	0	0.1	Distance from nearest radar station × Radar image
κ_3 (%$^{-1}$)	0	0.1	Radar beam blockage × Radar image

Kriging prediction - Predictions were made with global kriging using all observations. Since no standard implementation is available for non-stationary variance kriging, we developed our own code. We speeded up the algorithm by inverting matrices using Cholesky decomposition and by using parallel computing.

Simulated annealing - For SSA we used the R package spsann (Samuel-Rosa, 2017). The maximum distance that points could move was set to half the extent of the study area, the actual distance was drawn from a uniform distribution between zero and the maximum distance. The maximum distance in which the rain gauges can be moved becomes smaller as the number of iterations increases and converges to zero at the end of the process. The initial temperature was set to 0.1 with a cooling parameter α of 0.8. The maximum number of chains was set to 140 whereas the number of iterations within a chain was set to the number of observations, so that the total number of iterations is $185 \times 140 = 25{,}900$. The process stops if no improvement is made after 100 chains or when the maximum number of chains is reached. The prediction error variance was evaluated on a coarse grid (3 km × 3 km) to avoid excessive computing time. We used a Linux server 4.4.0-38-generic Ubuntu SMP with 48 cores, the total processing time for the SSA was approximately 580 hours.

3.3 Results

3.3.1 Parameter estimation

Figure 3.4 presents box plots of the estimated parameters $\boldsymbol{\beta}$, $\boldsymbol{\kappa}$, r_0 and a. Recall that the trend covariate radar image was standardized in order to be able to compare its estimated regression coefficients with the intercept coefficients. The trend coefficients associated with the radar-rainfall map are nearly always positive ($\beta_1 > 0$ for 92% of the days), indicating a positive effect of radar rainfall. This is as expected, since radar rainfall and rain gauge rainfall are positively correlated (their Pearson correlation coefficient is about 0.96). Note also that the distributions of the trend coefficient estimates are positively skewed. This can be explained from the skew distribution of the rainfall (see Fig. 3.2). Since the radar image covariate was standardized the trend coefficient estimates are likely to be large during days with high rainfall, in particular the trend intercept. This is confirmed by the scatter plots shown in Fig. 3.11).

The estimates of the regression coefficients associated with the standard deviation covariates are always greater or equal to zero because zero was taken as a lower

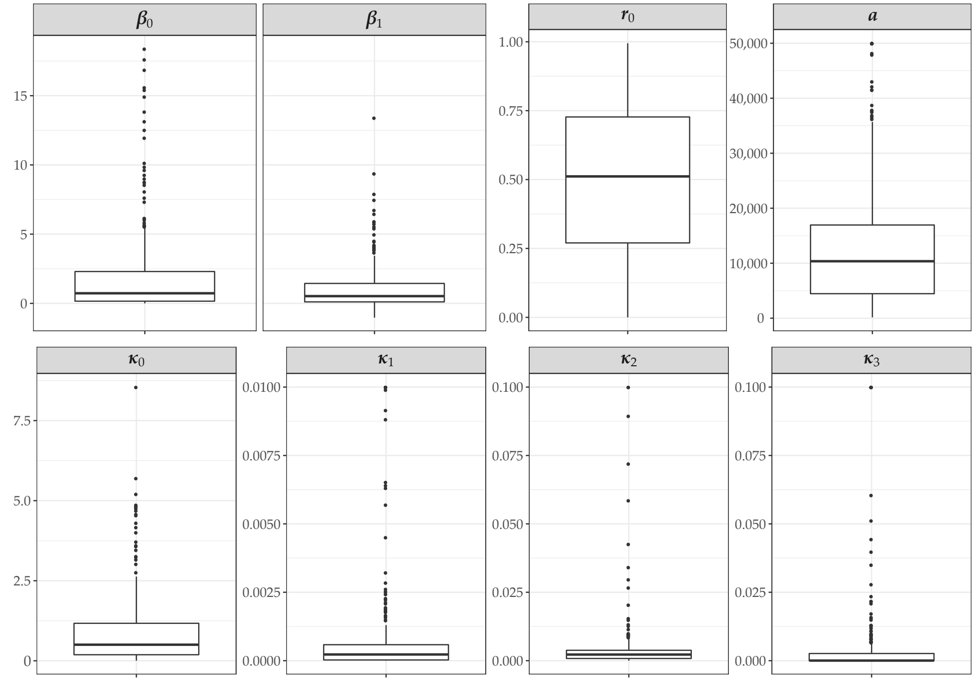


***Figure 3.4** – Box plots of estimated parameters for the mean ($\boldsymbol{\beta}$), standard deviation ($\boldsymbol{\kappa}$), micro-scale correlation (r_0) and correlation length (a). See Table 3.1 for associated covariates. Summary statistics are provided in Table 3.3 and daily values in Fig. 3.8 for $\boldsymbol{\beta}$, Fig. 3.9 for $\boldsymbol{\kappa}$ and in Fig. 3.10 for r_0 and a.*

bound (except for κ_0, which has a lower bound of 0.0001). It appears that the lower bounds for κ_1, κ_2 and κ_3 are a real restriction because the estimates are often almost equal to their lower bounds. For a few days estimates of κ_1, κ_2 and κ_3 are pushed to their upper bounds. This occurs mainly when the rainfall amount is close to zero (Fig. 3.11), which is not surprising because these are days where the standard deviation covariates are small. Note also that the contribution of each covariate to the standard deviation cannot be inferred by direct comparison of the coefficient estimates because the standard deviation covariates were not standardized.

The boxplots of the correlogram parameters in Fig. 3.4 show that for most days there is significant spatial correlation in the residuals. The micro-scale correlation parameter is symmetrically distributed around 0.50 whereas the correlation length parameter has a skew distribution with a median of about 10 km. For the exponential model this indicates a correlation up to about 30 km, which is not very large given the extent of the study area. Table 3.3 provides summary statistics of all parameter estimates.

3.3.2 Kriging

Figure 3.5 shows an example of three successive days with radar image, spatial trend (Eq. 3.2), prediction (Eq. 3.11), standard deviation of residuals (Eq. 3.3) and prediction error standard deviation (kriging standard deviation in Eq. 3.12), as obtained using the initial rain gauge network design. For all three days the spatial pattern of the predicted rainfall is very similar to that of radar rainfall, illustrating the strong effect of radar rainfall on the final prediction. This is confirmed by the high β_1 estimates for February 15 and February 16. The β_1 estimate for February 14 is much smaller, but this is because the average rainfall was low on that day (recall that the radar map was standardized while the rainfall data were not). Note that February 15 and 16 show an underestimation of the actual rainfall accounted for by β_0 (Table 3.2).

The spatial pattern of the standard deviation maps is correlated with the radar rainfall map for February 14 and February 15. For February 16, from the two rainfall events of the predicted map (south-west and north-west), only one appears in the standard deviation map. This can be explained from the Elevation and Distance from nearest radar station covariate maps (Fig. 2.2), which have low values in the south-west and high values in the north-west. The effect is even stronger in the kriging standard deviation map, because the rain gauge density is higher in the south-west. The maps show also that the rainfall pattern may change dramatically over the course of one day. This confirms that the temporal correlation at daily scale is not be very strong.

3.3.3 Optimization

Figure 3.6 shows the decrease of the prediction error variance as the sampling design is perturbed during SSA. The graph shows that several worsening designs are accepted at the beginning. After this initial phase the prediction error variance steadily decreases. After about 10,000 iterations, no substantial further reduction is achieved, suggesting that the algorithm reached a nearly optimum design, as was confirmed by running the algorithm again and obtaining a similar pattern (results not shown). Note that a marked decrease is observed at the very end of the process. We explain the cause of this in the Discussion. Overall the criterion drops from 4.41 to 4.15, which represents an improvement of about 5.8%. Figure 3.7 shows the initial and optimized sampling locations of the rain gauges with the associated spatial sampling density. The optimized design has a fairly uniform distribution of rain gauges with a higher density in the north-west and a lower density in a large band from north-est to south-est and in the south-west. The optimized sampling network

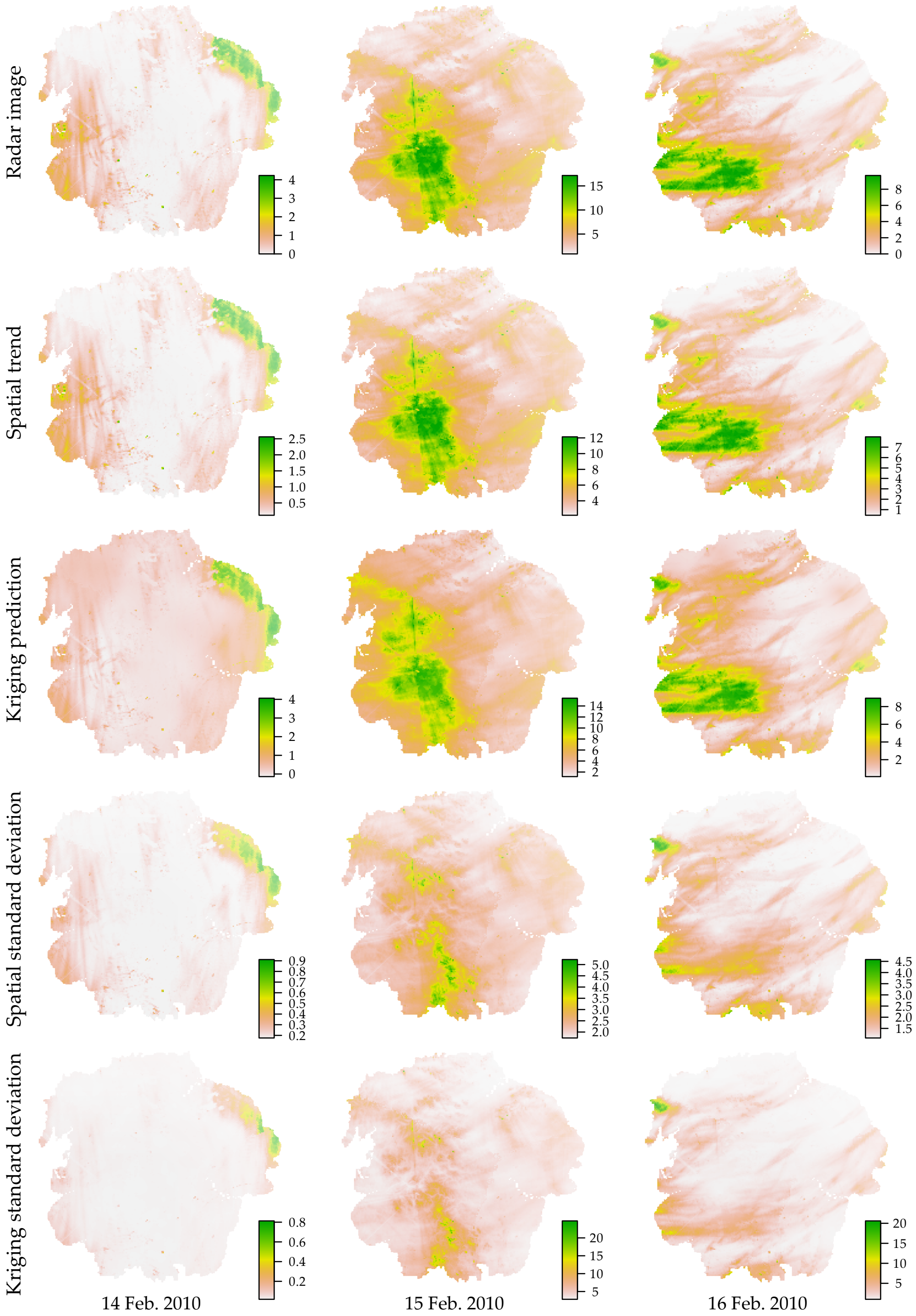


Figure 3.5 *– Radar image and maps of the trend, standard deviation of residuals (sd), kriging prediction and kriging standard deviation for three selected dates (14, 15 and 16 Feb. 2010).*

***Table 3.2** – Model parameter estimates for three example days.*

Day	r_0	a	β_0	β_1	κ_0	κ_1	κ_3	κ_4
February 14th	0.438	19,078	0.336	0.405	0.175	4.611^{-06}	1.831^{-06}	1.712^{-05}
February 15th	0.718	31,813	4.918	1.912	1.712	0.0002	1.704^{-06}	6.137^{-08}
February 16th	0.731	15,597	1.995	1.647	1.072	2.146^{-16}	4.627^{-06}	8.953^{-18}

***Table 3.3** – Summary statistics of estimated model parameters.*

	r_0	a	β_0	β_1	κ_0	κ_1	κ_3	κ_4
Mean	0.495	12,893	1.989	1.120	0.915	0.001	4.39^{-06}	0.005
Median	0.511	10,359	0.729	0.522	0.503	0.0002	2.23^{-06}	3.15^{-05}
SD	0.294	11,668	3.149	1.630	1.163	0.002	1.126^{-05}	0.016
Lower quartile	0.270	4,458	0.157	0.107	0.190	2.601^{-05}	8.014^{-07}	3.580^{-09}
Upper quartile	0.727	16,948	2.299	1.439	1.171	5.834^{-04}	3.793^{-06}	2.615^{-03}
Skewness	−0.151	1.534	2.813	3.034	2.558	3.965	6.577	5.028
Minimum	4.215^{-16}	154.5	0.001	−1.03	0.004	2.055^{-20}	5.847^{-23}	1.637^{-18}
Maximum	0.995	50,000	18.39	13.4	8.547	0.01	0.1	0.1

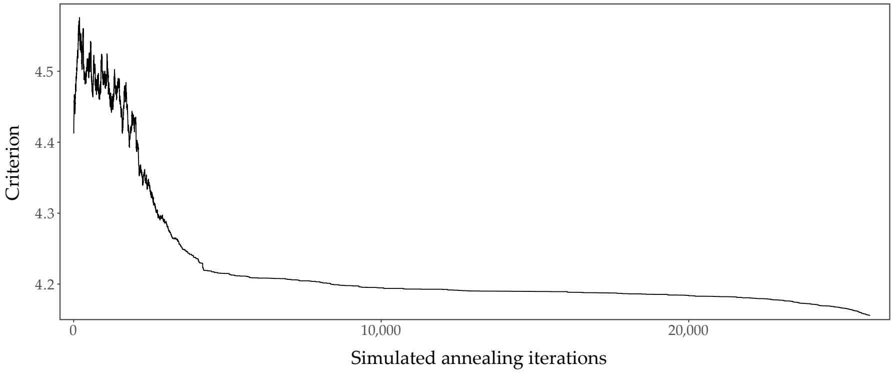


Figure 3.6 *– Trace of the minimization criterion, Eq. 3.13, during SSA (for a case of 25,900 iterations).*

also puts rain gauges towards the boundary of the study area. This is a well-known effect reported in Brus and Heuvelink (2007) and Van Groenigen et al. (1999).

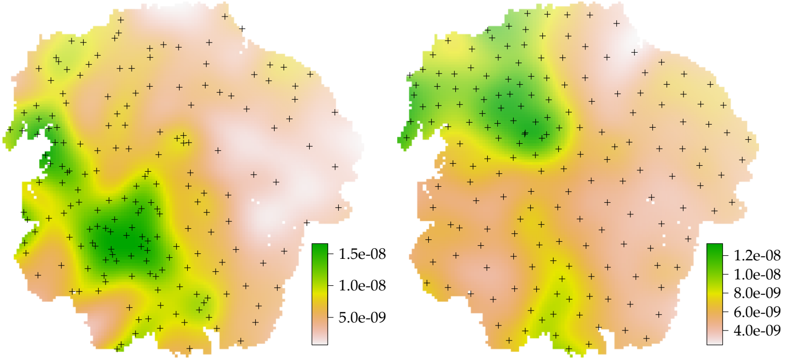


Figure 3.7 *– Initial (left) and optimized (right) rain gauge network with associated density of rain gauges. Density is calculated using a Gaussian kernel as defined in Baddeley and Turner (2005), using a standard deviation of 10 km. Values are expressed in rain gauge per grid cell (500 m × 500 m).*

3.4 Discussion

For the three example dates, the trend and kriging prediction maps have a very similar pattern to that of radar rainfall. The trend is taken as a linear function of the radar image. The trend map and the kriging prediction map are nearly the same. This shows that the kriging step does not add much, which is because the residual variance is small and the residual spatial correlation is often weak. Apparently, the

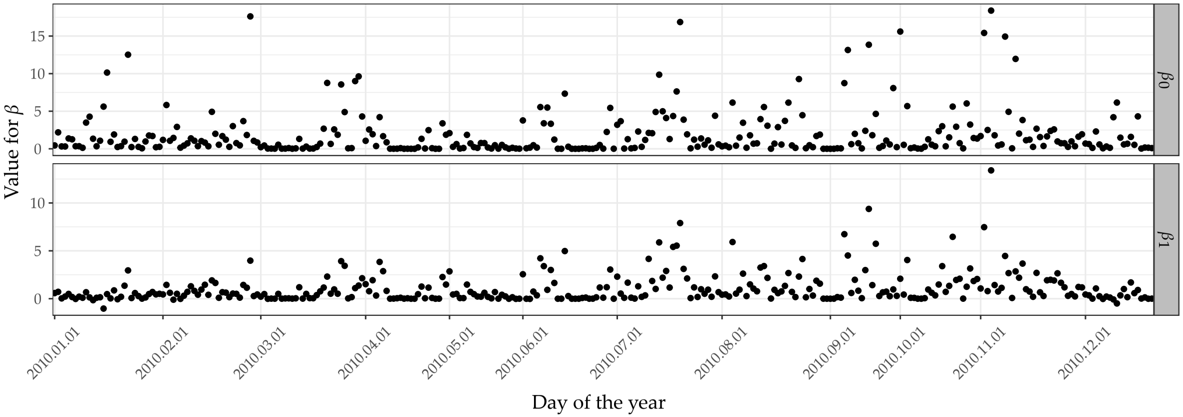


Figure 3.8 *– Values of spatial trend parameters.*

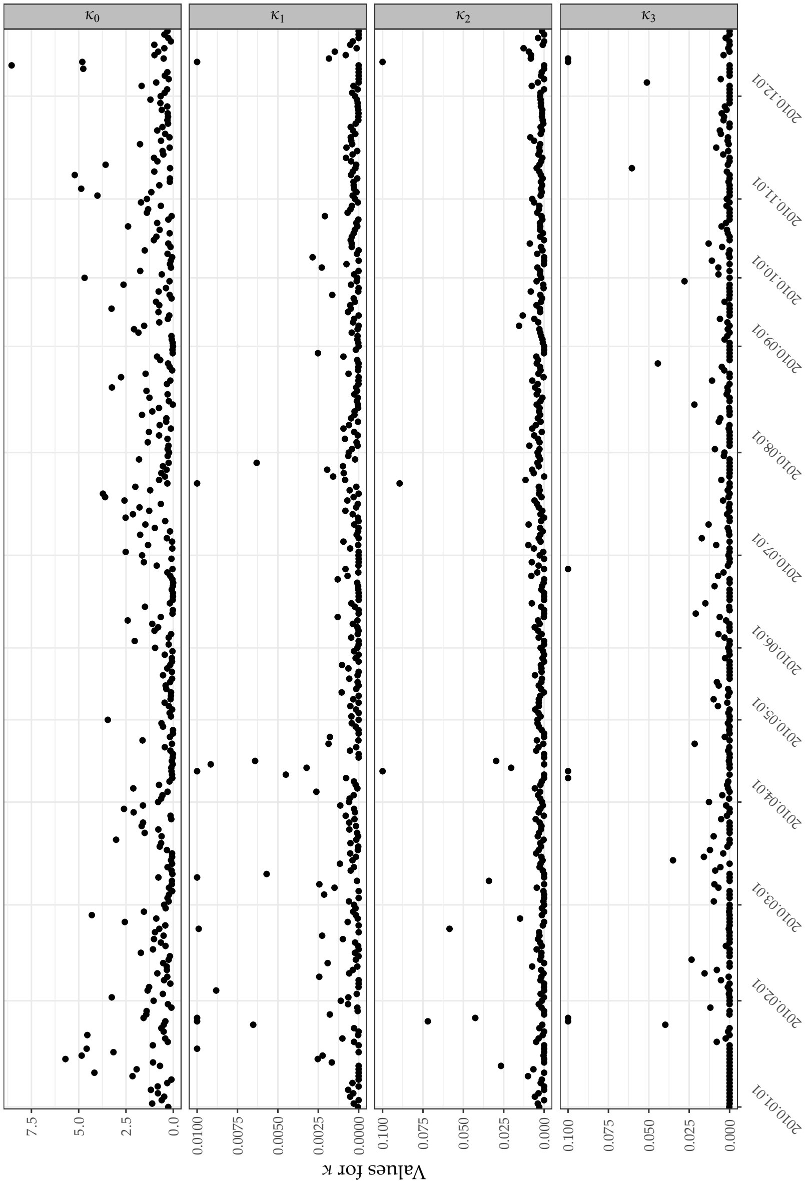


***Figure 3.9** – Values of spatial standard deviation parameters.*

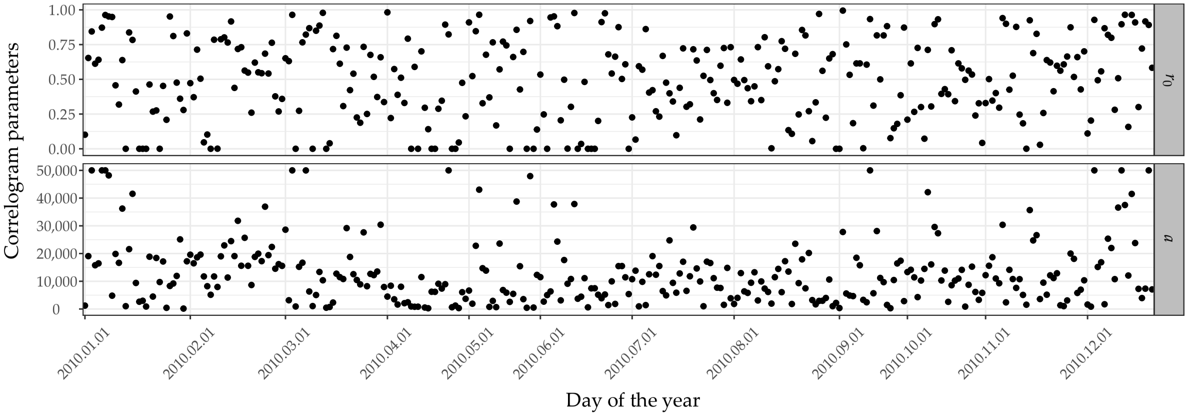


***Figure 3.10** – Values of correlogram parameters.*

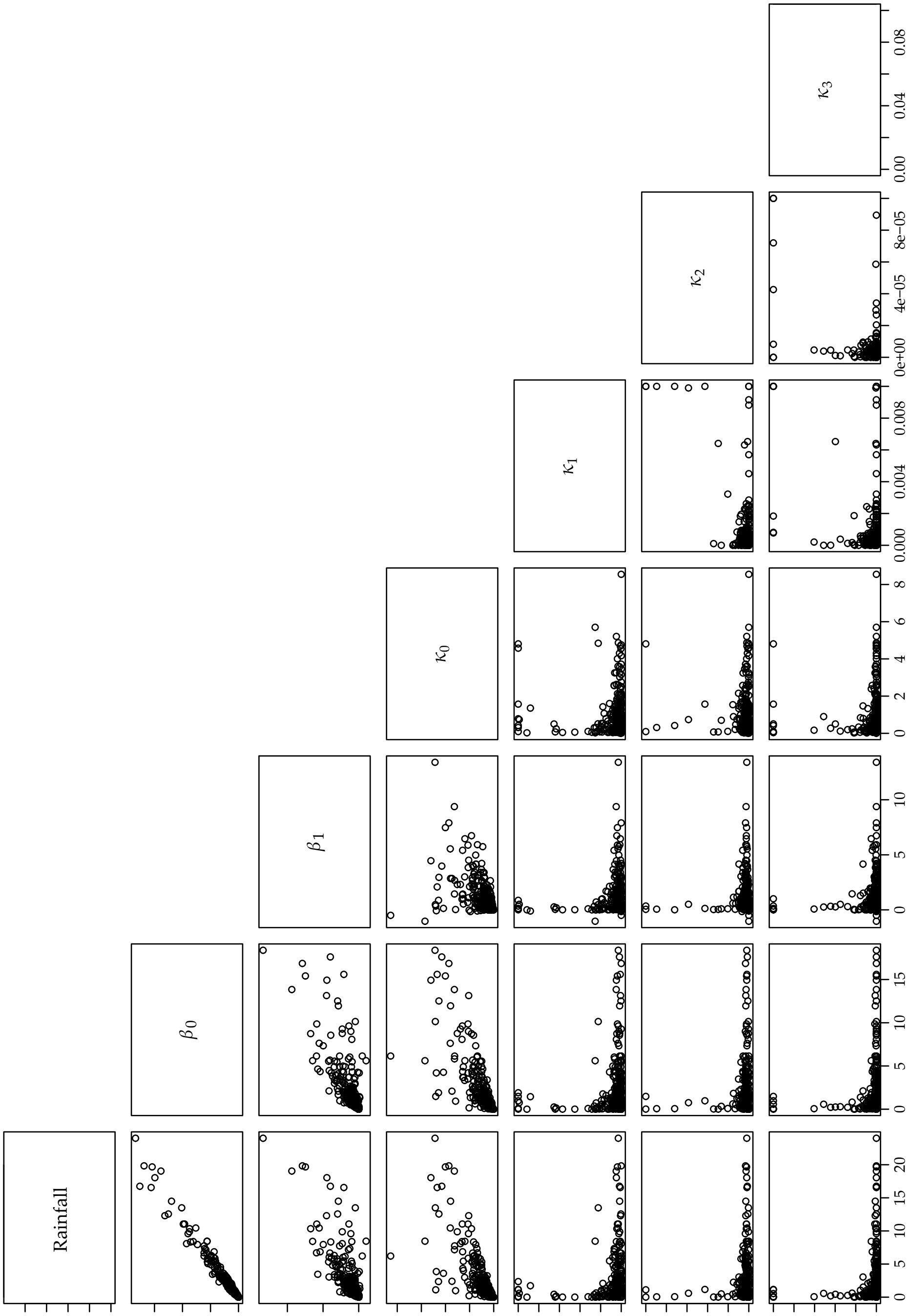


***Figure 3.11** – Cross-correlation matrix between parameters and daily averaged rainfall from rain gauges.*

***Table 3.4** – Pearson correlation coefficients between rain gauge density and standard deviation covariates.*

	Elevation	Distance	Beam blockage
Density initial network	0.31	−0.38	0.09
Density optimized network	0.71	0.29	0.36

radar signal is an important covariate and explains a large part of the rainfall spatial variation. This is not a surprising result that has been reported in many previous studies (e.g. Verworn and Haberlandt, 2011). The importance of radar rainfall is also confirmed by the trend regression coefficients (Fig. 3.4), which are large for the radar covariate. Temporal correlation in daily rainfall is weak and ignored in this example study, but might become more important in case of modelling at a finer time scale, such as required in urban hydrology applications (Muthusamy et al., 2017). Increase of temporal correlation would imply that rainfall at a previous time step becomes a significant covariate. In such case, a more elegant approach might be to replace spatial kriging as employed here by space-time kriging (Heuvelink et al., 2015; Gräler et al., 2016).

The standard deviation maps in Fig. 3.5 show that the radar image is also an important covariate to help explain the residual variance, but that other covariates, such as elevation and distance to the radar station, are important too. This is most obvious from the maps for February 16, where there are significant differences between the radar image and standard deviation maps, particularly in the south-west part of the study area. Comparison of the spatial standard deviation of the residuals and kriging standard deviation maps also shows that mapping does benefit from spatial interpolation of the residual: the kriging standard deviation is substantially smaller than the spatial standard deviation, particularly in areas with high rain gauge density and high rainfall.

Comparison of the estimated trend and standard deviation coefficients with the amount of rainfall in Fig. 3.11 reveals that the upper bounds of the standard deviation coefficients are reached when very little rainfall is recorded. This can be explained from the fact that the standard deviation covariates are small when the radar rainfall is small, and hence any residual variation has to be represented by increasing the coefficient estimates. In contrast, the trend parameters are nearly always high during days of heavy rainfall. This is also as expected because the trend covariates were standardized and higher rainfall and higher rainfall variation is then modelled through higher trend regression coefficients.

Overall, the micro-scale correlation and correlation length parameters of the correlogram is insensitive to the amount of rainfall and shows a seasonal (winter/summer) pattern. In summer there is a stronger micro-scale correlation and a larger correlation distance than in winter. This may be related to rainfall type which varies by season, i.e. frontal weather systems in winter and convective rainfall in summer. The currently used correlogram is assumed isotropic due to computational simplicity, but one might consider relaxing this assumption as daily rainfall often exhibits significant anisotropy (Gyasi-Agyei, 2016).

Figure 3.7 shows that after optimization rain gauges are placed fairly uniformly over the study area, but that some parts have up to four times higher sampling density than other parts. The high density areas are those that have on average large residual and kriging standard deviation. Since radar rainfall maps vary day by day and their annual average is nearly constant in space, high density sampling areas are correlated with the other standard deviation covariates, notably elevation and distance from nearest radar station. Elevation turns out to be most important, as indicated in Table 3.4 that shows the Pearson correlation coefficients between rain gauge density and standard deviation covariates. Distance from nearest radar station is least important, which comes as a surprise but might be explained from the fact that the training data (i.e. rainfall data from the initial network) do not cover the distance from the nearest radar station feature space entirely, and hence the relationship between distance from nearest radar station and residual variation may be difficult to detect.

The decrease of the space-time average prediction error variance that results from optimizing the network is relatively modest (i.e. 5%). It is smaller than that obtained in similar studies (e.g. Baume et al., 2011; Wang et al., 2014). The main reason for the modest decrease is that we imposed a static design that must do well for all days of the year. On the long run it performs better than the initial design but there will be days where the initial design (or any other design, for that matter) will do better, simply because prediction error variance is relatively large in those parts of the study area where there is substantial rain, and these vary day by day. If sampling design optimization were applied to dynamic designs then a stronger reduction of the criterion would have been achieved, but clearly this is not a realistic option. The costs of moving rain gauges would be too high, and moreover it is difficult to predict ahead of time where areas of high rainfall intensity will be. Even though the current 5% improvement is modest, it does improve the accuracy of the resulting maps. Alternatively, optimizing the sampling design could also be used to reduce sampling costs. We evaluated this by optimizing a sampling design that uses only 90% (166) of the rain gauges used in this study. This resulted in a slightly smaller criterion value than that of the initial design using 185 rain gauges. Thus, a 10% reduction in the number of rain gauge stations can be achieved without accuracy loss, provided these are placed optimally.

For this case study, the criterion decreases at the very end of the SSA iteration process when we expect it to stabilize (Fig. 3.6). This can be explained by the coarse prediction grid that we used for calculating the space-time average kriging variance. The distance over which sampling locations are shifted becomes smaller than the grid mesh towards the end of the iteration process. The algorithm then moves points to the centre of grid cells, since these are the points for which the kriging variance is computed. We tested this hypothesis by computing the mean shortest distance from gauge locations to grid cell centres. For the optimal design it was only 10% of the expected mean shortest distance for a random design (which is 1,150 m in case of a 3 km × 3 km grid). The final drop of the criterion is thus an artefact caused by using a coarse prediction grid. It can be eliminated by using a fine prediction grid, but this would increase computing time. Considering the extent of the study area, the artefact has no serious consequences for the optimal design, since these final shifts are relatively small.

Even with a coarse prediction grid the SSA algorithm took a lot of computing time. Alternative numerical optimization methods could be tried (e.g. genetic algorithms (Behzadian et al., 2009), particle swarm optimization (Jarboui et al., 2007) or metaheuristic search (e.g. NSGAII Deb et al., 2003), but another option is to reduce the computing time during each SSA iteration step. The bulk of the work is associated with solving the kriging system, which we did by Cholesky decomposition of the covariance matrix (Section 3.2.6). However, since each SSA iteration step only involves moving one station and hence only one row and column of the covariance matrix are changed, computations could be speeded up by using block inversion (Heesterman, 1983). This would become particularly attractive when the number of rain gauges n is large. However, it might conflict with the use of parallel computing solutions.

This study optimized a spatial sampling design for a case in which spatial variation was characterized by a non-stationary variance model. To the best of our knowledge this has not been done before, but it was important because it is not realistic to assume that rainfall spatial variation is stationary. We verified this by calculating and comparing the log-likelihood and Akaike information criterion (AIC) (Akaike, 2011) for a stationary variance and non-stationary variance model. The stationary variance model was obtained from the non-stationary variance model by setting parameters κ_1, κ_2 and κ_3 to zero. Parameters were optimized using REML as before. The log-likelihood and AIC results are shown in Table 3.5. They clearly show that the non-stationary variance model is more suitable. The non-stationary variance model had a larger log-likelihood for 305 out of 315 days, while the AIC was smaller for 257 out of the 315 days. The use of a non-stationary variance model did pose some additional problems, though. We used a model in which the standard deviation is a linear combination of multiple covariates. In this respect, we extended the work of Lark (2009) or Hamm et al. (2012) by using a more complex variance component. Kriging is sensitive to a near-singular covariance matrix and different approaches may be used to avoid it (Marchant and Lark, 2007b; Marchant et al., 2009). In our case, we initially avoided near-singularity during parameter estimation by rejecting parameter combinations suggested by the differential evolution algorithm if they lead to a reciprocal condition number (Golub and Van Loan, 2012) smaller than 0.2. However, this did not completely solve the problem because during SSA optimization, new network designs are tried. It then happened that near-singularity problems were introduced at this stage, while they did not occur for the initial network. We therefore imposed further restrictions on the standard deviation regression coefficients, by requiring that none can be negative and that the intercept must be greater or equal than a small positive threshold, as explained in Section 3.2.6. This solved the near-singularity problem, but at the expense of restricting the parameter search space. Alternatively, a solution might be to model the log-transformed standard deviation as a linear function of covariates.

Table 3.5 *– Log-likelihood and Akaike information criterion (AIC) summary statistics for the stationary variance and non-stationary variance models.*

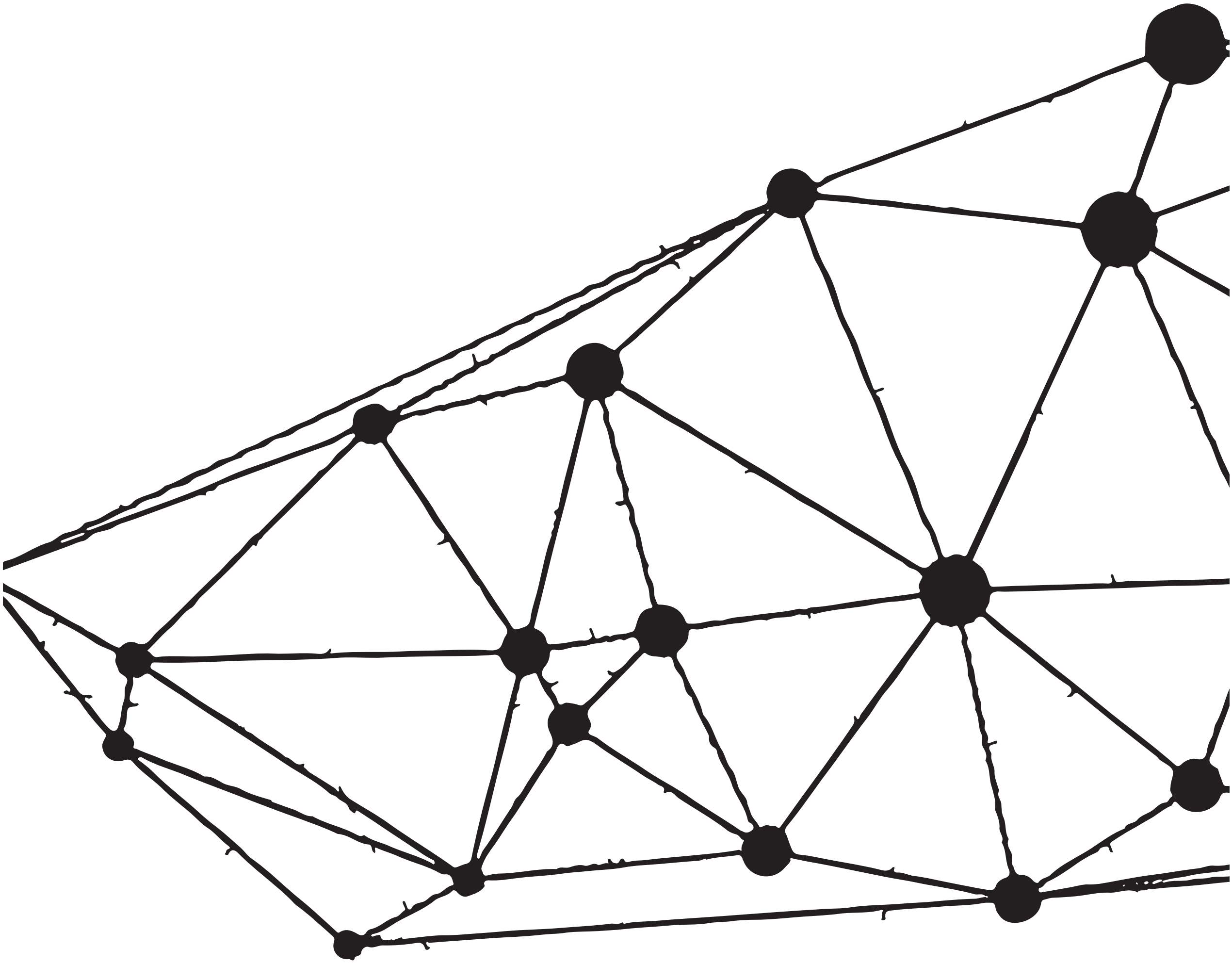
	Log-likelihood			AIC		
	Mean	Min.	Max.	Mean	Min.	Max.
Stationary variance model	-195	-672	507	400	-1004	1355
Non-stationary variance model	-152	-647	708	321	-1400	1311

Finally, we should note that while our approach included uncertainty in the trend regression coefficients, we ignored uncertainty about the standard deviation regression coefficients and correlogram parameters. The KED variance given in Eq. 3.12 may therefore underestimate the true uncertainty. In principle uncertainty about the covariance parameters can be included, such as by using a geostatistical approach (Zimmerman, 2006; Zhu and Stein, 2006; Diggle and Ribeiro, 2007a; Marchant and Lark, 2007a), but it is not obvious that the improved uncertainty assessment outweighs the substantial increase of computational complexity.

3.5 Conclusions

We extended geostatistical interpolation of rainfall data by employing a non-stationary variance KED model to predict rainfall by merging rain gauge and radar data. We optimized the rain gauge sampling design by minimizing the space-time average KED variance for a study area in England. The main conclusions are:

- Geostatistical prediction of rainfall from rain gauge data and radar data benefits from a model that incorporates non-stationarity in the mean and variance. This model matched real-world observations better than a stationary variance model, as shown by likelihood and Akaike Information Criterion statistics. Estimation of non-stationary variance parameters is hampered by (near-)singularity problems. This particular problem should be investigated more closely.
- In our case study we made spatial interpolation repeatedly over time, without accounting for the temporal structure. Temporal correlation of daily rainfall is small, but it might increase if a smaller temporal support is used, such as for hourly or 10-minute average rainfall. Space-time kriging might then be a more attractive approach.
- The standard deviation of rainfall residuals, i.e. rain gauge rainfall minus a trend mainly derived from radar rainfall, is positively correlated with radar rainfall, elevation, distance to radar station and beam blockage. In our case study the standard deviation depended more on elevation than on distance to radar station, which in turn was more important than beam blockage. Future studies may show whether this is a consistent finding or case dependent.
- Geostatistical optimization of a rain gauge network is feasible and yields plausible designs. The optimal design aims for a fairly uniform spatial distribution of the gauges, with an increased density in areas where the residual variance is large. In our case study this was in areas with high elevation, far from radar stations and near the study area boundary. The sampling density in densely sampled parts of the study area was four times higher than in sparsely sampled parts. Further work could include field accessibility in a multi-objective optimization procedure (Stumpf et al., 2016).
- Optimization of the rain gauge network leads to only a modest improvement of the space-time average prediction error variance. This is a consequence of using a static design, which cannot increase sampling density of subareas with heavy rainfall, because these subareas vary from day to day. Nonetheless, the achieved improvement is relevant and implies that savings on data collection costs could be achieved without compromising on prediction accuracy.



Chapter 4

Efficient sampling for geostatistical surveys

A geostatistical survey for soil requires rational choices regarding the sampling strategy. If the variogram of the property of interest is known then it is possible to optimize the sampling scheme such that an objective function related to the survey error is minimized. However, the variogram is rarely known prior to sampling. Instead it must be approximated by using either a variogram estimated from a reconnaissance survey or a variogram estimated for the same soil property in similar conditions. For this reason, spatial coverage schemes are often preferred, because they rely on the simple dispersion of sampling units as uniformly as possible, and are similar to those produced by minimizing the kriging variance. If extra sampling locations are added close to those in a spatial coverage scheme then the scheme might be broadly similar to one produced by minimizing the total error (i.e. kriging variance plus the prediction error due to uncertainty in the covariance parameters). We consider the relative merits of these different sampling approaches by comparing their mean total error for different specified random functions. Our results showed the considerable benefit of adding close-pairs to a spatial coverage scheme, and that optimizing with respect to the total error generally gave a small further advantage. When we consider the example of sampling for geostatistical survey of clay content of the soil, an optimized scheme based on the average of previously reported clay variograms was fairly robust compared to the spatial coverage plus close-pairs scheme. We conclude that the direct optimization of spatial surveys was only rarely worthwhile. For most cases, it is best to apply a spatial coverage scheme with a proportion of additional sampling locations to provide some closely spaced pairs. Furthermore, our results indicated that the number of observations required for an effective geostatistical survey depend on the variogram parameters.

Based on:
Wadoux, A. M. J.-C., Marchant, B. P. and Lark, R. M. (2019). Efficient sampling for geostatistical surveys. *European Journal of Soil Science*. In press.

4.1 Introduction

When mapping a continuous soil variable, geostatistical predictions at unobserved locations are made from a limited set of sampling units, called a sample. The spatial locations of those units, i.e. the sampling scheme or sampling design, has a key role in determining the cost of the survey and the quality of the predictions. Often, limited resources are available and one must adopt efficient strategies for the soil sample collection.

Several solutions have been proposed to select additional sampling sites optimally using ordinary kriging, a basic technique in geostatistics. These often require prior knowledge about the correlation function (i.e. variogram) of the target property. For example, Van Groenigen et al. (1999) proposed spatial simulated annealing (SSA) to optimize the sampling scheme so as to minimize the spatially averaged kriging variance as the objective function. This method leads to a space-filling distribution of observations, which are placed more or less evenly over the area of interest. A similar scheme can be obtained by the spatial coverage method described in Royle and Nychka (1998). They proposed a general geometric, space-filling criterion and published a point-swapping algorithm in S-plus to minimize this criterion. Brus et al. (1999) proposed the mean of the squared shortest distance (MSSD) as a geometric minimization criterion, so that it can be minimized by the fast k-means algorithm. Later this was implemented in the R language by Walvoort et al. (2010).

One advantage of coverage schemes is that they do not depend on the variogram of the soil property to be sampled. Coverage schemes are created by minimizing a criterion that is simply a function of the distance between sampling locations. Brus et al. (2007) showed that using a spatial coverage scheme led to only marginally larger mean ordinary kriging variances (MKV) than schemes where this quantity was minimized directly. The authors endorsed early geostatistical practice in soil science where sampling units were located on a regular grid (Yfantis et al., 1987).

However, regularly-spaced sampling schemes are inadequate to model the short-range variation of the soil property, which is critical for geostatistical analyses (Starks, 1986). A practical solution, as suggested for instance by De Gruijter et al. (2006, pp. 166-168), is to supplement the spatial coverage sample by a few additional units, located at short distances from the existing units. Recently, Lark and Marchant (2018) demonstrated that including such a short-distance subset markedly decreased the uncertainty of the kriging prediction for little additional effort in field data collection. Over a contrasting set of random variables, the authors proposed a simple rule that about 10% of the total sample size should be devoted to short distance units.

Using a more formal expression of the total error in a geostatistical survey, Marchant and Lark (2007a) optimized a sampling scheme by minimization of the sum of error contributions from the kriging variance and the effects of uncertainty in the variogram estimate. We refer to this objective function as the total error. The authors showed that the configuration of the optimized scheme varied according to the variogram, which was unknown prior to sampling, and used a Bayesian framework to account for a set of plausible values of variogram parameters. A similar approach was applied by Zhu and Stein (2006) for redesigning an air monitoring network. The authors noted that estimates of the variogram parameters were uncertain. They approximated the error covariance matrix of the parameters by the inverse of the Fisher information matrix, and used a Taylor series approximation of its effect on the prediction variance to account for it in their sampling objective function. For both studies, the resulting optimized schemes closely resembled the spatial coverage scheme with a small number of close-pairs of locations included, which are useful for estimating the spatial correlation over short distances. They showed that the number of close-pair locations depended largely on the variogram parameter values, and especially the variogram distance parameter.

However, the optimization procedure using a formal criterion for the minimization of the total error is complex and time-consuming. The formula for the total prediction error depends on the variogram and therefore it cannot be calculated exactly prior to sampling. Instead it must be approximated by using either a variogram estimated from a reconnaissance survey or a variogram estimated for the same soil property in similar conditions. Schemes based on approximate variograms are likely to be suboptimal. In such cases, spatial coverage sampling schemes (possibly with additional close-pairs) offer a viable, and relatively simple alternative to plan a soil survey with little or no prior information.

Surveyors must also consider the number of sampling units that are required to produce effective geostatistical predictions. The sample must be sufficient to estimate an accurate variogram function. Kerry and Oliver (2007) noted that it is generally accepted that 100 units are required to produce a reliable method of moments estimate of the variogram. This advice stems from a study of simulated random functions conducted by Webster and Oliver (1992). Kerry and Oliver (2007) subsampled four field-scale surveys of clay content and determined that a reliable residual maximum likelihood (REML) estimate of the variogram could be attained with fewer than 50 sampling units.

In summary, sampling scheme affects the uncertainty in the variogram parameters which can have an impact on the prediction error variance. Supplementing a spatial coverage sample by a simple rule of thumb reduces the prediction error variance, but the overall distribution of sample points in a scheme can be optimized, although this is laborious and requires some prior information. Whether a practical sampling strategy is markedly better when based on optimization rather than the simple rule remains an open question, and past work has not compared the approaches directly. That is what we address here.

In this research we examined empirically the difference between spatial coverage sampling schemes (sc), spatial coverage schemes supplemented with close-pairs of points (sc_+) and schemes optimized to reduce the total error. We compared these schemes with respect to the sample size required to obtain comparable results. Our objective was to show whether formal optimization is generally worthwhile, given the computational demands and the challenges of specifying prior values of variance parameters, and whether spatial coverage sampling with supplementary points is a robust practical strategy.

In our first scenario we minimized this error for a known hypothetical variogram and a given sample size. Then we determined the size of a spatial coverage scheme that would be required to achieve the same total prediction error. Similarly, we considered the size of a spatial coverage scheme plus 10% close-pairs that would also achieve the same total prediction error.

In addition to the spatial arrangement of sampling units we also considered the minimum number of units that were required to produce useful geostatistical predictions. For sample sizes larger than this minimum sample size the ordinary kriging predictor outperformed the simple random sample mean as a predictor of the values at points. For sample sizes smaller than this minimum there was no benefit from a geostatistical approach for mapping. We assumed that a geostatistical survey should, as a basic minimum requirement, ensure that local spatial predictions have an average prediction error variance that is smaller than the prediction error variance of the regional mean, estimated by design-based sampling. Webster and Lark (2012) discussed how the design-based mean can be treated statistically as a point prediction. We assumed that this design-based survey was the same size as the geostatistical survey, that the sampling units were selected according to a simple random scheme and that the corresponding design-based estimate of the mean was used as the prediction at each location.

In our second scenario we considered a geostatistical survey of soil clay content and the effect of using the average variogram of a set presented by Paterson et al. (2018) as a basis for a sampling scheme. We minimized the total prediction error variance given a sample size based on the average variogram and then repeated the tests conducted in the first scenario to find the size of the sc and sc_+ schemes that would be required to achieve the same total error as the optimized scheme for each of the clay variograms.

4.2 Materials and methods

4.2.1 Formulation of the objective function

Using the ordinary kriging formulation, we consider the situation in which the soil property (which is assumed to be a realization of a random function Z) has been measured at n locations $\mathbf{s}_i (i = 1, \ldots, n; \mathbf{s}_i \in \mathcal{A})$. The measurements $z(\mathbf{s}_i)$ are treated as realizations of $Z(\mathbf{s}_i)$ and prediction is done for Z at unobserved locations $\mathbf{s}_0$, with a known covariance parameter vector $\boldsymbol{\theta}$. Stacking the $z(\mathbf{s}_i)$ in a vector $\mathbf{z}$ and changing to matrix notation yields the ordinary kriging prediction equation (Webster and Oliver, 2007):

$$\tilde{Z}(\mathbf{s}_0|\boldsymbol{\theta}) = \boldsymbol{\lambda}^\top \mathbf{z}, \tag{4.1}$$

where $\boldsymbol{\lambda}^\top$ is the vector of kriging weights, obtained from the kriging equation:

$$\boldsymbol{\lambda} = \mathbf{A}^{-1}\mathbf{d}, \tag{4.2}$$

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \psi \end{pmatrix} = \begin{bmatrix} C(\mathbf{s}_1 - \mathbf{s}_1|\boldsymbol{\theta}) & C(\mathbf{s}_1 - \mathbf{s}_2|\boldsymbol{\theta}) & \ldots & C(\mathbf{s}_1 - \mathbf{s}_n|\boldsymbol{\theta}) & 1 \\ C(\mathbf{s}_2 - \mathbf{s}_1|\boldsymbol{\theta}) & C(\mathbf{s}_2 - \mathbf{s}_2|\boldsymbol{\theta}) & \ldots & C(\mathbf{s}_2 - \mathbf{s}_n|\boldsymbol{\theta}) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C(\mathbf{s}_n - \mathbf{s}_1|\boldsymbol{\theta}) & C(\mathbf{s}_n - \mathbf{s}_2|\boldsymbol{\theta}) & \ldots & C(\mathbf{s}_n - \mathbf{s}_n|\boldsymbol{\theta}) & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{bmatrix}^{-1} \times \begin{bmatrix} C(\mathbf{s}_0 - \mathbf{s}_1|\boldsymbol{\theta}) \\ C(\mathbf{s}_0 - \mathbf{s}_2|\boldsymbol{\theta}) \\ \vdots \\ C(\mathbf{s}_0 - \mathbf{s}_n|\boldsymbol{\theta}) \\ 1 \end{bmatrix}, \tag{4.3}$$

where ψ is the Lagrange multiplier introduced to allow minimization of the kriging variance subject to the constraint that the n weights $\lambda_1, \lambda_2, \cdots, \lambda_n$ sum to one. The covariance between the ith and jth locations is denoted by $C(\mathbf{s}_i - \mathbf{s}_j|\boldsymbol{\theta})$. The term $C(s_i - s_i)$ is the sill variance (*a priori* variance). Note that while $\mathbf{A}$ needs to be derived (and inverted) once if all observations are used for prediction at every target site, $\mathbf{d}$ must be computed for every prediction location $\mathbf{s}_0$.

From Eq. 4.1 and 4.3, the expected squared error of the prediction is given by:

$$\begin{aligned} \sigma^2_{\mathrm{OK}}(\mathbf{s}_0) &= \mathrm{var}\left(Z(\mathbf{s}_0) - \tilde{Z}(\mathbf{s}_0|\boldsymbol{\theta})\right) \\ &= C(\mathbf{s}_0 - \mathbf{s}_0|\boldsymbol{\theta}) - \boldsymbol{\lambda}^\top \mathbf{d}. \end{aligned} \tag{4.4}$$

In addition to the squared error of the prediction, Marchant and Lark (2007a) and

Zhu and Stein (2006) considered the effect of uncertainty in the estimated spatial model (variogram) parameters by a Taylor series approximation:

$$\mathrm{E}\left[\tau^2(\mathbf{s}_0)\right] = \sum_{i=1}^{q}\sum_{j=1}^{q} \mathrm{cov}(\theta_i, \theta_j) \frac{\partial \boldsymbol{\lambda}^\top}{\partial \theta_i} \mathbf{C} \frac{\partial \boldsymbol{\lambda}}{\partial \theta_j}, \tag{4.5}$$

where $\mathrm{cov}(\theta_i, \theta_j)$ is the covariance between the ith and jth parameters. This requires the variogram parameters $\theta_i (i, j = 1, \dots, q)$ to be known so that Eq. 4.5 can be approximated prior to sampling. The n-vector of partial derivatives of the kriging weights with respect to the ith variance parameter is denoted by $\frac{\partial \boldsymbol{\lambda}^\top}{\partial \theta_i}$ and can be obtained by (Marchant and Lark, 2007a):

$$\frac{\partial \boldsymbol{\lambda}}{\partial \theta_i} = \mathbf{A}^{-1}\left(\frac{\partial \mathbf{d}}{\partial \theta_i} - \frac{\partial \mathbf{A}}{\partial \theta_i}\mathbf{A}^{-1}\mathbf{d}\right). \tag{4.6}$$

The covariance between the variogram parameters can be approximated using the inverse of the Fisher information matrix $\mathbf{F}$ (Kitanidis, 1987):

$$\mathrm{cov}(\theta_i, \theta_j) \approx \mathbf{F}^{-1}(\theta_i, \theta_j) = \left(\frac{1}{2}\mathrm{Tr}\left[\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \theta_i}\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \theta_j}\right]\right)^{-1}, \tag{4.7}$$

where $\mathrm{Tr}[\cdot]$ denotes the trace of the matrix. The total error at locations $\mathbf{s}_0$, $\sigma^2_{\mathrm{P}}(\mathbf{s}_0)$, is given by the sum of the squared prediction error $\sigma^2_{\mathrm{OK}}(\mathbf{s}_0)$ and the spatial model parameter uncertainty $\mathrm{E}\left[\tau^2(\mathbf{s}_0)\right]$:

$$\sigma^2_{\mathrm{P}}(\mathbf{s}_0) = \sigma^2_{\mathrm{OK}}(\mathbf{s}_0) + \mathrm{E}\left[\tau^2(\mathbf{s}_0)\right], \tag{4.8}$$

where subsript P stands for parameter. This can be aggregated to obtain a spatial average:

$$\bar{\sigma}^2_{\mathrm{P}} = \frac{1}{\mathcal{A}}\int_{s\in\mathcal{A}} \left(\sigma^2_{\mathrm{OK}}(\mathbf{s}) + \mathrm{E}\left[\tau^2(\mathbf{s})\right]\right) \mathrm{d}s. \tag{4.9}$$

In practice, the integral $\bar{\sigma}^2_{\mathrm{P}}$ is numerically approximated by a discrete summation over a spatial grid.

4.2.2 Optimization of the sampling schemes

We start with an initial random set of sampling locations of size N, lying within the boundaries of study area $\mathcal{A}$. We assume that $Z(\mathbf{s}_i)$ is a stationary isotropic normally distributed random field, characterized by a constant mean and fitted correlation function $\rho(h)$ (h is the spatial lag or separation distance). The aim is to find the optimal sampling scheme, which minimizes the objective function (Eq. 4.9), given the parameters of $\rho(h)$. Many algorithms have been developed for solving optimization problems. We use simulated annealing (Kirkpatrick et al., 1983), extended for spatial optimization by Van Groenigen et al. (1999) for generating sequences of new possible schemes. A new sampling scheme is created by randomly shifting a randomly selected unit within the study area. This generates a new candidate scheme for which the objective function can be evaluated with Eq. 4.9, and compared with that of the previous scheme. The new candidate scheme is accepted if it has a smaller value of the objective function than the previous one. If the new scheme has a larger value of the objective function then it is accepted or rejected at random; the probability of acceptance is given by (Wadoux et al., 2017):

$$P(\text{accept}) = \exp\left(\frac{\bar{\sigma}_{\text{P}}^2(\text{old}) - \bar{\sigma}_{\text{P}}^2(\text{new})}{\alpha}\right), \tag{4.10}$$

where the control parameter α is a temperature parameter. The temperature is kept constant during a set of perturbations, called a chain, after which it is decreased to a value of $\beta \times \alpha$ for $\beta < 1$. In this way, the risk of the optimizer becoming trapped in a local but not a global minimum is reduced. We used the implementation provided by the R package spsann (Samuel-Rosa, 2017) through the optimUSER function. The initial temperature α was set to 3 with a cooling parameter β of 0.9. These were chosen so that P(accept) is close to 1 in the first chain and generally zero at the final chain. The maximum number of chains is set to 200, so that the total number of iterations is $N \times 200$. The process stops if the determined number of iterations ($N \times 200$) is reached or if the criterion remains constant for ten chains. The candidate locations are the centre of cells of a square grid.

4.2.3 Scenario 1

The first scenario considers the case where the variogram is known. We characterize the spatial correlation ρ by the second parametrization of the isotropic Matérn model (Matérn, 1986) given by (Stein, 2006, p. 31):

$$\rho(h) = \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2\nu^{\frac{1}{2}}h}{a}\right)^{\nu} \mathcal{K}_{\nu}\left(\frac{2\nu^{\frac{1}{2}}h}{a}\right), \tag{4.11}$$

where h is the separation distance, $\mathcal{K}_\nu$ is the modified Bessel function of the second kind of order ν (see Abramowitz and Stegun, 1972, pp. 374-379) and Γ is the gamma function. The correlation function $\rho(h)$ has parameters a and ν. Parameter a is the distance parameter which indicates how fast the correlation decays with increasing h and ν is the smoothness parameter. Stein (2006) noted that ν is the critical parameter in the Matérn correlation model. The larger is ν, the smoother is Z. We chose a Matérn model for its flexibility in modelling the spatial covariance with a small number of parameters (Minasny and McBratney, 2005).

For the first scenario, we generated a square area of 100 m × 100 m. Spatial coverage schemes of size $N = 60, 61, \ldots, 200$ are derived by discretization of the area into N geographical strata using the stratify method from the R package spcosa (Walvoort et al., 2010). The spatial coverage units are taken in the centroid of the strata, which is equivalent to minimizing the mean squared shortest distance between a location in the region and the nearest sampling location. In addition, we also generated samples of size $N = 60, 61, \ldots, 200$ in which the sampling locations were distributed according to a spatial coverage scheme, with a subset of 10% of units positioned at an arbitrary distance that was short relative to the spacing between neighbouring points in the basic spatial coverage survey. This arbitrary short distance was set to 2 m because of a mean spacing between neighbouring locations in the sc scheme of 6.6 m for $N = 60$ and 12.5 m for $N = 200$. These close-pair units were selected by simple random sampling without replacement in a randomly chosen direction from 0-360 degrees. We repeated the selection of close-pairs several times to determine the sampling variation in total variance. Since the latter was small, we did not pursue any further because this confirmed the very tight confidence intervals in the Lark and Marchant (2018) study.

We considered different sets of variogram parameter values, all of which had a total sill variance of one. Four values of ν were tested; $\nu = 0.5$ (equivalent to the exponential variogram), $\nu = 0.2$ (rougher than the exponential), $\nu = 1.1$ and $\nu = 2$ (smoother than the exponential). These four ν values were combined with each of three distance parameter: $a = 10, 20$ and 30 and three ratios of the nugget (c_0) to total sill variance ($c_0 + c_1 = 1$) for strong ($c_0 = 0$), moderate ($c_0 = 1/3$) and weak ($c_0 = 2/3$) spatial dependence. Note that we use the nugget to sill ratio to characterize the spatial dependence of a model with known parameters, but this should not be done when comparing empirical variograms because the magnitude of the nugget variance is likely to depend in part on the sampling design. Each of the $4{\times}3{\times}3 = 36$ scenarios were optimized for a fixed sample size $N = 90$, in the way described in the previous section. To speed up computations the criterion was evaluated at 34 × 34 locations on a regular square grid of spacing 3 m.

In this scenario we compared for each variogram the size of the sc and sc$_+$ samples required to attain the same value of the objective function as the optimized scheme of 90 units.

We also compared the average total prediction variance that resulted from the geostatistical survey of each random function with the prediction variance that would result from using an estimate of the simple random sample of the field mean as a predictor of the value at points. If the design-based survey consists of N locations selected by simple random sampling this prediction variance is equal to (Brus et al., 1992, Equation 7):

$$\overline{\sigma}^2_{\mathrm{DB}} = \sigma^2 \left(1 + \frac{1}{N}\right), \tag{4.12}$$

where σ^2 is the dispersion variance (the variance of the variable within the study area) and the σ^2/N term reflects the uncertainty in estimating the field-scale mean of the property of interest with simple random sampling (Brus and De Gruijter, 1993). Instead of the spatial variance (dispersion variance) for a single realization, we used the model expectation of the dispersion variance in Eq. 4.12, so that the model expectation of the spatial mean of the design-based estimation error variance at points was also obtained. For each set of variogram parameters, we determined the smallest sample size of a geostatistical survey which led to the average total prediction variance being less than this design-based prediction variance. We determined the dispersion variance for each random function from the average variance of 1000 lower-upper (LU)-simulations of the function at 2000 random locations across the study area.

4.2.4 Scenario 2

The second scenario considered a survey of soil clay content where no field-specific information about the variogram was available. In such a circumstance, McBratney and Pringle (1999) suggested that the average of previously published soil clay variograms should provide useful information for assessing soil sampling schemes.

Here we used data from a published study on field-scale variability of soil variograms. We used a compilation of soil clay variograms, provided by Paterson et al. (2018). They were gathered from the existing literature, based on untransformed data and physical measurements. We converted the exponential, spherical and linear clay variograms to a Matérn model (Eq. 4.11) by re-estimating their parameters using a least squares approach. In this way, we compared surveys using variograms with same number of estimated parameters. From the set of Matérn clay variograms, we derived an average experimental variogram as in McBratney and Pringle (1999). Each variogram for soil clay was evaluated at a set of closely-spaced lag intervals. Each value of semivariance was transformed to its fourth root. The average value of the fourth root of the variogram was computed at each lag interval over all the clay variograms and the resulting values were back-transformed to their fourth power. The fourth root is used to give a normally distributed variable even when the underlying variable includes extreme values (Cressie and Hawkins, 1980). Finally, a Matérn correlation function (Eq. 4.11) was fitted by non-linear least squares to the average experimental variogram. The estimated Matérn correlation function was similar to an exponential variogram ($\nu = 0.5$) with a nugget variance $c_0 = 2.6$, a partial sill $c_1 = 8.0$ and a distance parameter $a = 44.1$ m (effective range is about 85 m). We then optimized the distribution of 90 sample units within a 500 m × 500 m region, using the mean total prediction error variance, Eq. 4.9, as the objective function specifying the parameters of the average variogram. The objective function was evaluated at a centred square grid of 25 × 25 points with a spacing of 20 m. We then found, for the random function with parameters estimated for each clay variogram, the value of the objective function achieved by optimizing sample schemes of size $N = 60, 61, \ldots, 200$, and the corresponding number of observations in an sc and an sc$_+$ scheme required to match the value of the objective function achievable by optimization with the average clay variogram.

4.3 Results

4.3.1 Scenario 1

Figures 4.1, 4.3, 4.5 and 4.7 show 90 unit sampling schemes optimized to minimize the expected total error with different values of the nugget to sill ratio, different distance parameters a and smoothness parameters of 0.2, 0.5, 1.1 and 2, respectively. In all schemes, the sampling locations are generally evenly dispersed over the area with some close-pair units. When the nugget to sill ratio increases (larger c_0), the number of close-pairs tends to increase substantially. The pattern for larger values of the distance parameter a is reversed. The larger is a, the smaller are the transects of close-pairs. When $c_0 = 2/3$, $c_1 = 1/3$ and $a = 10$ the sample size seems insufficient to cover the whole area. This might indicate that for this variogram and study area, 90 units were insufficient to estimate both the variogram and predict the soil property across the region. All values of ν tested had comparable patterns for the optimized schemes.

Figures 4.2, 4.4, 4.6 and 4.8 show the values of the objective function for each variogram type for the sc or sc$_+$ schemes compared to the values of the objective function from the optimized 90-unit sample scheme. The values of objective function for the sc schemes show a rougher pattern than those of the objective function for the sc$_+$ schemes. The sc schemes performed poorly in most cases. The poor performance was less pronounced for large values of a when ν was 1.1 or 2. In such cases, sc schemes were only slightly worse than the optimized schemes. The sc$_+$

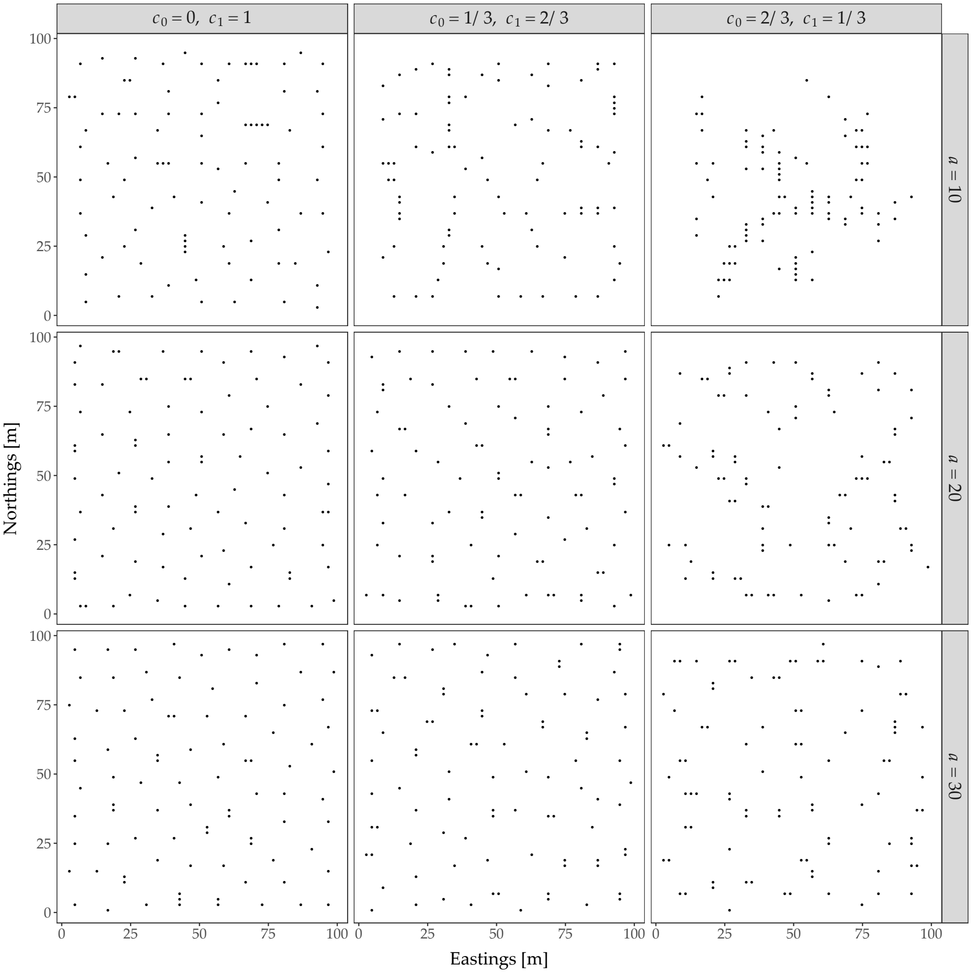


***Figure 4.1** – Optimized 90-sample schemes for different variogram parameters and $\nu = 0.2$.*

schemes always performed slightly worse than the optimized schemes. With increasing nugget to sill ratio, the sc$_+$ schemes needed an increasing number of additional units to reach the same value for the objective function as the optimized sample scheme. This was valid for all values of ν tested.

For each set of variogram parameters, Table 4.1 reports the number of additional samples necessary when using the sc$_+$ scheme to reach the objective function of the optimized 90-unit scheme. Overall, the sc$_+$ scheme needs at least 8 and a maximum of 59 additional units to achieve the objective function of the optimized 90-unit scheme. As mentioned previously, there is a clear trend associated with the nugget

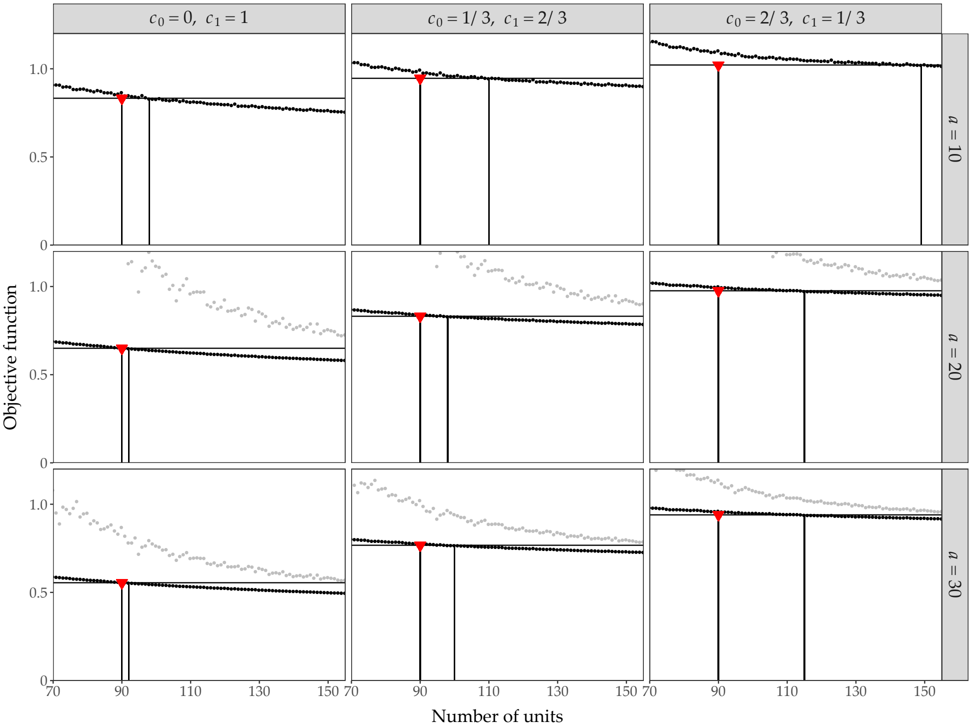


***Figure 4.2** – Value of objective function for sc$_+$ (black dots), sc (grey dots) and optimized (red triangle) schemes. The spacing between the two vertical lines indicates the number of extra units required for sc$_+$ to achieve an optimized objective function value for $\nu = 0.2$.*

***Table 4.1** – Additional number of sampling units in the sc$_+$ scheme required to achieve the same objective function as that of the optimized 90-unit survey.*

c_0	c_1	a	$\nu = 0.2$	$\nu = 0.5$	$\nu = 1.1$	$\nu = 2$
0	1	10	8	8	9	8
1/3	2/3	10	20	10	7	5
2/3	1/3	10	59	34	24	19
0	1	20	2	6	8	11
1/3	2/3	20	8	3	4	8
2/3	1/3	20	25	16	18	23
0	1	30	2	6	11	16
1/3	2/3	30	10	5	11	15
2/3	1/3	30	25	25	11	20

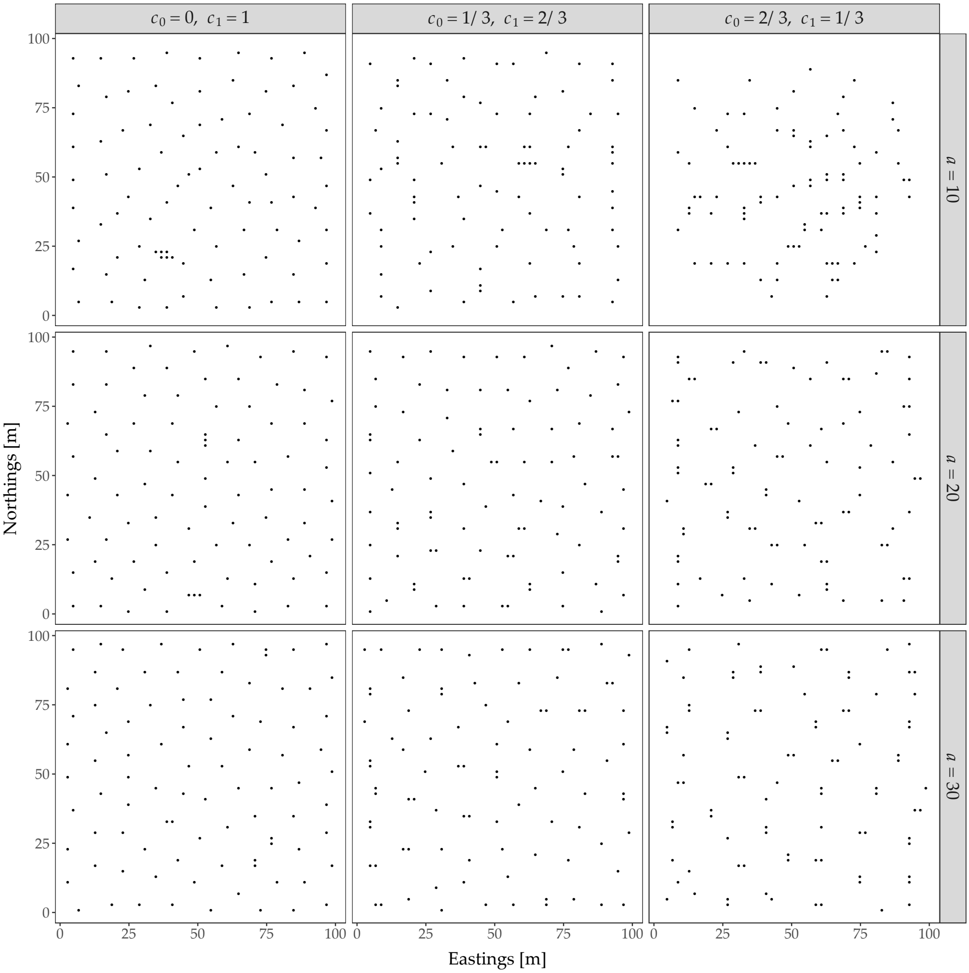


***Figure 4.3** – Optimized 90-unit schemes for different variogram parameters and $\nu = 0.5$.*

to sill ratio. The larger is the ratio, the larger is the number of additional units in the sc$_+$ scheme. This effect was slightly diminished for increasing values of a.

Figure 4.9 shows the objective function for the sc and sc$_+$ schemes for the case where the smoothness parameter $\nu = 0.5$ was either fixed (known) or estimated (with uncertainty) with parameters $c_0 = 0$, $c_1 = 1$ and $a = 20$. When the smoothness was estimated there was a marked difference between the total error variance for the sc and sc+ schemes when there were fewer than about 200 sample points in total. With larger sample sizes (above 220) the difference became negligible. When the smoothness is known (equivalent to assuming an exponential variogram), there

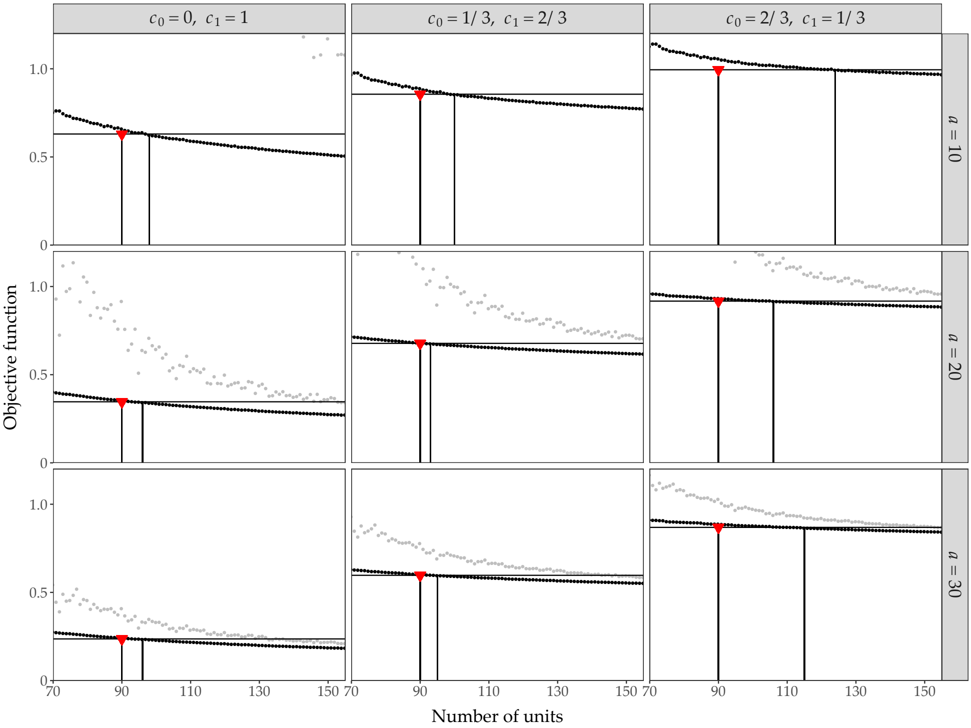


***Figure 4.4** – Value of objective function for sc$_+$ (black dots), sc (grey dots) and optimized (red triangle) schemes. The spacing between the two vertical lines indicates the number of extra units required for sc$_+$ to achieve an optimized objective function value for $\nu = 0.5$.*

were still minor differences between the sc and sc$_+$ scheme objective functions but they rapidly converged to the same values (from about 120 units).

Table 4.2 shows the minimum sample size required for the expected total variance to be smaller than the estimation variance of the target property that would result from a design-based survey of the same size. The sc$_+$ schemes needed on average fewer units than the sc schemes. There is a clear association between an increase in the required sample size, increase of nugget to sill ratio and decrease in the smoothness and distance parameters. When compared to the effective range of the target property (i.e. the distance at which the spatially correlated portion of the variogram attains 95% of the sill), the minimum number of units increased with decreasing values of the effective range. The dispersion variance (denoted σ^2 in Table 4.2) increased with larger values of nugget to the sill ratio and larger values of the distance parameter.

Table 4.2 *– Minimum number of units required for the expected total prediction error variance to be smaller than the estimation variance of the target property that would result from a design-based survey of the same size. The dispersion variance is derived by averaging the variance of 1000 simulations using the LU-decomposition (Davis, 1987). The simulations are realized using 2000 units, selected by simple random sampling. In addition, the effective ranges of the different variogram types, denoted r, are reported.*

c_0	c_1	a	$\nu = 0.2$				$\nu = 0.5$				$\nu = 1.1$				$\nu = 2$			
			σ^2	r	sc	sc$_+$	σ^2	r	sc	sc$_+$	σ^2	r	sc	sc$_+$	σ^2	r	sc	sc$_+$
0	1	10	0.97	22	>200	75	0.98	21	164	61	0.97	20	104	54	0.98	19	95	49
1/3	2/3	10	0.98	22	>200	79	0.99	21	>200	52	0.98	20	128	67	0.98	19	109	72
2/3	1/3	10	0.99	22	>200	>200	0.99	21	>200	83	0.99	20	>200	79	0.99	19	158	72
0	1	20	0.93	45	95	28	0.93	42	84	20	0.93	40	54	20	0.92	38	42	24
1/3	2/3	20	0.95	45	>200	77	0.95	42	95	24	0.94	40	62	24	0.94	38	48	20
2/3	1/3	20	0.98	45	195	145	0.98	42	124	66	0.97	40	163	65	0.97	38	163	>200
0	1	30	0.88	67	104	20	0.87	64	54	22	0.83	60	31	11	0.85	57	24	13
1/3	2/3	30	0.91	67	77	23	0.90	64	62	16	0.91	60	48	16	0.87	57	45	15
2/3	1/3	30	0.96	67	>200	136	0.95	64	92	147	0.94	60	72	27	0.93	57	73	46

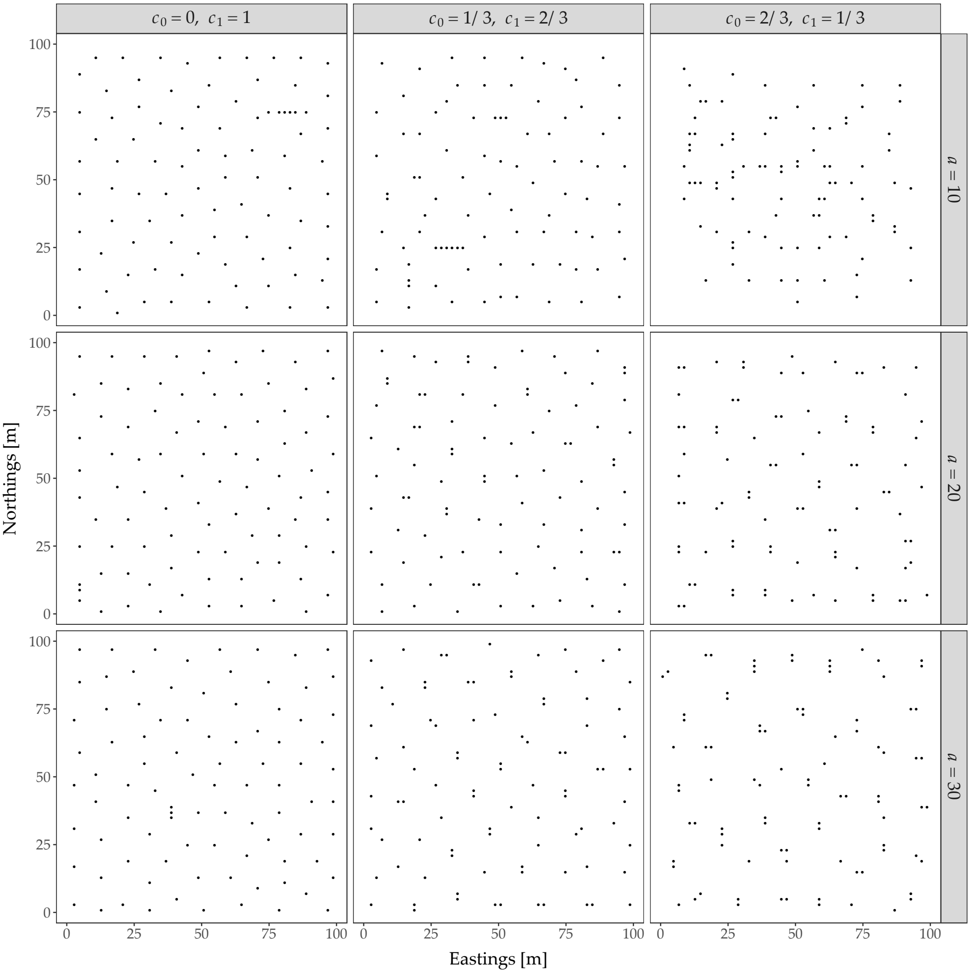


Figure 4.5 *– Optimized 90-unit schemes for different variogram parameters and* $\nu = 1.1$.

4.3.2 Scenario 2

Figure 4.10 shows an example of sc, sc$_+$, as well as the optimized 90-unit scheme obtained by minimization of the expected total error using the average soil clay variogram. The optimized scheme had sampling units dispersed evenly over the area with a number of close-pair units. The number of close-pair units seems slightly larger than that of the sc$_+$ scheme. While the sc$_+$ and optimized scheme share some similarity in the pattern of sampling locations, the sc scheme is very different from the optimized scheme.

This is confirmed by Fig. 4.11 which shows values of the objective function for sc$_+$, sc

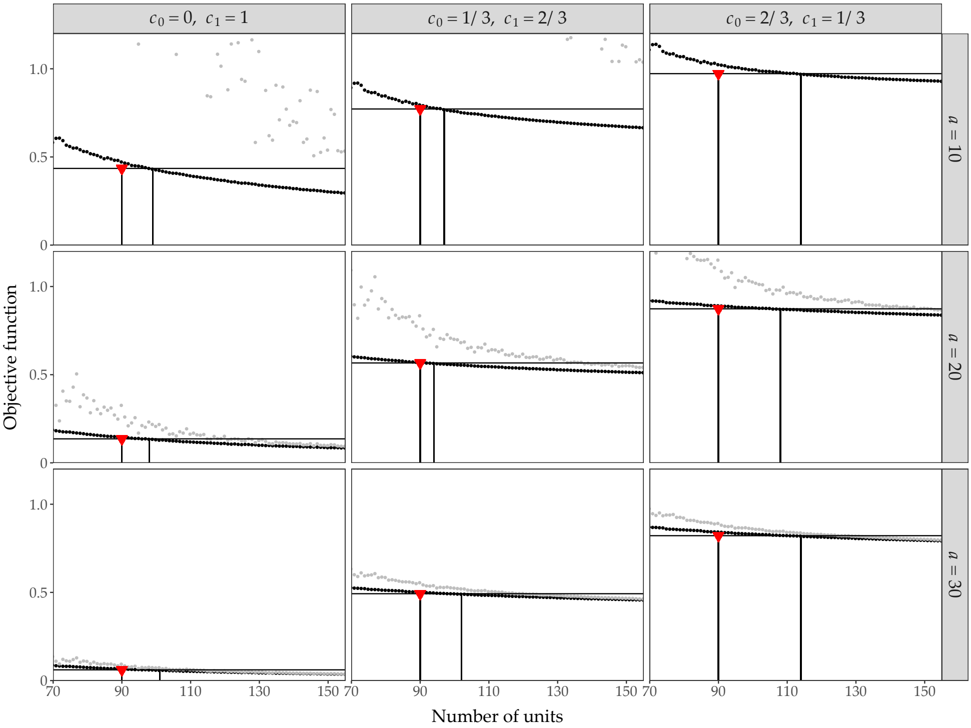


***Figure 4.6** – Value of objective function for sc$_+$ (black dots), sc (grey dots) and optimized (red triangle) schemes. The spacing between the two vertical lines indicates the number of extra units required for sc$_+$ to achieve an optimized objective function value for $v = 1.1$.*

and optimized schemes using the average clay variogram. The sc scheme performed poorly until about 200 units. In contrast, the sc$_+$ had objective function values closer to that of the optimized scheme. Twenty-two additional locations were required for the sc$_+$ scheme to reach the objective function of the optimized scheme, which was achieved with a total of 11 close-pairs in the sc$_+$ scheme (out of 112).

Figure 4.12 shows the standardized soil clay variograms and the average variogram. First, the average variogram was used to compute the optimized scheme. Second, we found the sample size for the sc$_+$ scheme for each separate clay variogram to achieve the total variance of the optimized scheme. Overall, the optimized scheme was fairly robust with contrasting standardized soil clay variograms because it gave about the same total variance for most of the individual variograms as for the average variogram. Fig. 4.12 shows that a large number of additional units were needed (> 100) when large sill values of the variogram were reached in a short distance. In addition, fewer units were needed (< -5) when the total sill was reached at large

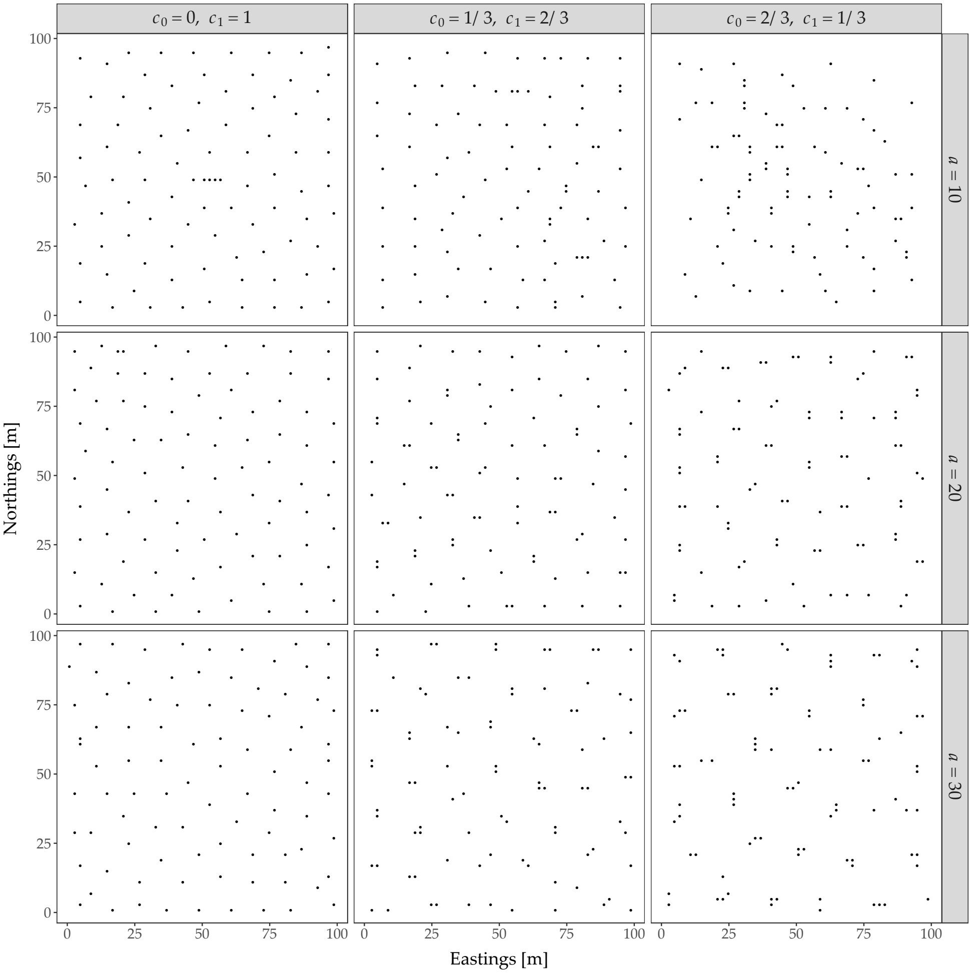


***Figure 4.7** – Optimized 90-unit schemes for different variogram parameters and* $\nu = 2$.

distances. For similar values of the distance parameter, more units were needed for larger values of the nugget variance, e.g. weaker spatial dependence at short distances (see for example the two clay variograms with similar distance parameters but different nugget values).

4.4 Discussion

For all optimized schemes, there was a number of close-pair units. This shows that sampling units at short distances had a critical effect on decreasing the total expected error (which encompasses uncertainty in the variogram parameters and

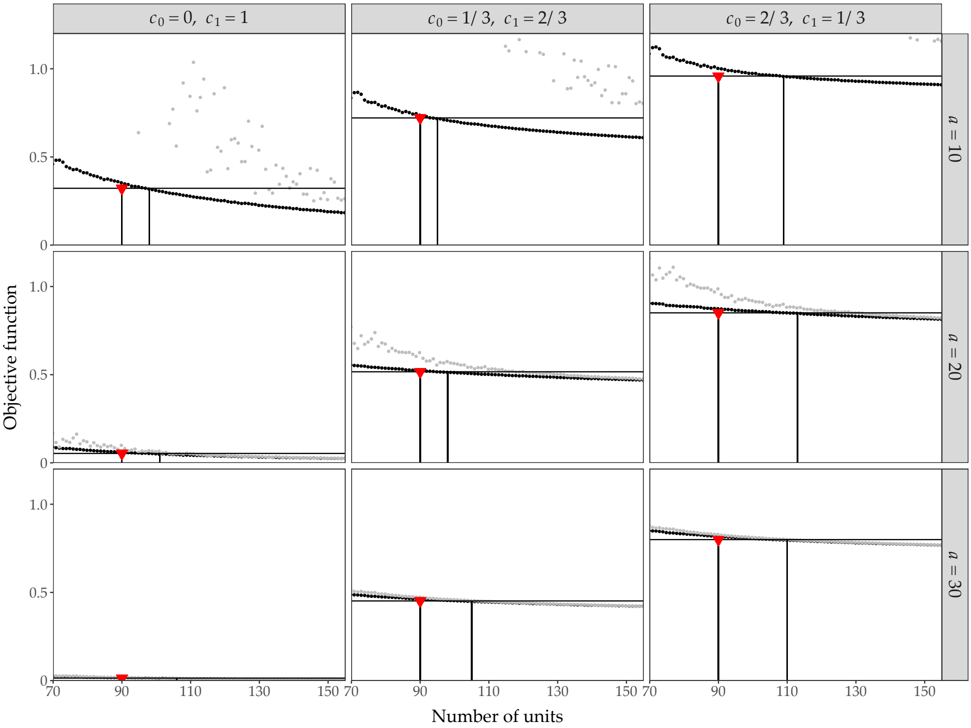


Figure 4.8 *– Value of objective function for sc$_+$ (black dots), sc (grey dots) and optimized (red triangle) schemes. The spacing between the two vertical lines indicates the number of extra units required for sc$_+$ to achieve an optimized objective function value for $v = 2$.*

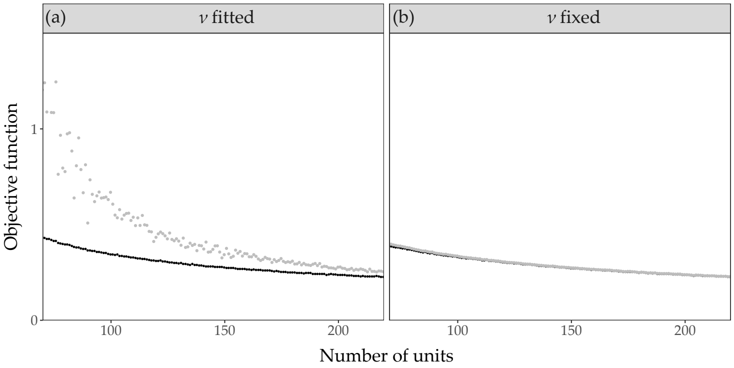


Figure 4.9 *– Value of objective function for sc$_+$ scheme (black dots) and sc scheme (grey dots) for $c_0 = 0$, $c_1 = 1$, $a = 20$ and $v = 0.5$. In (a) the smoothness parameter is to be estimated while in (b) it is assumed to be known.*

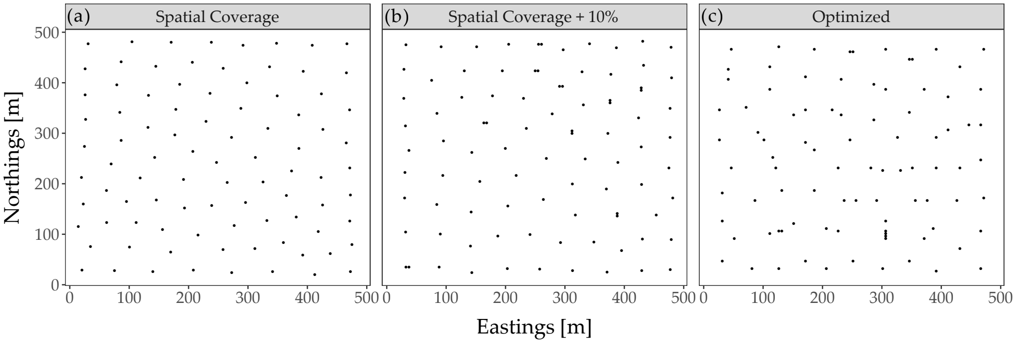


Figure 4.10 *– Example of 90-unit sc scheme (a), sc_+ scheme (b) and optimized for the average soil clay variogram (c).*

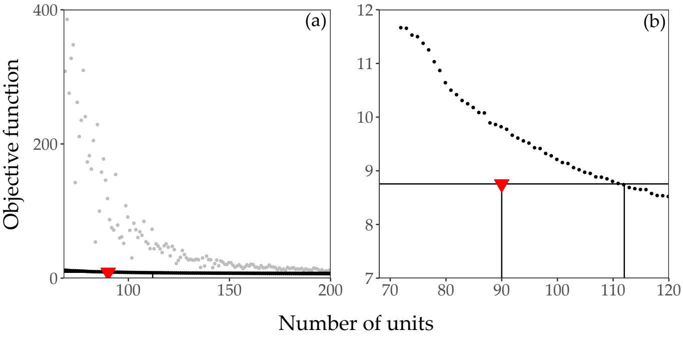


Figure 4.11 *– Values of objective function for sc_+ scheme (black dots) and sc scheme (grey dots) and optimized (the red triangle). The spacing between the two vertical lines in (b) indicates the extra units required for sc_+ to achieve an objective function value from the optimized scheme for the soil clay average variogram.*

kriging variance). The number of close-pair units increased according to the nugget to the sill ratio and to a lesser extent relative to the distance parameter. This was an expected result, because a random variable with a small spatial correlation distance and large nugget to sill ratio had to be sampled at a large number of short distance locations to ensure minimization of uncertainty in both variogram parameters and prediction error variances (Marchant and Lark, 2007a). This explains why sc schemes performed poorly in all cases. The sc schemes lacked close-pair units to estimate the spatial correlation over short distances which have a large effect on total expected error. Sampling schemes containing a subset of 10% as close-pairs (suggested by Lark and Marchant, 2018) provide a robust strategy to ensure a reasonably small total expected error. In the case of a small distance parameter or large nugget to sill ratio, 10% of close-pairs does not provide sufficient information and it is better to either increase the ratio of units taken at short distances or to use an

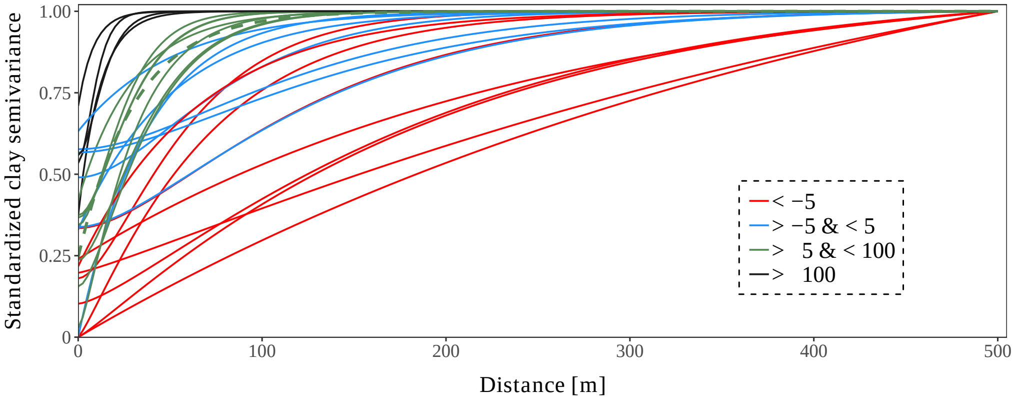


Figure 4.12 *– Standardized variograms of soil clay. The dashed line represents the average variogram. The colour defines the number of additional sc$_+$ units required to reach the same value of the objective function for each variogram as the scheme optimized with the average variogram.*

optimized scheme.

The test presented in Fig. 4.9 suggests that the importance of close pairs is reduced if the smoothness parameter is assumed to be known. In practice, however, this was not the case. Assuming a particular smoothness value (e.g. 0.5 for the exponential variogram) for a regularly sampled soil property led to a substantial proportion of the uncertainty being disregarded. This choice was somewhat subjective because it was related to the decision of the modeller and the range of possibilities we allowed in our model. We point out that close pairs are not only important when the nugget to sill ratio is large (Table 4.1) and the range of spatial correlation ($3 \times a$ if $\nu = 0.5$) is small relative to the size of the study area (Table 4.2), but also when one needs to estimate the additional Matérn model parameter ν (Fig. 4.9).

Results from our second scenario showed that the optimized scheme based on the average variogram was fairly robust for contrasting soil clay variograms. For several variograms, an sc$_+$ scheme outperformed the optimized scheme. This was an unexpected result at first sight. The reason is that the sampling scheme was optimized for the average variogram, and can therefore be suboptimal for an individual variogram. The results of the second scenario suggested that databases of variogram parameters (e.g. the one of Paterson et al., 2018) can be used to derive an average variogram, and that the latter can be used to guide sampling (McBratney and Pringle, 1999) or to predict a soil property from fewer units than usually required for estimating variogram parameters (Kerry and Oliver, 2004). An average variogram could also provide prior information for expert or Bayesian elicitation of the variogram (Cui et al., 1995).

In our two scenarios, close-pair units were taken at a fixed distance from one of the spatial coverage units. There might be room for further research on how these close-pair units should be selected. For example, in several optimized schemes, transects of several units can be seen. Further tests on our scenario 1 (not shown) suggested that selecting close pairs in a cluster might reduce substantially the number of additional units needed with an increasing nugget to sill ratio. Such a scheme would, however, rely heavily on the assumption of stationarity (i.e. that the short-scale variation in the cluster indicates the short-scale variation across the study area). Our results here were for ordinary kriging in which the local mean of the variable was assumed to be constant. Sampling to support universal kriging (to model a non-stationary mean), or to support kriging with external drift (to model dependency of the mean on covariates) introduces other considerations, specifically the estimation of the fixed effects parameters in the model. This requires further work. We speculate that the supplemented spatial coverage schemes that we have shown to be efficient for ordinary kriging would also be efficient for universal kriging, in that the spatial coverage points would ensure reliable estimation of trend parameters, and the close-pairs would similarly ensure that the variance parameters are estimated precisely

For the optimized schemes in scenario 1, derived from a variogram with a small distance parameter and large nugget to the sill ratio, the sample size seems too small compared to the size of the study area. Table 4.2 shows that this is indeed the case for several variogram types, and especially if the units are based on a spatial coverage scheme. This can lead to situations where the expected total variance is larger than the total sill variance. In such circumstances adding close-paired observations might resolve with the problem of parameter estimation, but the overall sampling scheme remains inadequate for the task of spatial mapping because the spacing between neighbouring observations in the spatial coverage scheme is not sufficiently small relative to the range of spatial dependence. If one kriges from a grid with spacing larger than the range, then the prediction error variance is equal to the sill variance plus the Lagrange parameter, which is equivalent to the second term for the prediction variance of the spatial mean as a point predictor in Eq. 4.12. This points us to the fact that, in these circumstances, where we cannot afford a grid with spacing that is small relative to the range, spatial prediction by kriging is not an option. In these circumstances point prediction might, in the worst case, be the regional mean of the variable, estimated by design-based sampling and with a prediction error variance computed from Eq. 4.12. It might be possible to do better by estimating mean values within subregions of the area of interest such as soil map units (Webster and Beckett, 1968), again by design-based sampling, or by undertaking design-based sampling to estimate parameters of a predictive relation between the soil property of interest and covariates such as data from remote sensors. We hope that this clarifies why we refer to design-based estimation in the paper. It is not the case that design-based sampling does not provide a basis for spatial prediction. Design-based simply refers to the sampling scheme (probability sampling) and the basis for estimation from the data. The resulting design-based mean (for a region, or subregion) may then be treated as a spatial prediction, as discussed by Webster and Lark (2012).

Table 4.2 also shows that for large value of smoothness ($\nu = 1.1$ or 2) and small nugget to sill ratios, the minimum number of units needed to make geostatistical analysis more accurate than a design-based estimate, on average, is surprisingly small. This can be explained by the relatively large values of the effective range ($r = 60$ and $r = 57$) for the case study (square of 100 m × 100 m); most sampling units were within the range of spatial correlation. However, for random functions with larger nugget to sill ratios, the design-based survey was more accurate even when the survey consisted of more than 200 units. Thus, the number of units required to estimate the variogram from a geostatistical survey depended on the degree of spatial correlation of the target property. We acknowledge that these total prediction variances are based upon a Taylor series approximation to the true variances.

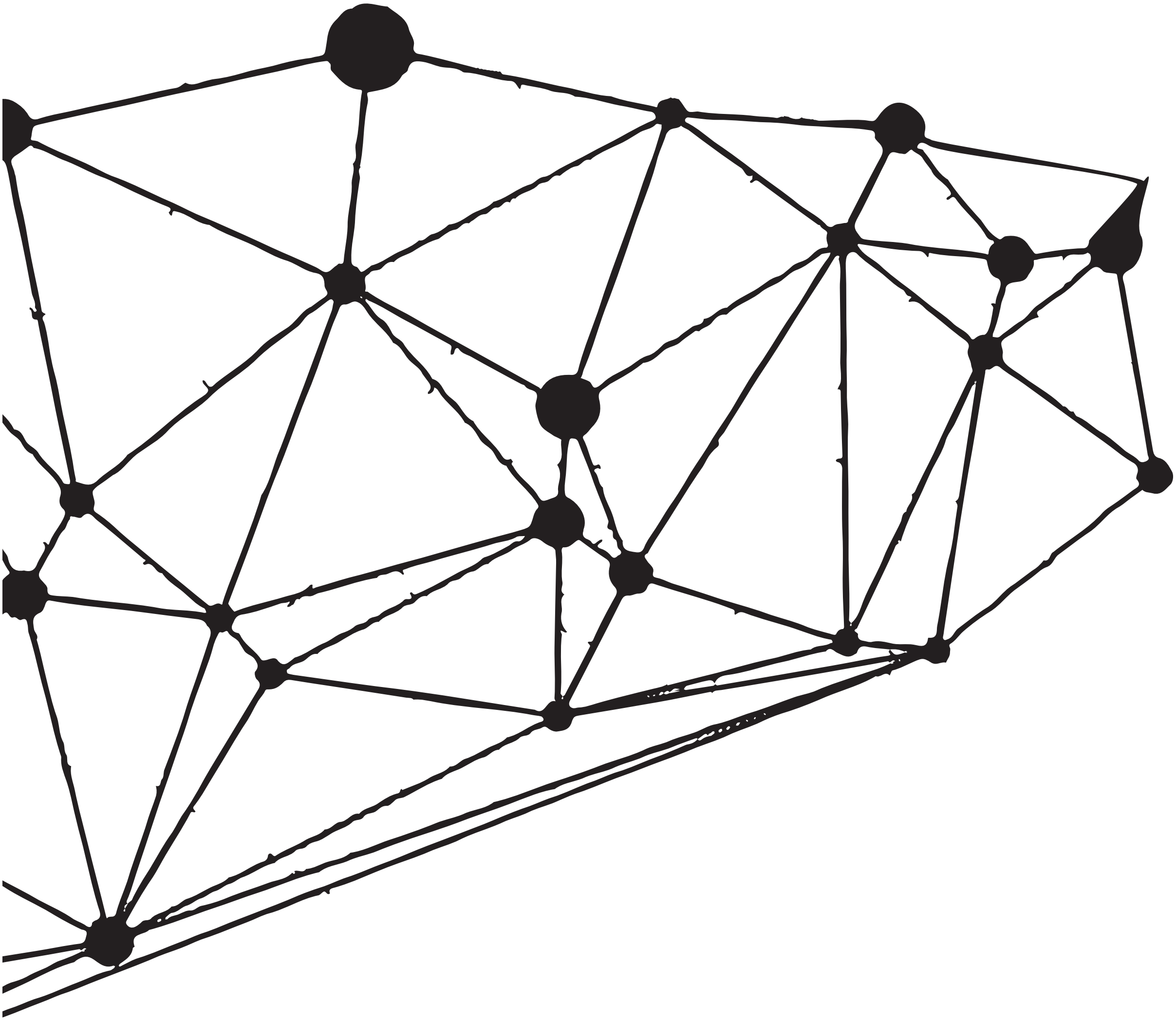
4.5 Conclusions

From the results and discussion we draw the following conclusions:

- The sc schemes performed poorly in almost all cases because of the lack of information at short distance to estimate the variogram parameters.
- Uncertainty of the sc scheme was mainly characterized by uncertainty of the smoothness parameter. Performance of the sc scheme can therefore be greatly improved by assuming that the smoothness is known, for example with an exponential variogram. However, in practice we have no justification for making such an assumption.
- The benefit of using an optimized scheme over an sc$_+$ scheme was clear but still generally modest. In addition, the optimization required the variogram parameters to be known.
- The benefit of using an optimized scheme over an sc$_+$ scheme became more important with an increasing nugget to sill ratio (weaker spatial dependence). In this case, geostatistical survey was unlikely to be effective.
- For a random variable with zero nugget and a large range of spatial correlation fewer than 15 observations were required to obtain average total prediction

variances that were smaller than the prediction variance of the design-based estimate of the regional mean, treated as a point prediction at each location. However, 200 observations of a random variable with a substantial nugget effect were insufficient to meet the same criterion.

- When the scale of spatial variation of the soil property was not known, using an average variogram for optimizing the sampling scheme is a robust strategy.
- Overall, the tests conducted showed that there was little evidence of large benefits from optimizing sampling schemes. Therefore, it is better in most cases to use a spatial coverage scheme supplemented by a subset of close-pair units unless prior knowledge of the variogram is available (e.g. reconnaissance survey).



Chapter 5

Sampling design optimization for soil mapping with random forest

Machine learning techniques are widely employed to generate digital soil maps. The map accuracy is partly determined by the number and spatial locations of the measurements used to calibrate the machine learning model. However, determining the optimal sampling design for mapping with machine learning techniques has not yet been considered in digital soil mapping studies. In this paper, we investigate sampling design optimization for soil mapping with random forest. A design is optimized using spatial simulated annealing by minimizing the population mean squared prediction error (MSE). We applied this approach to mapping soil organic carbon for a part of Europe using subsamples of the LUCAS dataset. The optimized subsamples are used as input for the random forest machine learning model, using a large set of readily available environmental data as covariates. We also predicted the same soil property using subsamples selected by simple random sampling, conditioned Latin Hypercube sampling (cLHS), spatial coverage sampling and feature space coverage sampling. The process is repeated several times using leave-group-out cross-validation so as to compute the calibration sampling distribution of the MSE of maps based on different sampling designs. Differences between MSE distributions are tested for significance using the non-parametric Mann-Whitney test. The process was also repeated for different sample sizes. We analysed the spread of the optimized designs in both geographic and feature space to reveal their characteristics. Results show that optimization of the sampling design by minimizing the MSE is worthwhile for small sample sizes. However, an important disadvantage of sampling design optimization using MSE is that it requires known values of the soil property at all locations and as a consequence is only feasible for subsampling an existing dataset. For larger sample sizes, the effect of using an MSE optimized design diminishes. In this case, we recommend to use a sample spread uniformly in the feature (i.e. covariate) space of the most important random forest covariates. The results also show that for our case study cLHS sampling performs much worse than the other designs for mapping with random forest.

Based on:
Wadoux, A. M. J.-C., Brus, D. J. and Heuvelink, G. B. M. (2019). Sampling design optimization for soil mapping with random forest. *Geoderma.* Accepted with revision.

5.1 Introduction

Conventional digital soil mapping (DSM) employs geostatistical techniques to predict a continuous soil property at unobserved locations from measurements of this property at a finite number of sampling locations. Prediction is usually improved by exploiting the quantitative empirical relationship between the soil property and one or several environmental covariates. This leads to kriging with external drift, a basic technique in geostatistics, in which a soil property is modelled as a sum of a linear combination of covariates and zero mean, spatially auto-correlated residuals. Kriging models the soil property in a comprehensive, statistically sound way, but has several limitations (Webster and Oliver, 2007). First, it typically assumes that the residuals are normally distributed, stationary and isotropic. Second, it considers that the model of spatial variation (i.e. the variogram) is estimated without error. Finally, the relation between the soil property and the covariates is usually assumed to be linear, and difficult to model when using a large number of correlated covariates.

As an alternative, in recent decades (supervised) machine learning (ML) techniques have been applied for spatial prediction and DSM. ML refers to a large class of nonlinear data-driven algorithms developed first for data mining and pattern recognition purposes. But ML is increasingly being used in other quantitative fields, such as in predictive soil mapping. ML techniques do not rely on rigid statistical assumptions about the distribution of the soil property and can handle numerous and correlated covariates as predictors, if at least a large calibration dataset is available. Examples on the use of ML techniques for DSM are Henderson et al. (2005) for mapping multiple soil properties at national-scale using decision trees, Behrens et al. (2005) for predicting soil units using artificial neural networks, and Grimm et al. (2008) to map soil organic carbon using random forest. In this study we use the latter technique, whose use for soil mapping was recently formalized in Hengl et al. (2018).

Mapping requires calibrating a model using a sample from the target population. In consequence, the map accuracy is partly determined by the sample size and spatial locations of the sampling units with measurements of the target property used to calibrate the model. Various sampling designs are potentially suitable, depending on the intended mapping technique. In most cases, the soil is mapped using a known model of spatial variation (e.g. a variogram when using kriging). In this context it is sensible to select a sample whose units are spread evenly throughout the area. This can be achieved by spatial coverage sampling (Royle and Nychka, 1998; Walvoort et al., 2010). If one assumes that the soil property is linked to environmental covariates, a robust strategy is to ensure that the measurements are also spread in the feature (i.e. covariates) space. This can be achieved using conditioned Latin Hypercube sampling (cLHS) (Minasny and McBratney, 2006) or feature space coverage sampling using the k-means (Hartigan and Wong, 1979) algorithm. The spatial coordinates can be added to the set of covariates so as to ensure a spread in both geographical and feature space. Brus (2019) noted that there is no single best sampling design, and that the best design depends on the technique used for mapping.

If the mapping technique is known beforehand, it is judicious to optimize a design for the intended use. In a model-based setting, we obtain an estimate of the prediction error variance, which can be minimized. For mapping with ordinary kriging, this leads to a fairly uniform spread of the measurements in the geographic space, which can be obtained using a spatial coverage design (Brus et al., 2007). If one or several covariates are used as a trend in the kriging model, the optimized design shows a spread of the measurements in both geographic and feature space (Heuvelink et al., 2006). For mapping using ML techniques with covariates, Brus (2019) recommends to select the sample using feature space coverage sampling (FSCS) or cLHS. Both cLHS and FSCS aim for an even sampling density in the multivariate feature space, but that is done in different ways. In cLHS it is done via the marginal distribution and correlation matrix of the covariates, in FSCS through minimization of a feature space distance criterion between sampling and predictions points using the k-means algorithm. This might be advantageous for ML techniques, which rely heavily on non-linear relations, but this has not yet been confirmed by experimental results. In machine learning, we do not have a model-based estimate of the prediction error variance. Optimizing the sampling design is not straightforward, although it is possible to optimize the design using a universal prediction accuracy measure (such as the mean square error (MSE) of the prediction). To the best of our knowledge, little has been investigated on optimal sampling design for mapping using random forest.

A relevant contribution was made in Tuia et al. (2013) which optimized the allocation of new climatological stations in a case study in Austria. In this study support vector regression and active learning were used to derive the optimal locations of new stations so as to select the most important sampling units to be included in the sample, i.e. units that are used as support vectors. However, active learning is a sequential re-design technique which is appropriate to improve an already-calibrated ML model. Tuia et al. (2013) provides little insight into where to select the sampling locations when there is no prior ML model. In consequence, the conclusions of this study are of little use for practitioners who wish to map soil properties using machine learning.

The objective of this study is to investigate what makes a design optimal (sample size and sampling locations) for mapping using RF. To achieve this, we (i) estimate the population MSE with various sampling designs (viz. simple random sampling, cLHS, spatial coverage sampling (SCS), feature space coverage sampling (FSCS) and MSE optimized); (ii) statistically test the differences in MSE distributions obtained with maps based on different designs; and (iii) reveal the characteristics of the optimal design by analysing the spread of the sampling locations in both geographic and feature space.

5.2 Materials and methods

5.2.1 Case study and data

We used the freely available soil database collected withing the framework of the European land use/cover area frame statistical survey (LUCAS) (Tóth et al., 2013). The LUCAS dataset is a sample of about 20,000 georeferenced topsoil (0-30 cm) measurements of thirteen soil properties spanning 23 European countries. The sampling density varies between 11 and 77 measurements per 10,000 km^2 with an average of 48. The sample was collected by a two-stage systematic sampling design (Gallego and Delincé, 2010) using a stratification based on seven land cover classes. The resulting sample is spread fairly uniformly in space and within the different land cover classes. A map of the sampling locations is provided in Orgiazzi et al. (2018, Figure 1a). We used as target soil property the soil organic matter (SOC) concentration in g kg^{-1} as measured by an automated vario MAX CN analyser (Elementar Analysensysteme GmbH, Germany) (Tóth et al., 2013). In this study, we treat the N LUCAS topsoil SOC measurements as our population of interest. This means that we ignore that the LUCAS units are a sample from the true area of interest, in our case the European countries included in the survey. The LUCAS topsoil SOC data are split randomly in three disjoint sets denoted calibration, validation and test sets. The purpose of each set is explained later in this section.

In addition to the LUCAS SOC sample, we used a set of 197 readily available continuous environmental variables at resolution 1 km × 1 km as covariates. The list of covariates is given in Hengl et al. (2017).

5.2.2 Random forest

Random forest (RF) is an ensemble machine learning method based on decision trees (Breiman, 2001). A single decision tree is built by repeating a binary recursive partitioning of the input training data. In the root node, the training data are grouped into a single partition. All possible binary partitions of the training data are evaluated using a splitting metric (Louppe, 2014). The binary split that has the smallest metric is selected. The newly created partitions undergo the same procedure, until a stopping criterion, the minimum node size, is met. The final prediction for continuous variable is taken as the average of the values of each the last decision tree node.

Breiman (1996) introduced the bagging technique. Bagging stands for bootstrap and aggregating, and aims at reducing the prediction error variance by building an ensemble of regression trees. A large number of trees is built based on bootstrap samples of the training data. All tree predictions are aggregated through averaging, and these averages are taken as the final predictions. The RF algorithm elaborates on this and introduces an additional random perturbation during the splitting of a tree (Breiman, 2001). In each split, the partitioning considers only a subset of size mtry from the original set of covariates.

The calibration of the RF model is therefore based on three user-defined parameters. The first is the number of trees ntree. To avoid computational load in fine-tuning ntree for each model, we fixed ntree = 200, as a compromise between accuracy and computational efficiency. Lopes (2015) showed that in many cases 150 trees is sufficient to obtain stable results, in particular when the number of covariates is smaller than the calibration sample size. The second parameter, mtry, is the number of covariates to randomly select at each split. By default, mtry is set to the rounded down square root of the total number of covariates. The third parameter is the minimal terminal node size (nodesize), which controls the minimum number of training data required to continue the process of tree growth. Parameter nodesize was set to its default value of 5.

5.2.3 Sampling designs

We compared five common spatial sampling designs.

Random: Simple random sampling (Cochran, 1977) of the soil property is the simplest form of sampling technique which does not require any prior knowledge on the soil property spatial variation. In simple random sampling, every unit of the population has equal probability of being selected and sampling units are selected independently. We used the sample function from the base package in the R language (R Core Team, 2018) for selecting simple random samples.

Spatial coverage sampling (SCS): A SCS design aims at dispersing the units throughout the study area as uniformly as possible (Royle and Nychka, 1998). Coverage designs are created by minimization of a criterion that is a function of the distance between sampling and prediction locations. Brus et al. (1999) proposed to compute the mean of the squared shortest distance (MSSD), denoted MSSD_G hereafter, between sampling locations and the centre cells of a fine prediction grid as criterion to obtain a spatial coverage design. This criterion can be minimized by the fast k-means clustering algorithm. We implemented it with the R base function kmeans, using the spatial coordinates of the whole study area as clustering variables. Since our population of interest is the LUCAS dataset, the selected sampling units are the LUCAS points closest (in geographic distance) to the centres of the geographic clusters.

Feature space coverage sampling (FSCS): A FSCS design follows the same principle as a spatial coverage design. However, in this case distances are measured in feature space instead of geographic space. Since covariates can have a very different scales, it is important to standardize them (zero mean and unit variance) so that the criterion to be minimized becomes the mean squared shortest standardized distance (MSSSD) (Brus, 2019), denoted MSSD_F hereafter. Sampling the centre of the k-means clusters ensures a uniform spread of the units in the multi-dimensional space of the covariates. We derive a FSCS design using the base R function kmeans. Similar to SCS, the LUCAS points closest (in standardized features space) to the centres of the clusters are used as sampling points.

Conditioned Latin Hypercube sampling (cLHS): cLHS (Minasny and McBratney, 2006) is a stratified random sampling procedure. For each covariate, n marginal strata are defined using the quantiles of the cumulative frequency distribution, with n being the sample size. Next, an optimization procedure minimizes the weighted sum of two components (O_1 and O_3) so that each covariate contains one unit per stratum in the multi-dimensional feature space (O_1) and the correlation between the covariate values in the sample and in the population is preserved (O_3). Note that we do not use component O_2 because our case study has no categorical covariates. In cLHS, the covariate marginal distribution of the sample is close to that of the population (Brus, 2019). Note that in this study cLHS designs were based on the 20 most important covariates for RF. These covariates were derived from a RF model calibrated using all LUCAS topsoil OC data (about 20,000 units). We refer to this design as the cLHS (20) design, and compare it to the FSCS optimized on the same 20 most important RF covariates (FSCS (20)). The most important covariates of the RF model are defined using the Gini impurity index (Nembrini et al., 2018) as implemented in the ranger package (Wright et al., 2017) in R. We used the R package clhs (Roudier, 2018) to obtain a cLHS sample from the population. The default implementation in the clhs package assigns equal weights to the O_1 and O_3 components.

MSE optimized: In this case an optimized design is obtained by minimization of the MSE between the predicted and measured SOC in the independent test set, from a RF model whose parameters are estimated using a calibration set. The choice of the minimization criterion is discussed more extensively in the Discussion. The minimization is achieved by spatial simulated annealing (SSA) (Van Groenigen and Stein, 1998; Wadoux et al., 2019a). For each iteration in SSA, a RF model is built, which is subsequently used to predict at the test set locations. If the MSE is smaller, the proposed sample is accepted, otherwise it is accepted with a probability that decreases during the optimization. We used the function optimUSER from the R package spsann (Samuel-Rosa, 2017). The total number of SSA iterations was set to 50 times the sample size.

Each of the five designs described above is evaluated by computing the MSE between the SOC prediction and observation for an independent validation set. The RF model is calibrated using the calibration set.

5.2.4 Estimation of the population MSE

We compute the population MSE for each design using a calibration sample of the LUCAS dataset. Several sources of randomness are involved when computing the population MSE, the first being the calibration sampling design: simple random sampling and cLHS are random designs, i.e. units are selected randomly. SCS and FSCS have some randomness due to their technical implementation (e.g. different initial solutions in k-means and convergence of the annealing schedule). The second source of randomness is due to the random split of the LUCAS dataset into calibration, test (for the MSE optimized design) and validation sets. Thus, by repeating the estimation of the MSE with a given sampling design and calibration/validation/test sets, we obtain a sampling distribution of the MSE.

The procedure for estimating the MSE sampling distribution, for a given sampling design and sample size $n = 100, 200, 500$ and 1000, is given as follows:

```
for $r = 1$ to $R$ do
    Split the LUCAS dataset fully randomly into $K$ disjoint subsets of equal
      size (validation subsets);
    for $k = 1$ to $K$ do
        Define the $k$-th subset as the validation subset. Merge the remaining
          $K - 1$ subsets and split the merged set fully randomly into $L$ disjoint
          subsets of equal size (test subsets);
        for $l = 1$ to $L$ do
            Define the $l$-th subset as the test dataset. Merge the remaining
              $L - 1$ subsets;
            for $m = 1$ to $M$ do
                    1. Select a sample of size $n$ from the merged $L - 1$ subsets.
                       This is the calibration dataset. In case of the MSE optimized
                       sampling design, the sample is selected such that it minimizes
                       the MSE of the test dataset. In case of the other four designs
                       the test dataset is not used but a sample is selected according
                       to the criterion of the design (i.e. simple random sampling,
                       SCS, FSCS, cLHS).
                    2. Calibrate the RF model using the sample selected by the
                       design.
                    3. Predict at the locations of the validation dataset and compute
                       the squared prediction errors for all validation locations.
            end
            Average the $M$ squared prediction errors at each validation
              location.
        end
        Average the $L$ averaged squared prediction errors at each validation
          location.
    end
    Average the final averaged squared squared prediction errors over all
      LUCAS locations, the outcome is a single estimate of the population
      MSE value.
end
Plot the distribution and print summary statistics of the $R$ estimates of the
  population MSE.
```

Hereafter, the distribution of the R estimates of the population MSE as obtained using this procedure for a given design and calibration sample size n is referred to as the “MSE sampling distribution”. The values of R, K, L and M are chosen based on the computational load and degree of randomness of each designs. When the design is more random, larger values of R, K, L and M are required. We chose $R = 10$ for all designs except for the MSE optimized design, where we used $R = 5$. K and L are set to 5 for all designs while $M = 20$ for the random and SCS designs, $M = 10$ for the FSCS design and $M = 1$ for the cLHS and MSE optimized designs.

5.2.5 Statistical hypothesis testing

Given the MSE sampling distributions for each design, we tested for all pairs of designs and all calibration sample sizes n whether the medians of the distributions are significantly different using the Mann-Whitney U test (Wilcoxon rank-sum test) (Mann and Whitney, 1947). The Mann-Whitney U test is a non-parametric test of the null-hypothesis that two distributions have the same median. Thus, under the null hypothesis a randomly selected value from one of the distributions has 50% chance of being smaller or greater than a randomly selected value from the other distribution. Contrary to the two independent samples t-test the Mann-Whitney U test does not require the normality assumption of the distributions that are compared. Significant differences between MSE sampling distributions are characterized by a significance threshold fixed at a p-value smaller or equal than 0.05.

5.2.6 Diagnostics of the designs

Sampling designs are not only evaluated by the resulting MSE, but also by the spread of the samples in the geographic and feature space. This is done with the aim to reveal the characteristics of the designs, in particular the MSE optimized design, which may help to design future surveys. Thus, all sampling designs are evaluated in terms of all criteria, not just MSE, but also MSSD_G, MSSD_F and O_1+O_3 as minimized in cLHS.

5.3 Results

5.3.1 MSE sampling distribution

Figure 5.1 shows the boxplot of the MSE values for all combinations of sampling design and sample sizes. As expected, the MSE optimized design is always more accurate than the other designs. This is particularly true for small sample sizes (e.g. 100 units) where the MSE optimized design has an MSE that is about 10% smaller than that of a simple random sampling design. For small sample sizes (e.g. 100), simple random sampling and cLHS have the largest MSE (median is 7208 and 7174 (g kg^{-1})^2 respectively) and FSCS has a somewhat smaller MSE (median is 7090 (g kg^{-1})^2). This pattern is preserved with larger sample sizes, but the differences in MSEs become negligible as the sample size increases. For instance, the difference in MSE between designs is smaller than 100 (g kg^{-1})^2 for a sample size of 1000. Note that with cLHS for all sample sizes tested, the sampling distribution of MSE not only has a large median value, about equal to that of simple random sampling, but also shows a large variability. For example, the standard deviation of the MSE distribution for the cLHS design and a sample size of 100 is 75 (g kg^{-1})^2 while it is only 13 (g kg^{-1})^2 for a simple random sampling design of the same size.

5.3.2 Statistical hypothesis testing

Table 5.1 shows the result of the statistical hypothesis testing. Sampling designs with median MSE that are not significantly different at $\alpha = 0.05$ have the same letter. The median MSE of the cLHS design is, for all sample sizes tested, not significantly different from the median MSE of the simple random design. In contrast, the median MSE based on the MSE optimized design is always significantly different from those of other designs. This is an expected result given that the corresponding MSE distributions shown in Fig. 5.1 do not overlap (but note that the boxplots shown in Fig. 5.1 are based on small samples of size 5 and 10 only). For sample size 100, the median MSE of the cLHS design is not significantly different from that of the SCS design, and the median MSE between FSCS designs using all or the 20 most important covariates are not significantly different. For sample size 500, the median MSE of the SCS design is not significantly different from that of the FSCS design using all covariates. Overall, it appears that parameters M and R were large enough to detect significant differences between designs.

5.3.3 Diagnostics of the designs

Figure 5.2 shows the distribution of the MSSD$_\mathrm{G}$ for all designs and sample sizes. Because the SCS design is optimized for this criterion it has always the smallest median MSSD$_\mathrm{G}$ compared to other designs, for the same sample size. FSCS designs (optimized on all or the 20 most important covariates) have relatively small MSSD$_\mathrm{G}$ values. This may be because the spatial coordinates are also included as covariates and hence used to optimize these designs. The simple random and MSE optimized designs have the largest MSSD$_\mathrm{G}$ values and also the largest MSSD$_\mathrm{G}$ variability (standard deviation of 7.28^9 and 1.51^{10} m^2 for a sample size of 100, respectively). The MSE optimized design has on average the least uniform spread in geographic space, as

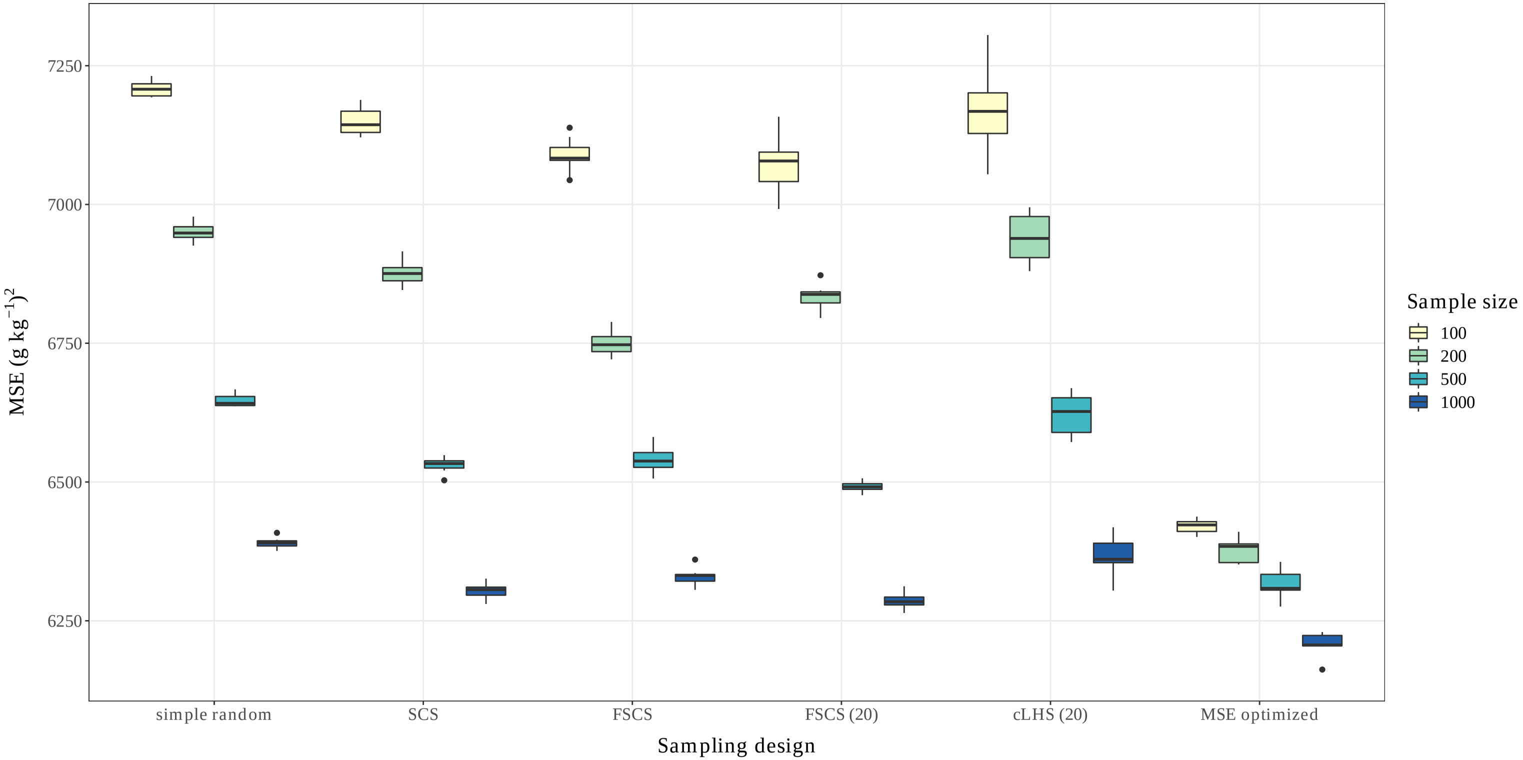


Figure 5.1 *– Boxplot of MSE values derived from the experimental design of Section 5.2.4, for each of the tested sampling designs and for different sample sizes. FSCS (20) and cLHS (20) refer to designs computed on the 20 most important covariates for the RF model, calibrated using all LUCAS topsoil OC data (about 20,000 units).*

***Table 5.1** – Mann-Whitney U test results for differences in median MSE obtained with random forest models calibrated with samples of various designs and sample sizes. Common letters indicate non-significant differences at significance level α of 0.05.*

	Sample size			
	100	200	500	1000
Simple random	a	a	a	a
cLHS (20)	a b	a	a	a
SCS	b	b	b	b
FSCS	c	c	b	c
FSCS (20)	c	d	c	d
MSE optimized	d	e	d	e

shown by the median MSSD_G. This is the case for all sample sizes, even though the differences in MSSD_G among designs are negligible for large sample sizes.

Figures 5.3 and 5.4 show the MSSD_F distributions, computed using all the covariates or a subset containing the 10 most important covariates of the RF models, respectively. Both figures show, as expected, that FSCS designs have the smallest MSSD_F compared to other designs. All other sampling designs have similar MSSD_F distributions. A similar pattern is observed in Fig. 5.4: all designs (except FSCS designs) have similar MSSD_F distributions. Note that simple random and SCS designs have a very large spread in the MSSD_F, while the MSE optimized design has narrower MSSD_F distributions and very few outliers.

Figure 5.5 shows the distribution of the $O_1 + O_3$ cLHS criterion computed for each of the designs and sample sizes. Note that in Fig. 5.5 the elements $O_1 + O_3$ are not computed as sums but as means. This is discussed more extensively in the Discussion. Since a cLHS design is optimized for this criterion, it has the smallest values for all sample sizes. For large sample size, the simple random sampling design is almost equivalent in terms of the cLHS criterion. The FSCS designs (using all or the 20 most important covariates) have always the largest value of the cLHS criterion.

Figure 5.6a indicates how often a sampling location is selected in the MSE optimized design, where red colours correspond to a case where a sampling location is selected more often than would be expected under a simple random sampling design and blue colours indicate the opposite. Fig. 5.6b shows the proportion of sampling locations used for calibration of the MSE optimized design. Red colours indicate that locally, a relatively large number of sites were used for calibration, while blue colours indicate that relatively few sites in the local neighbourhood were used for calibration of the MSE optimized design. Areas with fewer than five LUCAS sites within the

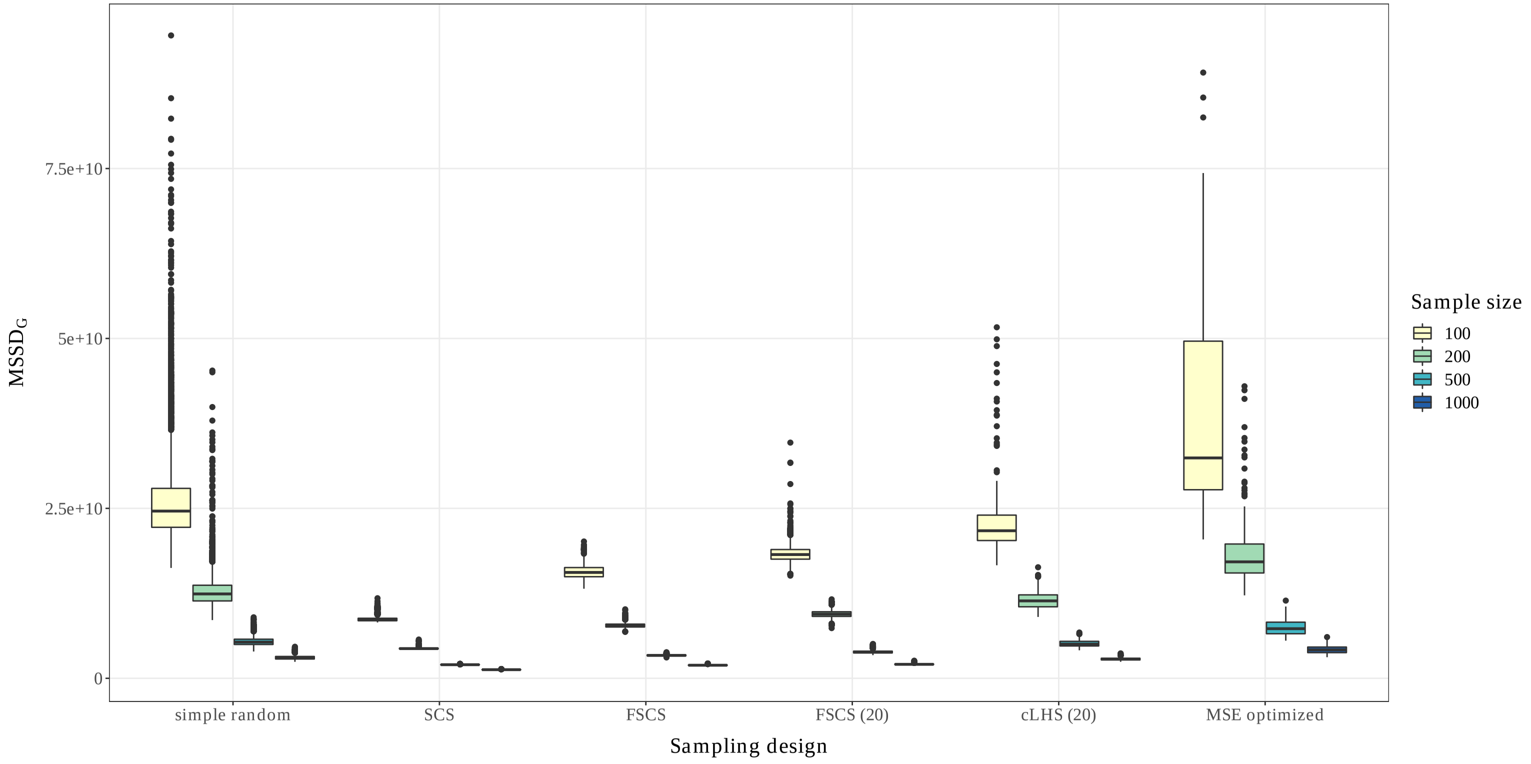


Figure 5.2 *– Boxplots of MSSD$_G$ for all sampling designs and sample sizes.*

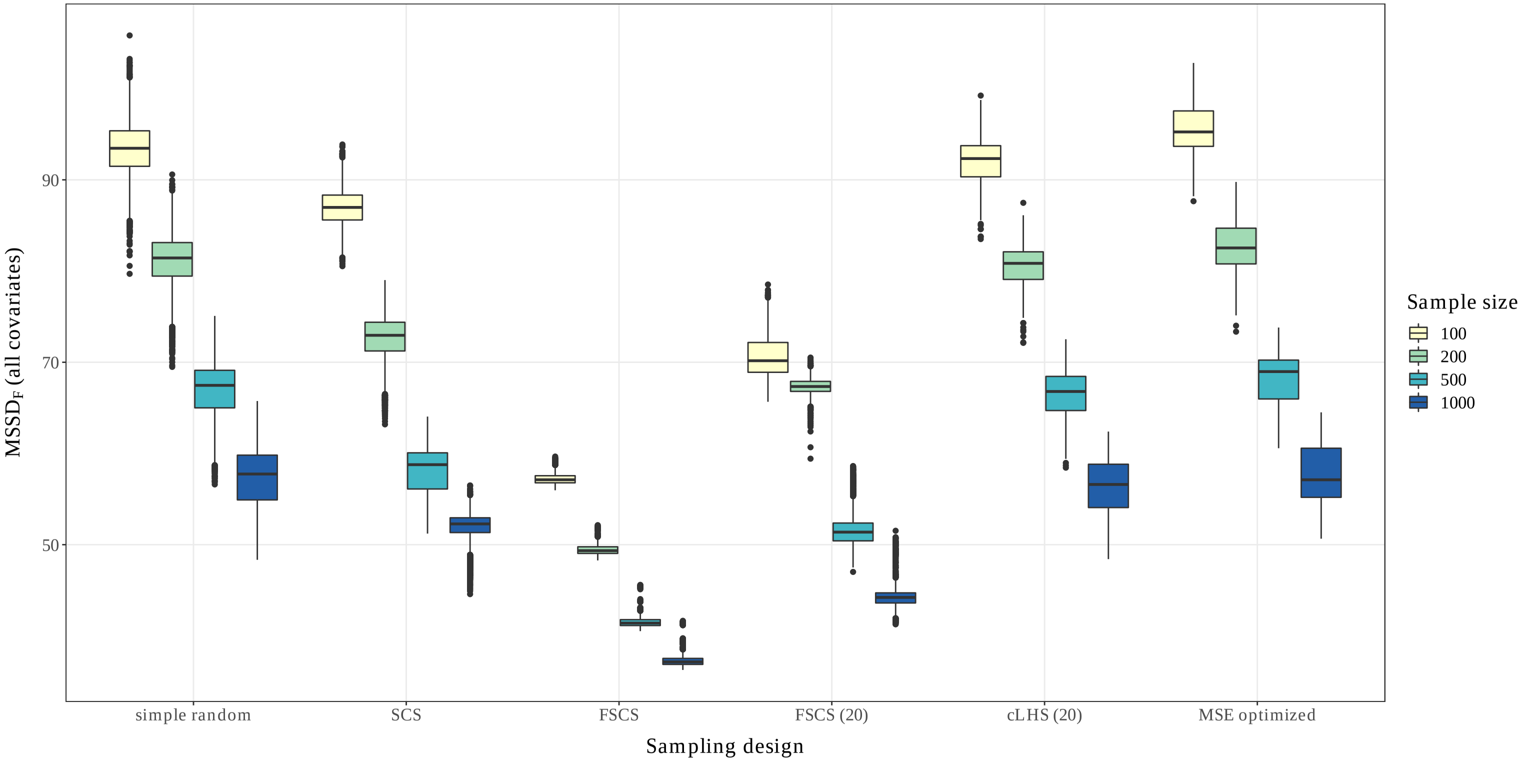


Figure 5.3 – *Boxplots of the MSSD$_F$ for different sample sizes and sampling designs.*

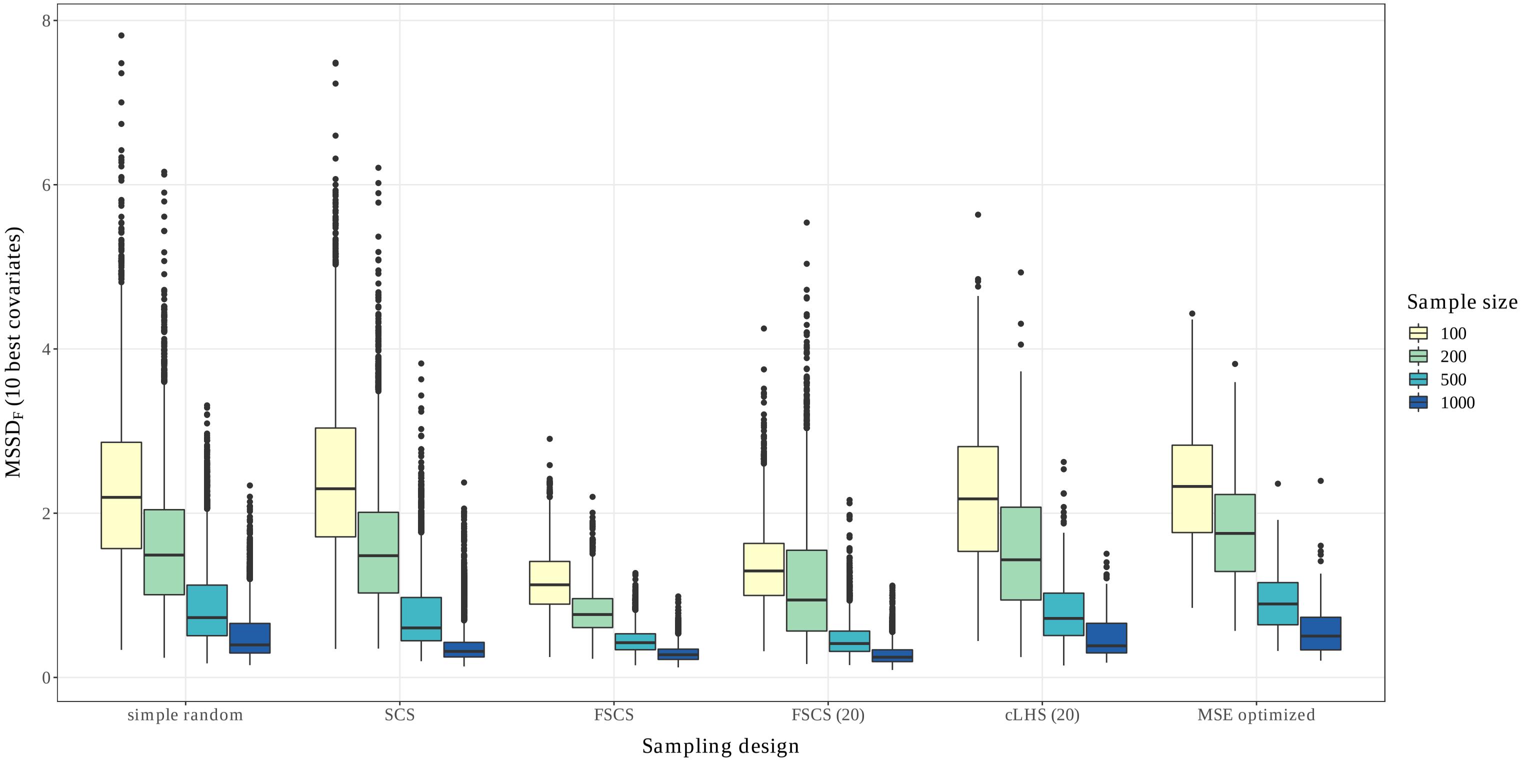


Figure 5.4 – *Boxplots of the MSSD$_F$, based on the ten most important covariates of each design.*

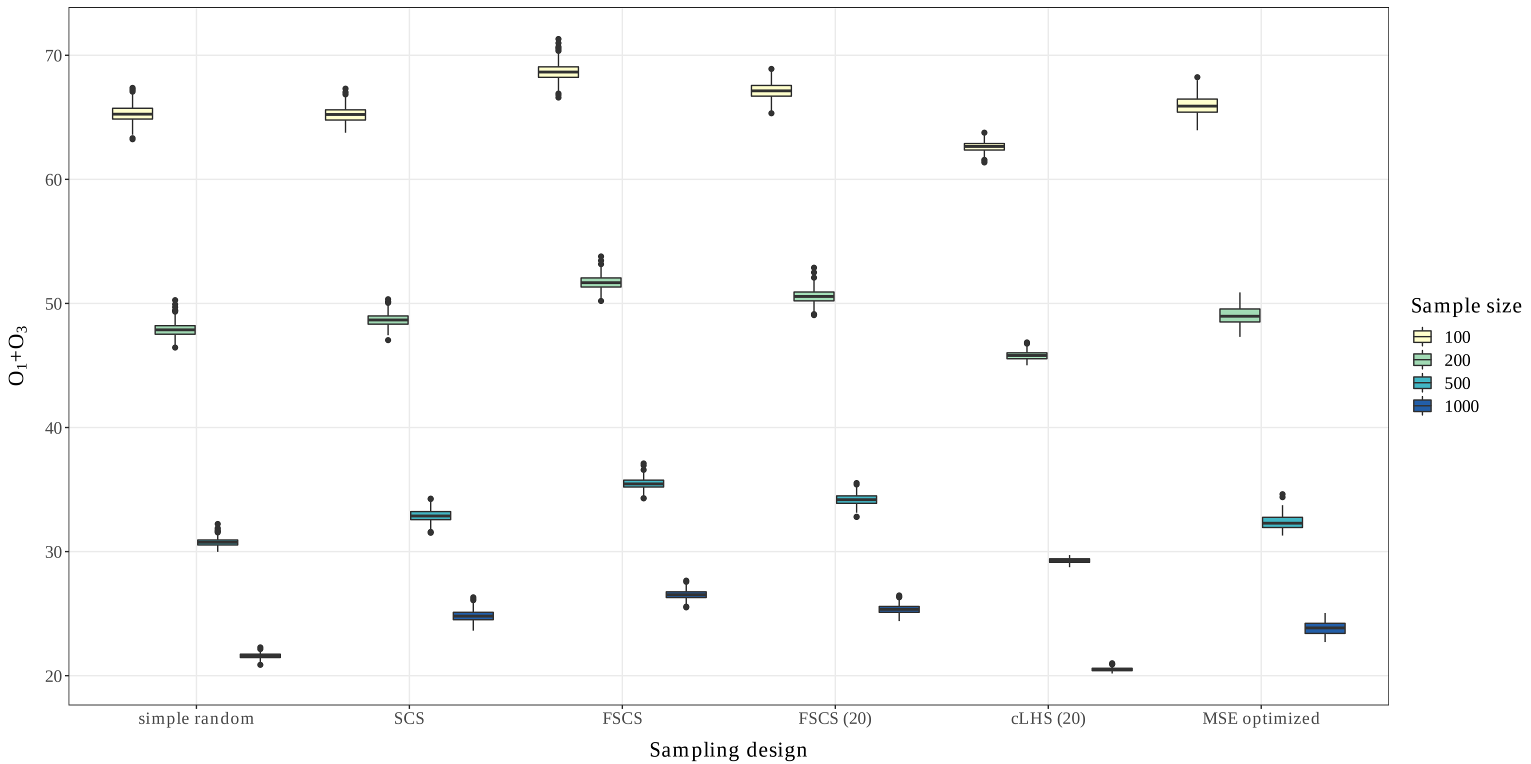


Figure 5.5 – *Boxplots of the O_1 + O_3 cLHS criterion for each of the sampling designs and sample sizes. The elements O_1 + O_3 are computed as means.*

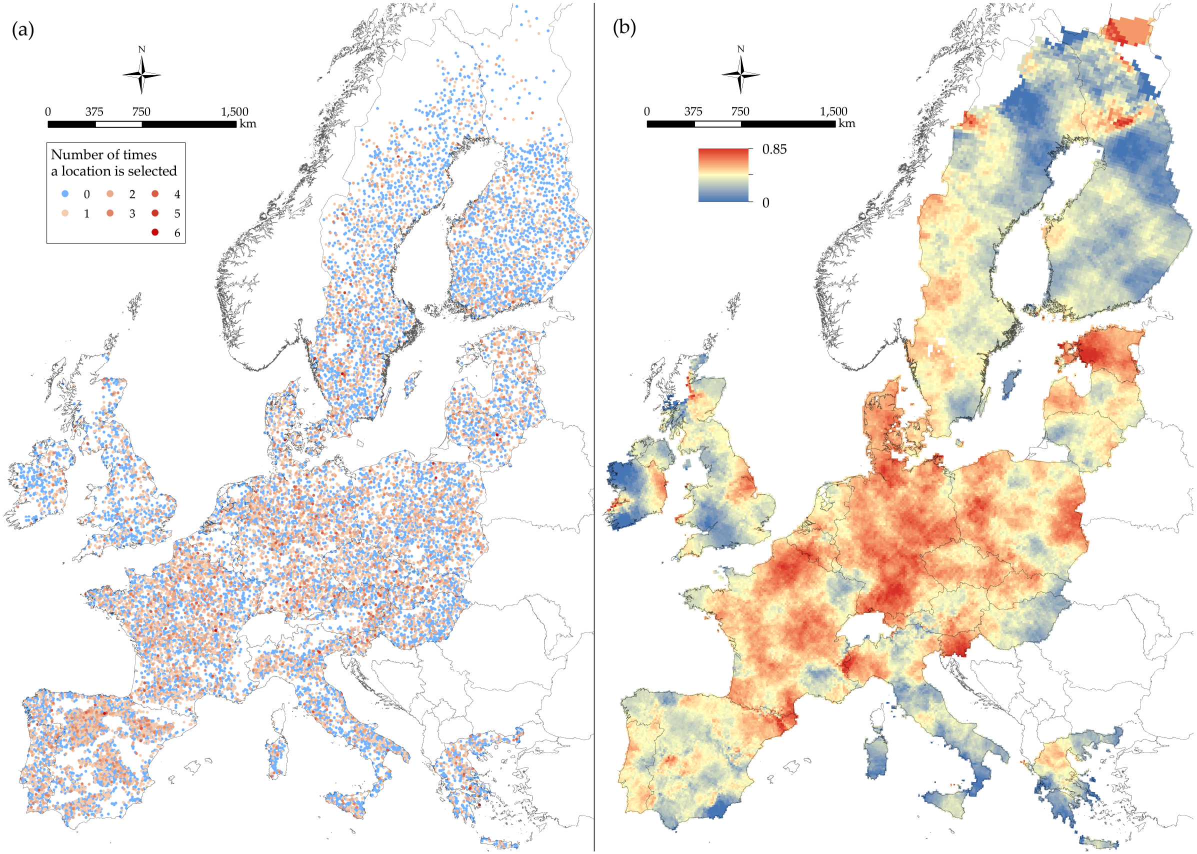


Figure 5.6 – *Number of times a sampling location is selected by the MSE optimized design (a). Red colours indicate that the location is selected more often than one would expect under a simple random sampling design, blue colours indicate that it is selected less often than expected under simple random sampling. Ratio of number of sampling sites used for calibration of the MSE optimized model and total number of sampling sites, as computed in a circular neighbourhood with radius 100 km (b). Red colours indicate regions for which sampling units are often included in the MSE optimized design, blue colours refer to regions for which sampling units are less often included in the MSE optimized design.*

local neighbourhood (100 km circular radius) were masked out. Fig. 5.6b shows that the MSE optimized design leads to a fairly high relative density in a geographic band spanning from France to Poland. Germany and Denmark have a high relative density across their entire country. Great Britain, Ireland, southern and northern Europe tend to have a lower relative density of sampling units included in the MSE optimized design, even though they might locally have a very high *absolute* density of sampling locations (e.g. North of Madrid).

5.4 Discussion

Impact of sampling designs on prediction accuracy

The sampling design had a significant impact on the accuracy of random forest predictions. In the case study mapping topsoil OC using RF in Europe, the MSE optimized design had the smallest mean squared prediction error, as shown in Fig. 5.1. This is because the MSE optimized design was optimized for this purpose, by minimizing the MSE of the test set. All other designs reach substantially higher MSE value than the MSE optimized design. However, the MSE optimized design can be used only when subsampling an existing dataset with known values of the target soil property at all locations. In other words, it may be used in a case where thinning of an existing sampling network is required, but not in a case where one needs to design a sampling scheme from scratch, such as in a reconnaissance survey. In this case, it is best to use a FSCS design which, for the case study, had the smallest prediction MSE of all other designs tested. This is not surprising because predictions made by machine learning methods rely on non-linear relationships with covariates, and estimation of these relationships benefits from a spread of the sampling units in feature space, as noted by Brus (2019). To our surprise the MSE values obtained with cLHS design were large (Fig. 5.1) and not statistically different from those obtained using simple random sampling (Table 5.1). This is discussed more extensively later in this Discussion. In spite of the differences in MSE between designs for small sample sizes, the MSE between designs for large sample sizes (in our case study larger than 1 unit per $4,159$ km^2) are negligible. This result applies to our case study using the LUCAS dataset as the population of interest, but is likely also valid more generally: increasing the sample size reduces the MSE differences between designs because the selected sample covers all cases sufficiently well.

Design diagnostics

In practice we cannot obtain an MSE optimized design because this requires that the target property is known at all locations in the study area. This is why it is useful to interpret and diagnose the MSE optimized designs obtained in the case study, because if general patterns can be derived then these may be used to design spatial sampling designs for DSM using RF. Diagnostics on the MSE optimized designs reveal that RF does not benefit much from a spread of the sampling units in geographic space (Fig. 5.2). One possible reason is that spatial location is ignored during the RF modelling process (Hengl et al., 2018) and in other machine learning techniques (Behrens et al., 2018b). Fig. 5.3 and 5.5 show that, in addition, RF neither benefits much from a spread of the sampling units in the feature (i.e. covariate) space, nor from reproducing the marginal distributions of the covariates. This is unexpected because many studies (e.g. Castro-Franco et al., 2015; Domenech et al., 2017; Brus, 2019) suggested that spread in the feature space is crucial. In fact, it is more subtle than that. We learn from Fig. 5.4 that the importance of the covariates used in the RF model must be taken into account as well. This is an important finding of this study: the predictions made by a RF model benefit from a design spread uniformly in the space spanned by the most important covariates. We acknowledge that this finding is based on a single case study and needs to be tested in further research. If this finding is confirmed by future studies, one can derive practical recommendations to design a soil survey for mapping with RF: (i) determine what are the most important covariates, either using a legacy sample, previous studies, pedological expertise or a two-stage sampling approach; and (ii) optimize the design using coverage sampling in covariate space for the important covariates (possibly using weights derived from the importance).

Conditioned Latin Hypercube sampling design

While it was shown above that RF benefits from a uniform spread of the sampling locations in the feature space of the most important covariates, predictions based on the cLHS design were as accurate as those based on a simple random sampling design and worse than predictions obtained using all other designs. Apparently, sampling the marginal distribution of the covariates is not a useful strategy for mapping with RF. The criterion values in Fig. 5.3 and 5.5 show that the cLHS and FSCS designs are very different in the way they spread the sampling units in feature space. We showed in this study that this has a major impact on the resulting prediction accuracy, and that cLHS sampling is not recommended for mapping using RF. Note that in this study we used the cLHS implementation from the R package of Roudier

(2018) following the Minasny and McBratney (2006) paper, where the O_1 and O_3 components are computed as sums, not as means. The resulting criterion is therefore affected by the magnitude of the O_1 and O_3 components. To solve this problem, other implementations (e.g. Samuel-Rosa, 2017) compute O_1 as the mean of the absolute deviations between the marginal strata sample size and targeted sample size, while O_3 is computed as the mean of the deviations over all off-diagonal entries of the correlation matrix. Taking the latter into account will potentially have an impact on the performance of the cLHS design. However, we did not consider it in this study.

Optimization criteria

In our case study, the MSE optimized design was derived based on the MSE between predicted and measured SOC values in the test dataset. The MSE is a universal criterion which can be computed for any mapping method, also in a case where we do not have a model-based estimate of the prediction error variance. If a model-based estimate of the prediction error variance is available, we can use a function of the prediction error variance as minimization criterion. Obvious candidates for such function are the spatial mean (Brus and Heuvelink, 2007) and maximum (Van Groenigen et al., 1999) prediction error variance. For the RF model used in the case study, the prediction error can be quantified by quantile regression forest (QRF) (Meinshausen, 2006), for instance using the width of the 90% prediction interval. We explored this and used the average width of the QRF 90% prediction interval over the study area (i.e. the 23 EU countries included in this study) as a minimization criterion. However, we observed that the sampling units of the optimized design had a narrow SOC distribution and small SOC variance. These sampling units were selected because this resulted in narrow QRF predicting intervals and hence a small criterion value. As a result, validation of the quantified uncertainty (e.g. using accuracy plots Deutsch, 1997; Wadoux et al., 2018) showed that the uncertainty was systematically and severely underestimated. Thus, we did not pursue this any further.

Sampling for other machine learning techniques

Finally, there is a need to further investigate whether a design that is optimal for RF modelling is also optimal for other machine learning models. Our results were obtained for a tree-based model. We hypothesize that a design that is optimal for RF may also be efficient for modelling and predicting using other tree-based models (e.g. CART Breiman, 2017), because they are comparable in their basic structure and splitting metrics. Sampling to support other machine learning models (e.g. support vector machine or deep neural network) introduces additional considerations and deserves further investigation. For example, Pozdnoukhov and Kanevski (2006) and Tuia et al. (2013) optimized a network for mapping using support vector machine. They specifically aimed at minimizing the "risk" of selecting new sampling units that do not have a valuable contribution to the model (by becoming support vectors). Recently, Wadoux (2019) showed how a deep neural network can be used for soil mapping, and how the minimized loss function can be modified to include additional information (e.g. to quantify the prediction uncertainty). Formulating a loss function that searches for optimal units to be measured using the feature (i.e. covariates) space has been tackled by MacKay (1992). How much a design optimal for a neural network model would differ from that of a RF model requires further study. This would certainly make a valuable contribution to future DSM studies.

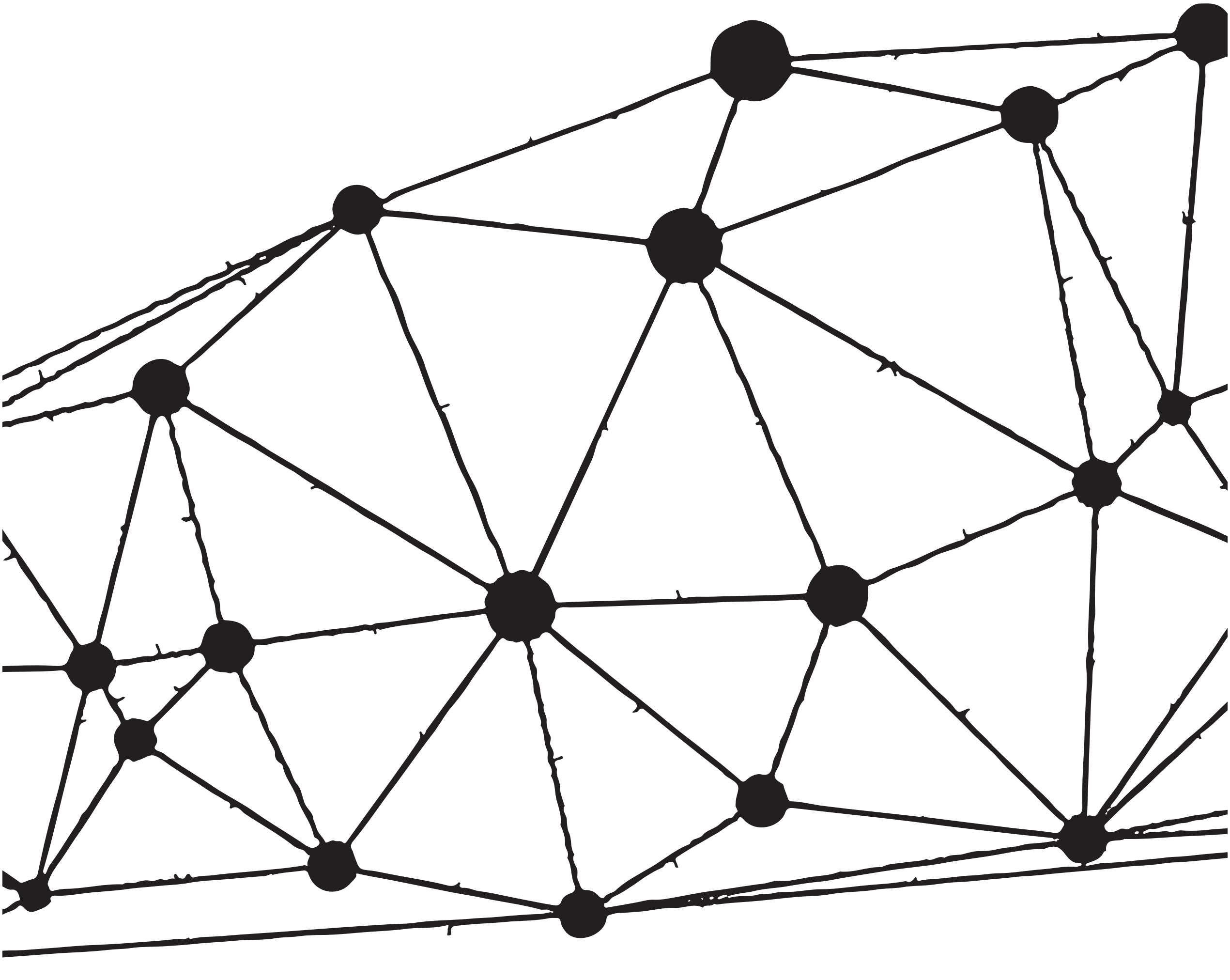
5.5 Conclusion

We computed an MSE optimized design for mapping with RF and compared it to several commonly used sampling designs. We compared the designs in terms of both prediction accuracy and spread of sampling units in geographic and feature space. In a case study, we used the LUCAS topsoil OC measurements as our population of interest, from which subsamples were collected. From the results and discussion we draw the following conclusions:

- An MSE optimized design provides the smallest mean squared prediction error. However this is feasible only in case of subsampling an existing dataset with known values of the target soil property at all locations.
- In terms of accuracy, a sample selected by feature space coverage sampling of the most important covariates had the closest match with the MSE optimized design.
- For large sample sizes, the differences between prediction accuracies of different designs become negligible. In our continental scale case study, this was for a sampling density greater than 1 sampling unit per about 4,000 km^2.
- A conditioned Latin Hypercube sampling (cLHS) design is not a good choice for mapping using RF. In our case study, predictions based on a cLH sample had the poorest prediction accuracy, similar to that of predictions based on a simple random sample.
- Diagnostics on the MSE optimized design showed that for RF the optimal sampling design is not achieved by a uniform spread of the sampling units in the

geographic and/or feature (i.e. covariate) space, nor from reproducing the marginal distributions of the whole set of covariates.

- Further diagnostics of the MSE optimized design showed that the importance of the covariates used in the RF model must be taken into account when optimizing the spatial sampling design. RF benefits from a spread of the sampling units uniformly in the feature space of the most important covariates of the RF model. The most important covariates can be selected using a sample from a reconnaissance survey, by pedological expertise or by a two-stage sampling strategy.



Chapter 6

Optimization of rain gauge sampling density for discharge prediction using Bayesian calibration

Stream discharges are often predicted based on a calibrated rainfall-runoff model. The major sources of uncertainty, namely input, parameter and model structural uncertainty must all be taken into account to obtain realistic estimates of the accuracy of discharge predictions. Over the past years, Bayesian calibration has emerged as a suitable method for quantifying uncertainty in model parameters and model structure, where the latter is usually modelled by an additive or multiplicative stochastic term. Recently, much work has also been done to include input uncertainty in the Bayesian framework. However, the use of geostatistical methods for characterizing the prior distribution of the catchment rainfall is underexplored, particularly in combination with assessments of the influence of increasing or decreasing rain gauge network density on discharge prediction accuracy. In this paper we integrate geostatistics and Bayesian calibration to analyse the effect of rain gauge density on discharge prediction accuracy. We calibrated the HBV hydrological model while accounting for input, initial state, model parameter and model structural uncertainty, and also taking uncertainties in the discharge measurements into account. Results for the Thur basin in Switzerland showed that model parameter uncertainty was the main contributor to the joint posterior uncertainty. We also showed that a low rain gauge density is enough for the Bayesian calibration, and that increasing the number of rain gauges improved model's prediction until reaching a density of one gauge per 340 km^2*. Based on the results, we make recommendations on how to handle input uncertainty in Bayesian calibration for discharge prediction.*

Based on:
Wadoux, A. M. J.-C., Heuvelink, G. B. M., Uijlenhoet, R. and de Bruin, S. (2019).
Optimization of rain gauge sampling density for discharge prediction using Bayesian
calibration. *Water Resources Research.* Accepted with revision.

6.1 Introduction

Uncertainty analysis has garnered considerable attention in hydrological modelling during the past decades (e.g. Pappenberger and Beven, 2006; Han and Coulibaly, 2017). There is agreement on the necessity to provide (realistic) uncertainty bounds to end-users and practitioners (Beven, 2006; Andréassian et al., 2007). Brown and Heuvelink (2005) define uncertainty as an expression of confidence about how well we know the "truth". Similarly, Maskey (2004) defines uncertainty as a measure of the information about an unknown quantity to be measured or a situation to be forecast, and discusses the nature of different potential sources of uncertainty and their effect on flood forecasting.

It is generally recognized that three principal sources of uncertainty cause uncertainty in model output: model input uncertainty (including initial state and boundary conditions), model parameter uncertainty and model structural uncertainty (Refsgaard et al., 2007; Van der Keur et al., 2008). For example, Højberg and Refsgaard (2005) found that parameter uncertainty cannot fully cover model structural uncertainty, while Tian et al. (2014) showed that parameter uncertainty for three rainfall-runoff models tested on two catchments has a larger contribution to model output uncertainty than model structural uncertainty. Kavetski et al. (2006) found that input (rainfall) uncertainty has a considerable effect on the predicted outflow and output prediction intervals. In addition to these three main sources, there is usually also uncertainty in the measurements of the model output (Di Baldassarre and Montanari, 2009). This source of uncertainty must be taken into account if these measurements are used to calibrate the model.

Explicit integration of all sources of uncertainty is not an easy task. This problem has been tackled using approaches such as the pseudo-Bayesian generalized likelihood uncertainty estimation (GLUE) methodology (Beven et al., 2000), the integrated Bayesian uncertainty estimator (IBUNE) (Ajami et al., 2007) and by using Bayesian total error analysis (BATEA) (Kavetski et al., 2006). In general, Bayesian analysis has received wide attention because it provides a comprehensive and general framework to specify uncertainty explicitly using probability distributions. It also fosters easy updating of distributions when additional information comes available. The main steps of a Bayesian uncertainty framework are summarized as follows (Kennedy and O'Hagan, 2001): (1) an explicit probability model is specified for each uncertainty source (input, model parameters, model structure), based on prior information; (2) measurements of the model output are used to update the prior distributions to posterior distributions, typically using Markov chain Monte Carlo techniques; (3) the posterior distributions are used to propagate uncertainty in model input, model parameters and model structure to model output for (future) cases where model output is not observed; (4) results are tested against independent validation data to evaluate whether the assumptions made as part of the procedure are realistic.

Numerous studies on Bayesian uncertainty analysis for distributed and physically based hydrological models have been conducted and published. There is general agreement that rainfall-runoff data are often insufficient for supporting reliable inference for complex models involving many spatially distributed physical catchment processes (Beven, 2006; Renard et al., 2010). Wagener et al. (2001) refers to "non-identifiability" leading to "ill-posed" inference of the parameters, which can be avoided by using simpler (lumped) hydrological models with fewer parameters. Lumped hydrological models consider the quantity of interest (e.g. discharge) to be derived from catchment-averaged inputs (e.g. rainfall, potential evapotranspiration). Model inputs often contain substantial error which affect model output. In general, input uncertainty in lumped models is mainly caused by measurement and interpolation errors. For instance, rainfall measurements obtained using rain gauges are not error-free (Habib et al., 2001), while interpolation error is added when rain gauge measurements at point locations are aggregated to spatial averages as needed in lumped rainfall-runoff models. In the case of rainfall, radar images provide time series of spatial rainfall fields, thus avoiding interpolation error, but these often suffer from complex spatio-temporal errors which make them inaccurate in some circumstances (Cecinati et al., 2017). Thus, rainfall point observations remain a major source for estimating catchment-average rainfall.

There is a recent trend towards a decrease of hydrometric network density (Mishra and Coulibaly, 2009; Keum and Kaluarachchi, 2015). Yet, the uncertainty in average rainfall strongly depends on rain gauge sampling density (Xu et al., 2013; Terink et al., 2018). Hence, a reduction of the rain gauge density will increase the uncertainty about the discharge predicted by the rainfall-runoff model. Renard et al. (2011) used a geostatistical model to infer the catchment-average rainfall and the associated uncertainty from the rain gauges using block kriging. Next, they used the block kriging conditional distribution as a prior in the Bayesian calibration of a lumped rainfall-runoff model. Taking prior knowledge on input uncertainty into account overcomes ill-posedness and significantly improved the accuracy of the runoff predictions (Renard et al., 2010). Clearly, the higher the rain gauge density the narrower the block kriging prior. Thus, a different sampling density leads to a different prior and posterior and ultimately to a different output uncertainty distribution. To the best of our knowledge, little has been done to investigate the effect of rain gauge density on the model output uncertainty within a Bayesian framework.

The objective of this work is to evaluate the effect of rain gauge sampling density on uncertainty in the output of a lumped rainfall-runoff model. The methodology used relies on geostatistics to quantify prior input uncertainty and on Bayesian calibration for model parameter and model structural uncertainty quantification. We calibrate the lumped HBV model (Lindström et al., 1997) using a Bayesian uncertainty framework that accounts for input, parameter, output observation, initial state and model structural uncertainty. Model residuals comprising model structural uncertainty and discharge error are characterized using a time series model, while Markov chain Monte Carlo methods are used to obtain posteriors of the input, model and initial state parameters. The propagation of uncertainties associated with the model input, model parameters and model structure is then analysed using regular Monte Carlo methods. Several rain gauge density scenarios were tested, each time recalibrating the model and providing discharge predictions with uncertainty intervals. Rainfall posterior intervals as well as model predictive abilities were assessed and discussed. The approach was tested in a case study using ten-day average rainfall and discharge data of the 1696 km^2 Thur basin in Switzerland for the years 2004 to 2011.

6.2 Methods

6.2.1 Rainfall-runoff model

Consider a hydrological model H that predicts stream discharge from catchment average rainfall. Let $\mathbf{y} = [y_1\ y_2 \ldots y_T]^T$ and $\overline{\mathbf{z}} = [\overline{z}_1\ \overline{z}_2 \ldots \overline{z}_T]^T$ be time series of measured discharge and (known) catchment average rainfall, respectively. We assume that the relation between $\mathbf{y}$ and $\overline{\mathbf{z}}$, which is governed by the model H, is affected by multiplicative measurement and model structural errors, which after log-transformation gives:

$$\log(\mathbf{y}) = \log(H(\overline{\mathbf{z}}, \varphi)) + \boldsymbol{\varepsilon} + \boldsymbol{\eta} \tag{6.1}$$

where φ is a vector comprising model parameters and the initial state, $\boldsymbol{\varepsilon} = [\varepsilon_1\ \varepsilon_2 \ldots \varepsilon_T]^T$ is log-transformed model structural uncertainty and $\boldsymbol{\eta} = [\eta_1\ \eta_2 \ldots \eta_T]^T$ is log-transformed discharge measurement error. Uncertainty in model input (i.e. $\overline{\mathbf{z}}$) and model parameters (i.e. φ) will be introduced in the next section. We assume that the η_t ($t = 1 \ldots T$) are independent and identically distributed normal variates, with zero mean and constant variance σ_η^2.

It is unrealistic to assume temporal independence for $\boldsymbol{\varepsilon}$ and hence we represent it by a first-order autoregressive model (AR(1)):

$$\varepsilon_t = \beta_0 + \beta_1 \cdot \varepsilon_{t-1} + \delta_t, \quad t = 1 \ldots T \qquad \varepsilon_0 \sim \mathcal{N}(\mu_0, \sigma_0^2) \tag{6.2}$$

where the δ_t ($t = 1 \ldots T$) are independent and identically distributed normal variates, with zero mean and constant variance σ_δ^2. The parameters that characterize $\boldsymbol{\varepsilon}$ are merged into a parameter vector $\theta = \{\beta_0, \beta_1, \sigma_\delta^2, \mu_0, \sigma_0^2\}$. In this study, μ_0 and σ_0^2 are assumed known since their effect is typically negligible after a few time steps. We also assume that $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$ are mutually independent.

To simplify notation we define $\mathbf{u} = \log(\mathbf{y}) - \log(H(\bar{\mathbf{z}}, \varphi))$ and obtain:

$$\mathbf{u} = \boldsymbol{\varepsilon} + \boldsymbol{\eta} \tag{6.3}$$

6.2.2 Bayesian uncertainty framework

Conventional estimation of parameters φ, θ and σ_η^2 using Bayesian calibration (Beven and Freer, 2001; Kavetski et al., 2006) starts by using Bayes' law to derive that the posterior distribution of the parameters is proportional to the product of the likelihood of the observed residuals $\mathbf{u}$ and the prior distribution of the parameters:

$$p(\varphi, \theta, \sigma_\eta^2 | \mathbf{u}) = \frac{p(\mathbf{u}|\varphi, \theta, \sigma_\eta^2) \cdot p(\varphi, \theta, \sigma_\eta^2)}{p(\mathbf{u})} \propto p(\mathbf{u}|\varphi, \theta, \sigma_\eta^2) \cdot p(\varphi, \theta, \sigma_\eta^2) \tag{6.4}$$

where $p(\varphi, \theta, \sigma_\eta^2)$ is a joint prior for the rainfall-runoff model parameters and the model structural and measurement uncertainty parameters. We further assume independence between the priors, i.e. $p(\varphi, \theta, \sigma_\eta^2) = p(\varphi) \cdot p(\theta) \cdot p(\sigma_\eta^2)$.

We extend the conventional approach by considering the case where the model input (i.e. the catchment-average rainfall) is also uncertain. Since model output and input are not independent (as is obvious from Eq. 6.1) we have $p(\bar{\mathbf{z}}|\mathbf{y}) \neq p(\bar{\mathbf{z}})$, which implies that model output observations (or observed residuals) should be used to update the probability distribution of the model input, just as these are used to update the distributions of the parameters of the rainfall-runoff model and the model structural and measurement uncertainty. This can be achieved by adding $\bar{\mathbf{z}}$ to the parameters, so that Eq. 6.4 is replaced by:

$$p(\varphi, \theta, \sigma_\eta^2, \bar{\mathbf{z}}|\mathbf{u}) \propto p(\mathbf{u}|\varphi, \theta, \sigma_\eta^2, \bar{\mathbf{z}}) p(\varphi) p(\theta) p(\sigma_\eta^2) p(\bar{\mathbf{z}}) \tag{6.5}$$

Note that here we further assume that (unconditional to $\mathbf{u}$) the parameters φ, θ and σ_η^2 are independent of $\bar{\mathbf{z}}$. The model parameter priors $p(\varphi)$ can be derived from expert judgment and may be centred around deterministically calibrated parameter values, while uninformative (i.e. wide) priors are typically chosen for θ and σ_η^2. The remaining terms of the right-hand side of Eq. 6.5 are the likelihood $p(\mathbf{u}|\varphi, \theta, \sigma_\eta^2, \overline{\mathbf{z}})$ and the prior for $\overline{\mathbf{z}}$. We work these out in the next two subsections, starting with the latter.

Rainfall prior

The rainfall priors for all time instances were derived using a geostatistical approach. Let $z_t(s)$ denote the rainfall at location s in the catchment $\mathcal{A}$ at time $t \in \{1 \dots T\}$. We treat $z_t = \{z_t(s)|s \in \mathcal{A}\}$ as a realization of a random field Z_t. We further assume that $\log(Z_t)$ is a stationary normally distributed random field, characterized by a (constant) mean and isotropic variogram $\gamma(h)$ (where h is geographical distance). Since it is unrealistic to assume that $\log(Z_t)$ has the same statistical properties for all times t, in the case study we classified all times into a finite number of classes that are judged sufficiently homogeneous with respect to rainfall intensity, and assumed constant statistical properties within each class. Note that while we include spatial autocorrelation, we ignore temporal autocorrelation. In other words, we assume that the correlation between $\log(Z_t(s))$ and $\log(Z_{t+v}(s+h))$ is zero if $v \neq 0$. Ignoring temporal autocorrelation is acceptable if the temporal support of rainfall data is sufficiently large (Cecinati et al., 2017) (in the case study we consider rainfall accumulated over ten days). Estimation of the variogram $\gamma(h)$ for each rainfall intensity class may be done using the conventional Methods of Moments estimator and by pooling sample variograms derived for all time instants within the same rainfall intensity class (Muthusamy et al., 2017).

To sample from the distribution of Z_t we use conditional sequential Gaussian simulation (cSGS) (Cressie, 2015). For each time instant t, we first simulate fields of log-transformed rainfall, conditional to the observations $\log(z_t(s_i))$, $\quad i = 1 \dots n$, where n is the number of rain gauge locations. Next we back-transform these fields to the original scale to obtain conditional rainfall simulation fields $z_t^l = \{z_t^l(s)|s \in \mathcal{A}\}$, $\quad l = 1 \dots L$, where L is the number of simulated fields. Finally, each simulated field is spatially aggregated to obtain catchment average rainfall simulations:

$$\overline{z}_t^l = \frac{1}{|\mathcal{A}|} \int_{s \in \mathcal{A}} z_t^l(s) \mathrm{d}s \tag{6.6}$$

In practice, the integral in Eq. 6.6 is approximated by a summation over a (sufficiently dense) spatial grid. The simulations z_t^l are also generated on this same grid.

The set of L simulations of catchment average rainfall provides an empirical representation of the prior distribution of $\overline{z}_t$ for all t, which is an accurate approximation of the true prior if L is sufficiently large. The empirical prior cumulative distribution of $\overline{z}_t$ is then given by:

$$F_{\overline{z}_t}(a) = \frac{1}{L}\sum_{l=1}^{L} I(\overline{z}_t^l \leq a) \tag{6.7}$$

where I is an indicator function equal to 1 if its argument is true and 0 otherwise.

Likelihood function

Deriving the likelihood $p(\mathbf{u}|\varphi, \theta, \sigma_\eta^2, \overline{\mathbf{z}})$ is a major step in the Bayesian model calibration. For notational convenience we will drop the conditioning information in this subsection and write $\mathbf{u}$ instead of $\{\mathbf{u}|\varphi, \theta, \sigma_\eta^2, \overline{\mathbf{z}}\}$, but note that because of the conditioning information $\mathbf{u}$ satisfies Eq. 6.3, with all parameters of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$ known. We start by writing the joint distribution of $\mathbf{u}$ as a product of conditional distributions:

$$p(\mathbf{u}) = p(u_0) \cdot p(u_1|u_0) \cdot p(u_2|u_0, u_1) \ldots p(u_T|u_0 \ldots u_{T-1}) \tag{6.8}$$

Because $\mathbf{u}$ is the sum of $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$ it cannot be written as an AR(1) model and does not satisfy the Markov property (Durbin and Koopman, 2012). The conditional distributions can therefore not be reduced to a simple form. Instead, we use a Kalman filter approach (Kalman, 1960) to evaluate the conditional distribution in Eq. 6.8. Since all stochastic variables involved are normal the conditional distributions are also normal, leaving only their means and variances to be determined. These are derived in a recursive way ($t = 1 \ldots T$):

$$\hat{\varepsilon}_0^+ = \mathrm{E}\left[\varepsilon_0\right] = \mu_0, \quad \sigma_0^{2+} = \mathrm{var}(\varepsilon_0) = \sigma_0^2 \tag{6.9}$$

$$\hat{\varepsilon}_t^- = \beta_0 + \beta_1\hat{\varepsilon}_{t-1}^+ \tag{6.10}$$

$$\sigma_t^{2-} = \beta_1^2\sigma_{t-1}^{2+} + \sigma_\delta^2 \tag{6.11}$$

$$\hat{\varepsilon}_t^+ = \hat{\varepsilon}_t^- + k_t(u_t - \hat{\varepsilon}_t^-) \tag{6.12}$$

$$\sigma_t^{2+} = (1 - k_t)\sigma_t^{2-} \tag{6.13}$$

$$k_t = \frac{\sigma_t^{2-}}{\sigma_t^{2-} + \sigma_\eta^2} \tag{6.14}$$

Here, $\hat{\varepsilon}_t^- = E[\varepsilon_t|u_0 \dots u_{t-1}]$ is the time update, $\hat{\varepsilon}_t^+ = E[\varepsilon_t|u_0 \dots u_t]$ the measurement update and k_t is the Kalman gain. The prediction error variances associated with $\hat{\varepsilon}_t^-$ and $\hat{\varepsilon}_t^+$ are given by σ_t^{2-} and σ_t^{2+}, respectively. Obtaining the mean and variance of $\{u_t|u_0 \dots u_{t-1}\}$ is now easy and given by:

$$\mathrm{E}\left[u_t|u_0 \dots u_{t-1}\right] = \hat{u}_t = \hat{\varepsilon}_t^- \tag{6.15}$$

$$\mathrm{var}(u_t|u_0 \dots u_{t-1}) = \sigma_t^{2-} + \sigma_\eta^2 \tag{6.16}$$

The log-transformed conditional distribution at time $t > 0$ is thus given by:

$$\log(p(u_t|u_0, \dots, u_{t-1})) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_t^{2-} + \sigma_\eta^2) - \frac{1}{2}\left\{\frac{(u_t - \hat{u}_t)^2}{\sigma_t^{2-} + \sigma_\eta^2}\right\} \tag{6.17}$$

while for $t = 0$ we have $u_0 \sim \mathcal{N}(\mu_0, \sigma_0^2 + \sigma_\eta^2)$. Taking the logarithm of Eq. 6.8 implies that we must sum over all time steps so that the log-likelihood is given by (defining $\hat{\mu}_0 = \mu_0$ and $\sigma_0^{2-} = \sigma_0^2$):

$$\log(p(\mathbf{u})) = -\frac{T+1}{2}\log(2\pi) - \frac{1}{2}\sum_{t=0}^{T}\left(\log(\sigma_t^{2-} + \sigma_\eta^2)\right) - \frac{1}{2}\sum_{t=0}^{T}\left(\frac{(u_t - \hat{u}_t)^2}{\sigma_t^{2-} + \sigma_\eta^2}\right) \tag{6.18}$$

Markov chain Monte Carlo

The proportionality sign in Eq. 6.5 means that the posterior on its left-hand side differs from the right-hand side by a multiplicative constant. This constant is unknown (that is to say, it can only be obtained using formidable computer power) and hence the posterior cannot be determined explicitly. To overcome this problem a common approach is to sample from the posterior distribution $p(\varphi, \theta, \sigma_\eta^2, \bar{\mathbf{z}}|\mathbf{u})$ using Markov chain Monte Carlo (MCMC). In this paper we use the Metropolis algorithm, which may not be the most efficient approach but perfectly valid and relatively easy to implement. It is described in detail in Brooks et al. (2011). Thus, a large sample N of the joint posterior distribution of $(\varphi, \theta, \sigma_\eta^2, \bar{\mathbf{z}})$ is generated, where N is typically taken in the range 10^4 to 10^5. Convergence can be assessed by running several independent Markov chains and checking that the sample distributions are sufficiently similar (Gelman et al., 2014). Another important performance indicator is the acceptance rate, i.e. the number of accepted proposals divided by the total number of proposals. We manually tuned the size of the jump in the parameter space to obtain an acceptance rate between 0.25 and 0.5 (suggested by Rosenthal et al., 2011) to obtain sufficient exploration of the parameter space without grossly deteriorating efficiency. The acceptance rate was calculated after removing the first set of proposals, also called the burn-in phase.

6.2.3 Prediction

Once the joint conditional posterior distribution of all parameters and the catchment-average rainfall (Eq. 6.5) has been obtained, it can be used for discharge prediction. To derive the prediction, it is important to distinguish between the true discharge $\mathbf{d}$ and the measured discharge $\mathbf{y}$. From the model defined in Eq. 6.1 it follows that the log-transformed true discharge $\log(\mathbf{d}) = \log(\mathbf{y}) - \boldsymbol{\eta}$ is the sum of the log-transformed model output $H(\overline{\mathbf{z}}, \varphi))$ and the log-transformed model structural uncertainty $\boldsymbol{\varepsilon}$. Using the law of total probability, the probability distribution of the discharge can be written as:

$$p(\mathbf{d}) = \iiint p(\mathbf{d}|\varphi, \theta, \overline{\mathbf{z}}) \cdot p(\varphi, \theta, \overline{\mathbf{z}}) \, \mathrm{d}\varphi \mathrm{d}\theta \mathrm{d}\overline{\mathbf{z}} \tag{6.19}$$

This multi-dimensional integral is usually solved numerically using Monte Carlo sampling. Since $\log(\mathbf{d}) = \log(H(\overline{\mathbf{z}}, \varphi)) + \boldsymbol{\varepsilon}$, sampling from $p(\mathbf{d}|\varphi, \theta, \overline{\mathbf{z}})$ involves a deterministic run of the rainfall-runoff model and simulating a realization from the AR(1) model of $\boldsymbol{\varepsilon}$. One might think that realizations of the posterior $p(\varphi, \theta, \overline{\mathbf{z}})$ are already available from the MCMC sampling, described in Section 6.2.2, but this is not the case. The problem is that realizations of this posterior are only available for the calibration period, i.e. a time period with discharge measurements. Since prediction is only needed for time periods without discharge measurements, we consider a case in which there are no discharge measurements at or near the prediction time. Hence, there are no realizations of the posterior $p(\varphi, \theta, \overline{\mathbf{z}})$ for a time period without discharge measurements either. To make the distinction between the calibration and prediction periods explicit, let the prediction period be from $t = T + V + 1$ to $t = T + V + W$, where V is typically larger than the catchment response time. We denote the catchment-average rainfall for the prediction period by $\overline{\mathbf{z}}_{+}$. Thus, we derive $p(\mathbf{d}_{+})$ from the (posterior) distribution of φ, θ and $\overline{\mathbf{z}}_{+}$ using Eq. 6.19, with $\mathbf{d}$ replaced by $\mathbf{d}_{+}$ and $\overline{\mathbf{z}}$ replaced by $\overline{\mathbf{z}}_{+}$.

We consider three approaches to derive a "posterior" distribution $p(\varphi, \theta, \overline{\mathbf{z}}_{+})$:

1. For φ and θ, use the MCMC sample of their joint posterior as explained in Section 6.2.2. For $\overline{\mathbf{z}}_{+}$, apply a linear correction to its prior mean as follows:

$$\mu_{\overline{\mathbf{z}}_{+}^{po}} = a\mu_{\overline{\mathbf{z}}_{+}^{pr}} + b \tag{6.20}$$

where the coefficients a and b are derived by fitting a linear regression between the means of the rainfall posterior distribution $\mu_{\bar{z}_{po}}$ and the means of the rainfall prior distribution $\mu_{\bar{z}_{pr}}$ for the calibration period. A similar approach is used in Huard and Mailhot (2008, Section 5.1). Thus, realizations $\bar{z}_t^l$ of the rainfall prior are converted to realizations $\tilde{\bar{z}}_t^l$ of the "posterior" by shifting the means, while keeping the same shape and standard deviation:

$$\tilde{\bar{z}}_t^l = \bar{z}_t^l - \mu_{\bar{z}_{pr}} + \mu_{\bar{z}_{po}} = \bar{z}_t^l + (a-1)\mu_{\bar{z}_{pr}} + b \tag{6.21}$$

This approach has the disadvantage that only the mean is corrected, using a simple linear transform. It is not obvious how the correction can be improved. Another disadvantage is that the posteriors of φ and θ are decoupled from that of $\bar{\mathbf{z}}_+$. In other words, they are made statistically independent.

2. For φ and θ, use the MCMC sample from their joint posterior as explained in Section 6.2.2. For $\bar{\mathbf{z}}_+$, let the posterior distribution be identical to its prior. This approach has the disadvantage that the posteriors of φ and θ are again decoupled from that of $\bar{\mathbf{z}}_+$.

3. Ignore that discharge measurements inform catchment-averaged rainfall and exclude $\bar{\mathbf{z}}$ from the Bayesian calibration. Thus, the prior distribution of $\bar{\mathbf{z}}_+$ as derived using block kriging is used as in Approach 2, but unlike in Approach 2 the parameters φ and θ are calibrated without including $\bar{\mathbf{z}}$ in the calibration procedure. The disadvantage of this approach is that it ignores that rainfall and discharge are dependent, but it has two important advantages. First, there is no interference between calibration of model parameters and rainfall input, which causes problems in the prediction period. Second, the overall number of parameters to be calibrated is much smaller compared to the case in which rainfall input is to be calibrated as well. While this approach essentially boils down to Eq. 6.4, rainfall uncertainty during the calibration period must be taken into account. This is achieved by integrating the likelihood in Eq. 6.4 over all realizations of the rainfall $\bar{\mathbf{z}}$:

$$p(\mathbf{u}|\varphi,\theta,\sigma_\eta^2) = \int p(\mathbf{u},\bar{\mathbf{z}}|\varphi,\theta,\sigma_\eta^2)\,\mathrm{d}\bar{\mathbf{z}} = \int p(\mathbf{u}|\varphi,\theta,\sigma_\eta^2,\bar{\mathbf{z}})\cdot p(\bar{\mathbf{z}})\,\mathrm{d}\bar{\mathbf{z}} \tag{6.22}$$

Note that here we used the fact that $\bar{\mathbf{z}}$ is independent of the model parameters.

6.2.4 Validation measures

To evaluate the methodology it is necessary to statistically validate the discharge predictions and associated prediction uncertainty using independent discharge measurements. We do so for the prediction period, since discharge measurements during this period were not used for calibration and prediction. However, since discharge measurements are not error-free, the validation must take discharge measurement error into account. Discharge measurement error was defined by η in Section 6.2.1. The distribution of η is characterized by a single parameter σ_η^2, the calibration of which was explained in Section 6.2.2.

With little modification, prediction equation Eq. 6.19 can be rewritten to include discharge measurement error:

$$p(\mathbf{y}_+) = \iiiint p(\mathbf{y}_+|\varphi, \theta, \overline{\mathbf{z}}_+, \sigma_\eta^2) \cdot p(\varphi, \theta, \overline{\mathbf{z}}_+, \sigma_\eta^2)\, \mathrm{d}\varphi \mathrm{d}\theta \mathrm{d}\overline{\mathbf{z}}_+ \mathrm{d}\sigma_\eta^2\,, \qquad (6.23)$$

Here, $\mathbf{y}_+ = [y_{T+V+1}, y_{T+V+2}, \ldots, y_{T+V+W}]^T$ denotes the modelled discharge measurements for the prediction period. The predictions and prediction intervals of $\mathbf{y}_+$ can now be compared to the actual discharge observations to assess the quality of the model. To distinguish between modelled and observed discharge measurements, in this section we denote the first by $\mathbf{y}_+^m$ and the second by $\mathbf{y}_+^o$.

Several measures are employed to assess the quality of the model and the effect of the rain gauge density on the discharge prediction intervals. These are the mean error (ME), the root mean squared error (RMSE), the Nash-Sutcliffe model efficiency coefficient (NSE) and the prediction intervals coverage probability (PICP). The PICP is the percentage of observations covered by a defined prediction interval (Shrestha and Solomatine, 2008).

$$\mathrm{ME} = \frac{1}{W} \sum_{t=T+V+1}^{T+V+W} (\overline{y}_t^m - y_t^o), \qquad (6.24)$$

where $\overline{y}_t^m$ is the arithmetic mean of the simulated discharges at time t.

$$\mathrm{RMSE} = \sqrt{\frac{1}{W} \sum_{t=T+V+1}^{T+V+W} (\overline{y}_t^m - y_t^o)^2}, \qquad (6.25)$$

$$\mathrm{NSE} = 1 - \frac{\sum_{t=T+V+1}^{T+V+W} (\overline{y}_t^m - y_t^o)^2}{\sum_{t=T+V+1}^{T+V+W} (y_t^o - \overline{y}^o)^2}, \qquad (6.26)$$

where $\overline{y}^o = \frac{1}{W} \sum_{t=T+V+1}^{T+V+W} y_t^o$.

$$\mathrm{PICP} = \frac{100}{W} \sum_{t=T+V+1}^{T+V+W} I_t \qquad (6.27)$$

with

$$I_t = \begin{cases} 1 & \text{if } y_t^m(low) \leq y_t^o \leq y_t^m(up) \\ 0 & \text{otherwise.} \end{cases} \tag{6.28}$$

where $y_t^m(low)$ and $y_t^m(up)$ are the lower and upper limits of the prediction interval for y_t as computed by the model. In the case study, we will compute the PICP both for the 50% and 90% prediction intervals.

6.2.5 Sampling density scenarios

To investigate the effect of rain gauge density on the uncertainty of the discharge predictions, several scenarios were developed. Each scenario comprises a number of rain gauges that are optimally selected from the existing rain gauge locations in the study area. The optimal locations were derived using spatial simulated annealing (SSA), by minimizing the time-averaged block kriging prediction error variance:

$$\frac{1}{T+W}\{\sum_{t=1}^{T} \sigma_{OK}^2(\mathcal{A},t) + \sum_{t=T+V+1}^{T+V+W} \sigma_{OK}^2(\mathcal{A},t)\} \tag{6.29}$$

where $\sigma_{OK}^2(\mathcal{A},t)$ is the rainfall ordinary block-kriging variance at time t and using the entire study area $\mathcal{A}$ as a “block”. Because there is no obvious analytical means to compute the block kriging variance of a lognormally distributed variable, the block kriging variance was approximated by computing the variance of 200 rainfall fields simulated using cSGS, in a similar way as described in Section 6.2.2. For more details about SSA we refer to Van Groenigen and Stein (1998) and Wadoux et al. (2017).

6.3 Data and model

6.3.1 Study area

The study area is the Thur basin ($1,696$ km^2), located in the North-East of Switzerland (Fig. 6.1). The Thur river is a tributary of the Rhine river and is the largest non-regulated river in Switzerland (Lopez and Seibert, 2016). The elevation within the basin ranges from 357 to 2437 meters above sea level (m.a.s.l.) with an average height of 765 m.a.s.l. The Thur basin has been subject of several previous studies (e.g. Melsen et al., 2016) and data availability is large. Three datasets are used in this study:

- Daily average temperature data for the period 2004-2011 from the Swiss Federal Office for Meteorology and Climatology (MeteoSwiss). Daily temperature

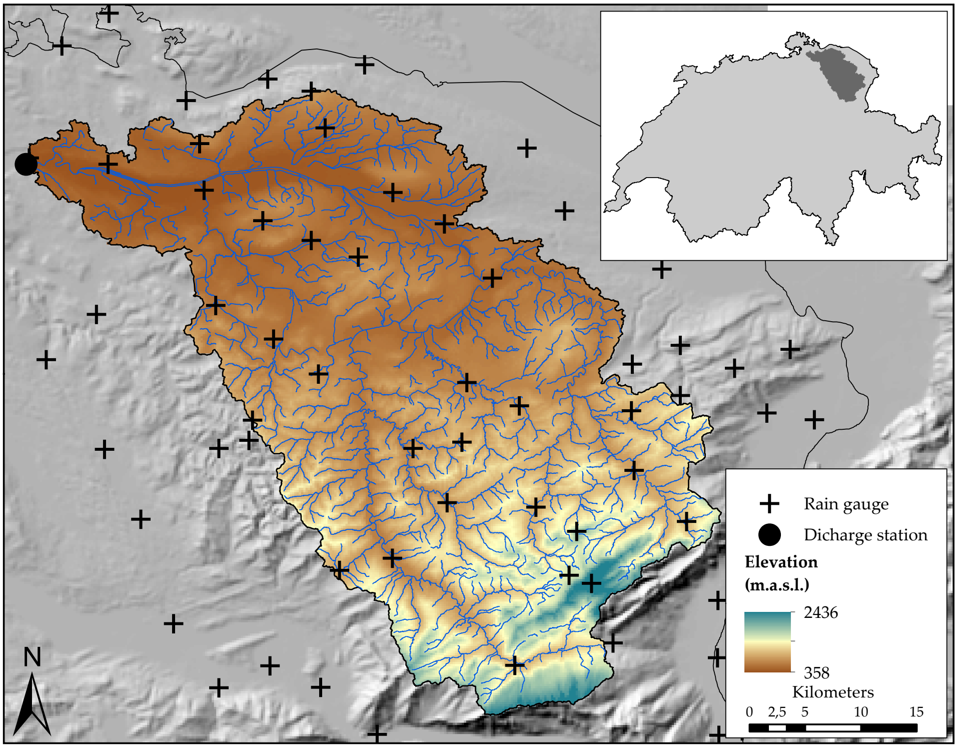


***Figure 6.1** – Map of the Thur River Basin with locations of rain gauges and discharge station.*

is provided as a spatial grid of about 2300 m × 2300 m resolution based on an interpolation between meteorological stations (Frei, 2014).

- Daily tipping bucket rain gauge data from MeteoSwiss. Combining manual and automatic gauges, a total of 29 rain gauges also measuring snowfall are available for the period 2004-2011. For the purpose of this study, we included another set of 40 gauges that are within a maximum distance of 20 km from the basin boundary.
- Daily cumulative discharge data for the period 2004-2011 from the Swiss Federal Office for the Environment (FOEN). The discharge measuring station is located at the outlet of the basin at Andelfingen, at an altitude of 356 m.a.s.l.

6.3.2 The HBV model

The HBV model (Lindström et al., 1997) is a conceptual lumped rainfall-runoff model developed by the Swedish Meteorological and Hydrological Institute. We chose the

HBV model because of its low input data requirement and because it includes a snow melt routine. The required input data consist of time series of catchment-averaged rainfall and air temperature. The model is structured in different routines such as snow melt, evaporation, soil moisture and groundwater. Channel routing is described by a triangular hydrograph. For more detailed information about HBV, we refer to the original publication of Lindström et al. (1997) and to Heistermann and Kneis (2011) for the specific version used in the case study.

6.3.3 Application to the case study

Rainfall-runoff model - We decided to implement a simplified version of the HBV model from the R package RHydro (Reusser et al., 2017). Time series of rainfall and temperature were split into calibration (2004-2007) and validation (2008-2011) periods. The first year of the two periods (2004 and 2008) was considered as a warm-up period and discarded from the results. For practical convenience the daily time series were aggregated to 10-day averages. This is discussed more extensively in the discussion. Prior to the Bayesian calibration and uncertainty analysis, a deterministic calibration was performed. We used a differential evolution algorithm to minimize the mean squared error (MSE) between measured and predicted discharge for the calibration period. The estimated parameters are shown in Table 6.1 and were used to help define plausible ranges for the priors of the model parameters.

Rainfall input - We defined ten rainfall intensity classes based on the 10-day catchment averaged rainfall amounts and fitted exponential variogram models for each class. Rainfall intensity increases with class number. Variogram fitting was based on all rainfall observations that are in the same class, both using rain gauges inside and outside the basin and using an approach described in Muthusamy et al. (2017). A plot of the fitted variograms is presented in Fig. 6.2. Periods with an average rainfall of less than 0.1 mm were not interpolated and considered as dry. Log rainfall was simulated 500 times on a 1 km × 1 km resolution grid using the class-specific variogram and were conditioned on the rain gauges inside the catchment only. Processing was done using the R package gstat (Gräler et al., 2016). Log rainfall simulations were back-transformed at point locations and spatially averaged (Heuvelink and Pebesma, 1999).

Eight rain gauge scenarios were considered, comprising 1, 2, 5, 10, 15, 20, 25 and 29 rain gauges, respectively. For each scenario, the rain gauges were selected using SSA by thinning the existing network. The minimization criterion was the average block kriging variance computed by discretization of the area into 500 sub-areas. Implementation was done with the R package spsann (Samuel-Rosa, 2017). The

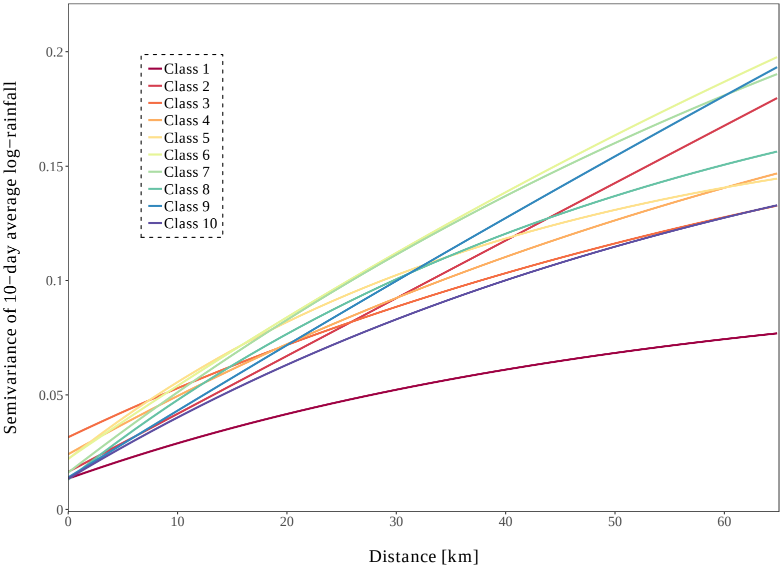


***Figure 6.2** – Fitted exponential variograms for each of the ten rainfall classes.*

initial SSA temperature was set to 0.1 and the cooling factor was set to 0.8. The total number of SSA iterations was fixed at 10,000. Five out of the eight tested scenarios are reported in this study.

Bayesian inference - Model parameters and their priors are shown in Table 6.1. Prior parameter distributions were chosen based on expert knowledge, previous work in the same basin and optimized parameter values of the deterministic calibration. For each rain gauge scenario the Bayesian inference was performed. The number of MCMC iterations was fixed at 10^6 for Approaches 1 and 2 and to 10^4 for Approach 3 (6.2.3). The process was repeated several times to ensure convergence of the parameter estimates.

6.4 Results

6.4.1 Calibration

Figure 6.3 shows the posterior distribution of the calibrated model parameters for Approaches 1 and 2, for different rain gauge density scenarios, along with their prior distribution. For most parameters the posterior distribution is narrower than the prior distribution. This particularly holds for parameters associated with the routing routine (maxbas, n, k), the initial state (snow, sm, suz, slz) and the error model (β_0,

***Table 6.1** – Model parameters and error model parameters with initial values and prior distributions. The implementation of the HBV model is based on Heistermann and Kneis (2011).*

Parameter name	Definition	Initial value through deterministic calibration[1]	Prior distribution of parameter[2]
Model parameters			
CFMAX	degree day factor for snow melt [mm/°C/d]	2.789	*Beta*[3, 5, 0, 10]
TT	temperature threshold below which precipitation falls as snow [°C]	0.7787	*Beta*[3, 2, -3, 3]
FC	field capacity [mm]	142.4	*Beta*[1, 4, 0, 200]
MINSM	minimum soil moisture for storage [mm]	2.190	*Beta*[1, 4, 0, 200]
BETA	parameter to control the fraction of rain and snow melt partitioned for groundwater recharge [-]	0.5580	*Beta*[2, 1, 0, 1]
LP	fraction of soil moisture-field capacity-ratio above which actual evapotranspiration equals potential evapotranspiration [-]	0.7817	*Beta*[2, 1, 0, 1]
CET	correction factor for potential evapotranspiration [-]	0.1153	*Beta*[4, 4, -10, 20]
KPERC	percolation coefficient [1/d]	2.499	*Beta*[4, 4, 0, 5]
K0	fast storage coefficient of soil upper zone [1/d]	0.4994	*Beta*[1, 4, 0, 0.5]
UZL	threshold above which soil upper zone storage empties at rate computed by storage coefficient K0 [mm]	1.522	*Beta*[3, 4, 0, 60]
K1	slow storage coefficient of soil upper zone [1/d]	0.3883	*Beta*[1, 4, 0, 0.5]
K2	storage coefficient of soil lower zone [1/d]	2.8×10^{-4}	*Beta*[1, 4, 0, 0.1]
MAXBAS	length of (triangular) unit hydrograph [d]	4.067	*Beta*[1, 4, 0, 6]
etpmean	mean evaporation [mm/d]	2.596	*Beta*[1, 3, 0, 50]
tmean	mean temperature [°C]	7.259	*Beta*[4, 4, -20, 30]
n	(real) number of storages in linear storage cascade	1.997	*Beta*[1, 2, 0, 7.5]
k	decay constant for linear storage cascade	0.4540	*Beta*[1, 2, 0, 5]
Initial state parameters			
snow	snow storage [mm]	178	*Beta*[1 ,3 , 0, 200]
sm	soil moisture storage [mm]	369	*Beta*[1, 3, 0, 200]
suz	soil upper zone storage [mm]	117	*Beta*[1, 3, 0, 200]
slz	soil lower zone storage [mm]	44	*Beta*[1, 3, 0, 200]
Error model parameters			
β_0	constant in the AR(1) in Eq. 6.2	-	U[0,1]
β_1	coefficient for the AR(1) in Eq. 6.2	-	U[0,1]
σ_δ^2	standard deviation for the AR(1), Eq. 6.1	-	U[0,1]
σ_η^2	standard deviation for the measurement error, Eq. 6.1	-	U[0,1]
$\bar{\mathrm{z}}$	vector of length $T + W$ for the rainfall parameters	-	$F_{\bar{\mathrm{z}}}$, see Eq. 6.7

[1] A deterministic calibration is performed prior to the Bayesian calibration. The parameters are optimized with differential evolution. The objective function is the MSE between the predicted and measured discharge.

[2] *Beta*[a,b,c,d] represents a beta distribution in the interval [c,d] with shape parameters a and b. U[a,b] is a uniform distribution over the interval [a, b].

β_1, σ_δ^2, σ_η^2). Note that while the posterior distribution of some parameters (e.g. lp, k perc, k0, k1 and k) is comparable for different rain gauge density scenarios, all other parameter distributions show (large) differences between different rain gauge density scenarios. The distributions do not seem to be narrower in case of higher rain gauge density.

Figure 6.4 shows the posterior and prior distribution of the calibrated parameters for Approach 3, for different rain gauge scenarios. The posterior distributions are much narrower than the prior distributions, particularly for the parameters of the snow routine (cfmax, tt, fc), the routing routine (maxbas, n, k), initial state and error model. For parameters lp, k0 and k1, the posterior distributions are very similar to the priors. Some posterior distributions are multimodal. In contrast to Fig. 6.3, all parameters have very similar posterior distributions for different rain gauge densities.

6.4.2 Prediction

Figure 6.5 shows the rainfall prior and corrected prior ("posterior") for the prediction period. Recall that the prior distribution was directly sampled for Approaches 2 and 3 while the rainfall "posterior" distribution was sampled for Approach 1. As expected, the prediction interval width increases when using a smaller number of rain gauges. Both rainfall prior and posterior distributions showed very similar prediction interval widths and mean values. Small differences can be observed when using a small number of rain gauges (i.e. 1 or 2 rain gauges).

Figure 6.6 shows the 90% prediction intervals of the discharge for the three approaches and including/excluding various uncertainty sources. For the case where all uncertainty sources are accounted for (plots (a)), there is a clear pattern towards smaller width of the prediction intervals with an increase of the number of rain gauges. While there is no clear difference in terms of prediction intervals between the three approaches, Approach 3 provides a larger interval width at a certain time period (e.g. events at 2009-07 and 2010-08) for the scenario involving 15 rain gauges. Note that the differences between rain gauge scenarios are most pronounced for high flows and become negligible for low flows.

Figures 6.6a do not provide information about the separate effect of input, model parameters and model structural uncertainty on the joint predictive uncertainty. Therefore we also performed an uncertainty propagation analysis that includes/excludes the various uncertainty sources for the cases where: (b) model structural uncertainty was ignored; and (c) model structural and model parameter (comprising initial state and error model parameters) uncertainty were ignored.

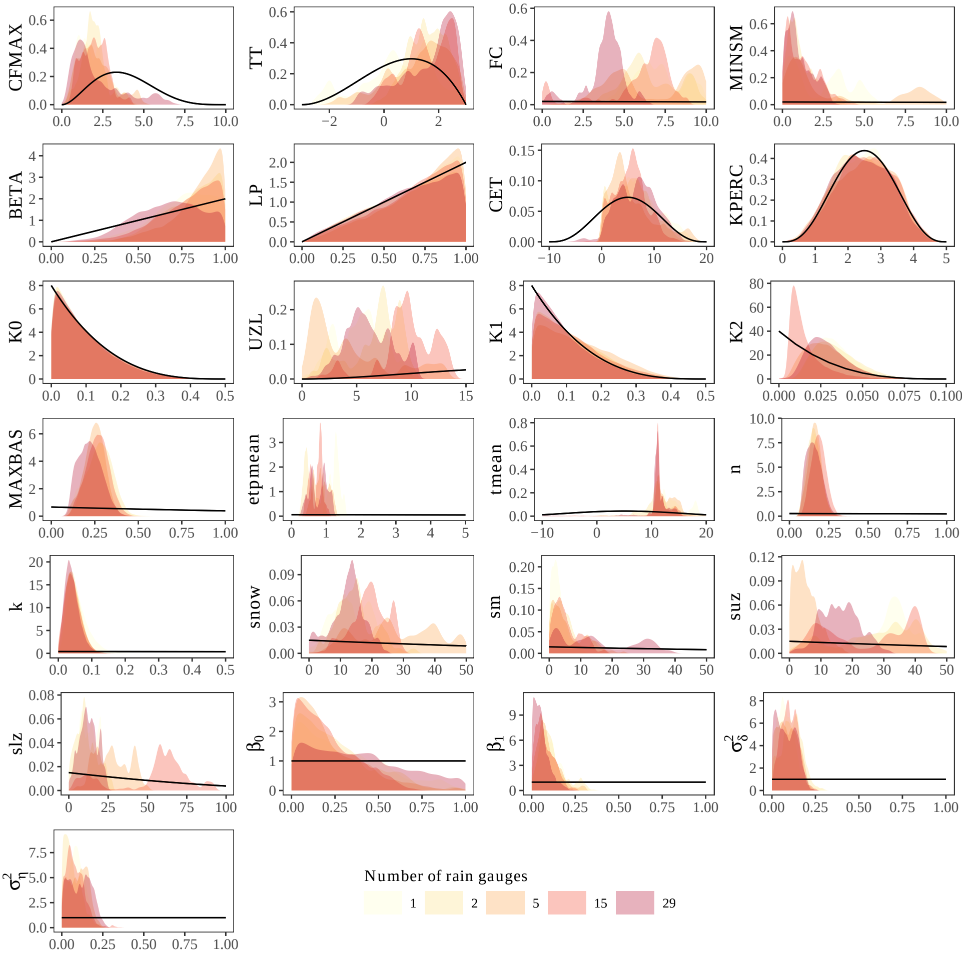


***Figure 6.3** – Parameters estimated by Bayesian calibration for Approaches 1 and 2. Black lines represent prior distributions and coloured shapes posterior densities for different rain gauge density scenarios. Rainfall parameter results not shown.*

When model structural uncertainty is ignored (b), the prediction intervals are similar, except for Approach 3, for which the uncertainty decreases. With the additional effect of model parameter uncertainty removed (c), the prediction intervals become much narrower, with large differences for different rain gauge densities, i.e. the larger the number of rain gauges, the smaller the prediction interval width. The latter reduction is particularly visible when increasing the density from 5 to 15 rain gauges, and becomes marginal when using between 15 to 29 rain gauges. As noted before, the largest difference is obtained for high flow periods.

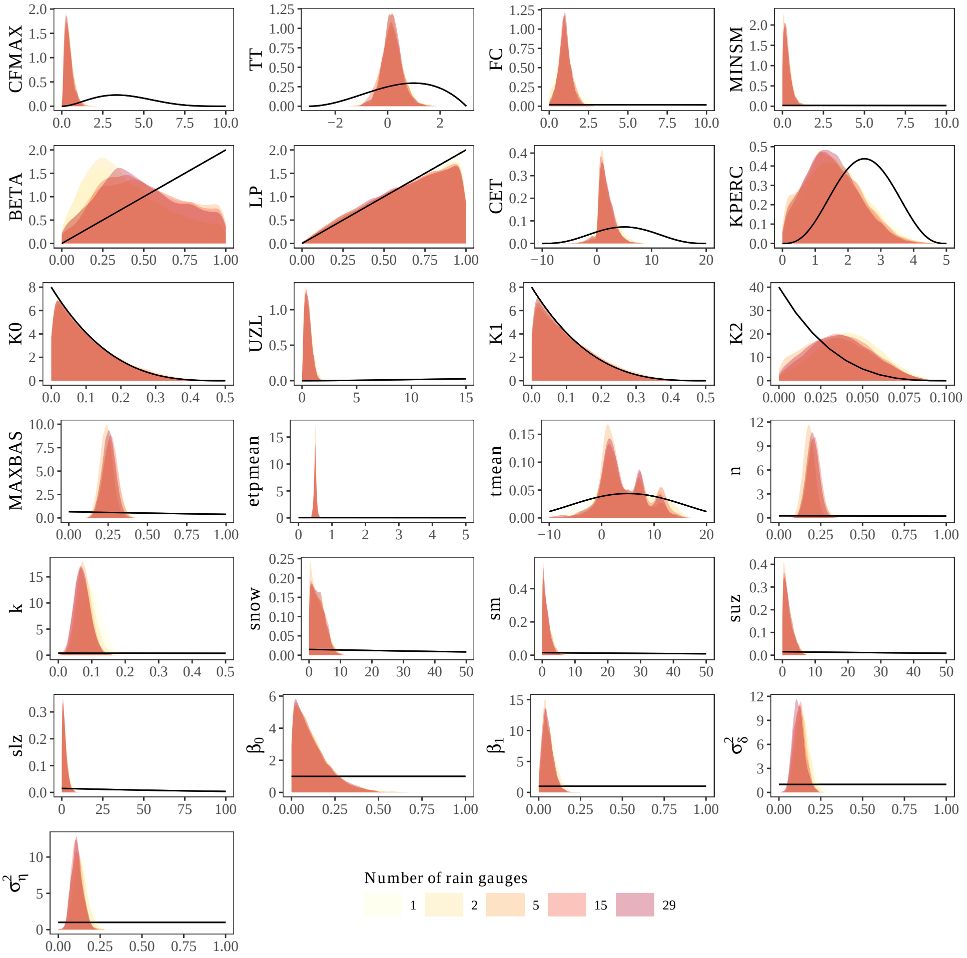


Figure 6.4 *– Parameters estimated by Bayesian calibration for Approach 3. Black lines represent prior distributions and coloured shapes posterior densities for different rain gauge density scenarios.*

6.4.3 Validation

Table 6.2 shows the validation statistics for the three approaches and the five tested rain gauge scenarios. For the three approaches, increasing the number of rain gauges led to an increase of the predictive power of the model (increase of the NSE). This increase was generally modest, except for Approach 3. It was accompanied by a modest decrease of the residual as characterized by the RMSE. The model predictions were practically unbiased (biggest ME deviation from zero equals -0.22), which shows that the prediction inaccuracy was mainly due to random error, not systematic error. There is no clear pattern regarding the PICP with increasing rain gauge

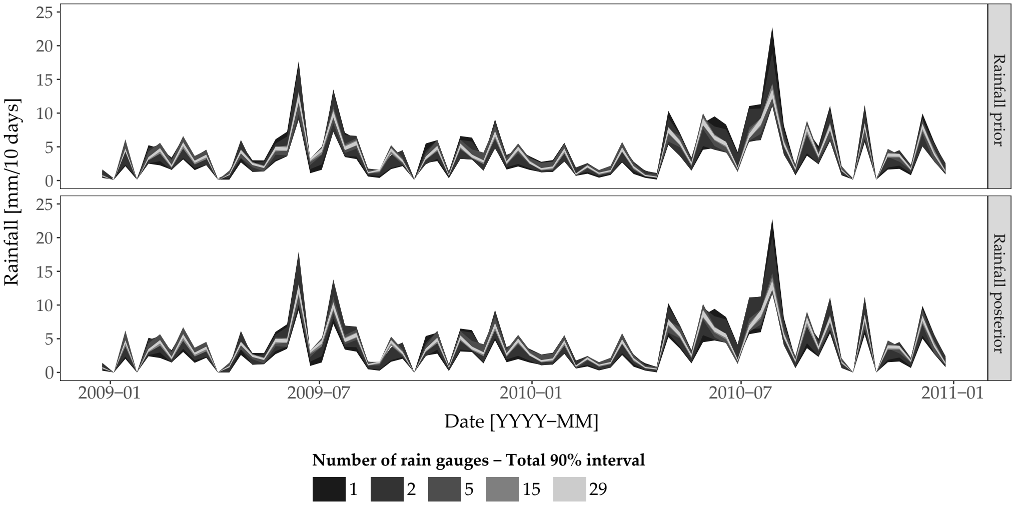


***Figure 6.5** – Rainfall priors for the prediction/validation period. The rainfall posterior is the corrected prior for Approach 1 for the prediction/validation period. Approaches 2 and 3 sample from the rainfall prior for prediction and validation.*

density. For both intervals (50% and 90%), the percentage of observations covered by the interval was within a reasonable range of variation. Approaches 1 and 2 were very similar in terms of validation statistics, particularly when a large number of rain gauges was used. This is according to expectations, as the effect of the linear correction on the prior diminishes with an increasing number of rain gauges.

6.5 Discussion

6.5.1 Consistency of parameter estimates

Our experiments suggest that several model parameters estimated in Approaches 1 and 2 might be weakly identified because of their wide posterior distribution. Gelman et al. (2014) stressed that the concept of identification is not so important in the Bayesian perspective and that one must rather look at how much information is supplied by the data, i.e that the joint parameter posterior must occupy less space than the joint prior distribution. In our case, Fig. 6.3 and Fig. 6.4 show that posteriors were narrower than the priors. This indicates that information was supplied by the data and explains why the parameter posteriors were actually accurate predictors, as shown by the NSE value being higher for Approaches 1 and 2 than for Approach 3, despite the wider posterior distributions of parameters of the third approach.

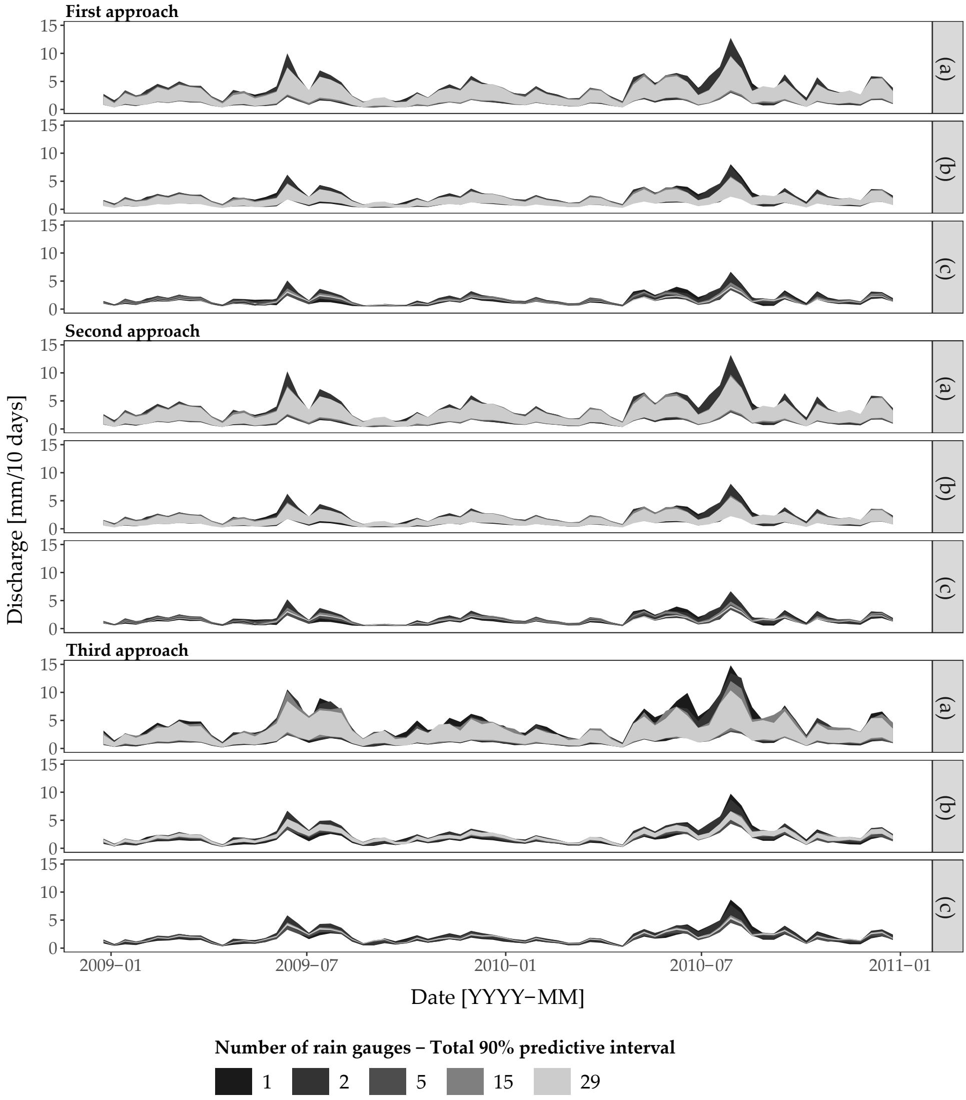


***Figure 6.6** – Prediction of the discharge using the three approaches for the cases (a) all sources of error are accounted for, (b) model structural uncertainty is ignored and (c) model structural uncertainty and model parameter uncertainty (excluding the rainfall input parameters for Approaches 1 and 2) are ignored.*

The posterior of parameter β_1 of the AR(1) model suggests that temporal correlation of the model structural error was weak to moderate, which agrees with Huard and Mailhot (2008). Parameters of the model structural error were well identified. Realistic assumptions regarding the model structural error model formulation play a major role to distinguish between model structural and input uncertainty. Since

***Table 6.2** – Validation measures for the three approaches and five rain gauge densities, computed over the period 2009-2011.*

	Number of rain gauges				
	1	2	5	15	29
Approach 1					
NSE	0.43	0.44	0.47	0.48	0.49
RMSE	1.17	1.16	1.13	1.12	1.10
ME	−0.04	−0.06	−0.19	−0.16	−0.08
PICP 50%	60.81	56.76	58.11	59.46	56.76
PICP 90%	90.54	90.54	90.54	91.89	91.89
Approach 2					
NSE	0.43	0.43	0.46	0.48	0.49
RMSE	1.17	1.17	1.14	1.12	1.11
ME	−0.11	−0.08	−0.22	−0.14	−0.08
PICP 50%	60.81	60.81	57.76	60.81	56.76
PICP 90%	87.84	87.84	91.89	91.89	91.89
Approach 3					
NSE	0.19	0.28	0.40	0.37	0.42
RMSE	1.39	1.31	1.20	1.23	1.19
ME	0.18	0.14	−0.07	0.13	0.09
PICP 50%	52.70	56.76	52.70	58.11	55.40
PICP 90%	90.54	93.24	87.84	86.48	91.89

model structural uncertainty is incorporated explicitly it is unlikely that input uncertainty compensates for deficits in the model structure (Thyer et al., 2009). We followed the approach of Beven and Freer (2001) that model structural uncertainty can be described by a first order autoregressive model. We acknowledge that more complex structures of model residuals can be formulated, such as by using an ARMA or ARIMA model, but this would be at the expense of increasing parameter space dimensionality.

Note that Eq. 6.1 can be extended to impose $\mathrm{E}[e^{\varepsilon}] = \mathrm{E}[e^{\eta}] = 1$. In this case, the mean μ_η of the η_t is no longer zero but must satisfy the identity $\mu_\eta = -\frac{1}{2}\sigma_\eta^2$. Similarly, to ensure $\mathrm{E}[e^{\varepsilon}] = 1$, it is not difficult to show that the identity $\frac{2\beta_0}{1-\beta_1} + \frac{\sigma_\delta^2}{1-\beta_1^2} = 0$ must hold, implying that one of the three parameters is determined by the other two (e.g. we could impose $\beta_0 = -\frac{1}{2}\frac{1-\beta_1}{1-\beta_1^2}\sigma_\delta^2$). While we think this would be a preferred approach, it did not matter much for our case study because the values of the mean structural error and mean measurement error were in most cases close to 1, and never greater than 1.6.

6.5.2 Prediction

In contrast to many studies reported in the literature (e.g. Kavetski et al., 2006), analysis of the predictive uncertainty shows that in the case study the contribution of rainfall uncertainty is relatively small and that discharge predictive uncertainty is mainly dominated by model structural and model parameter uncertainty. As a consequence, the effect of the rain gauge density diminishes if model parameter and model structural uncertainty is accounted for. Large model parameter and model structural uncertainty offsets small input rainfall uncertainty. For Approach 3, model structural uncertainty is clearly the largest contributor to the total predictive uncertainty. Several investigations obtained similar results. For example, Engeland et al. (2005) showed that model structural uncertainty is larger than model parameter uncertainty for simple conceptual models with few well-defined parameters. We also confirm the study of Talamba et al. (2010) for a lumped hydrological model. Talamba et al. (2010) showed, for a fully distributed hydrological model, that accounting for input rainfall uncertainty did not lead to a substantial change in terms of estimated parameters and model performance, because other sources of uncertainty dominated the total predictive uncertainty. This is similar to our case study, where in all three approaches the prediction intervals have a similar width and range. The validation measures show that Approaches 1 and 2, i.e. the case where rainfall parameters are calibrated, outperform Approach 3.

6.5.3 Differences between the three approaches

Comparison of the three approaches revealed that Approach 3 has larger prediction intervals and poorer model performance, despite the fact that in Approach 3 most parameters have a well-defined unimodal posterior distribution. Note also that the choice of approach leads to different posterior ranges of parameter estimates. The smaller number of parameters to calibrate in Approach 3 (i.e. the input rainfall parameters are not calibrated) suggests that inference benefits from the reduced dimensionality of the parameter space. The validation results show that this was not the case when a small number of rain gauges is used (for instance, NSE = 0.29 using 1 rain gauge). Thyer et al. (2009) and Huard and Mailhot (2008) reported similar results when calibrating time-dependent rainfall input parameters. They showed how calibrating input rainfall parameters for each time step compensates for the situation where a rainfall event is not recorded by a small number of rain gauges, and how this can lead to a near-perfect match between the observed and predicted discharge. In the latter case, Huard and Mailhot (2008) demonstrated that, since model and input rainfall parameters are estimated jointly, it is likely that the input rainfall parameters compensate for structural deficits of the model. In our case this was avoided by: (i) explicitly accounting for model structural uncertainty; and (ii) defining meaningful priors for the input rainfall parameters using geostatistical analysis. In addition, Fig. 6.6 shows that model structural error was larger for Approach 3, while model parameter uncertainty was larger for Approaches 1 and 2.

From a numerical perspective, a major difference between Approaches 1 and 2 and Approach 3 is the number of parameters to calibrate. Hydrologists tend to shy-away from high-dimensional rainfall input parameter space, because it often leads to question the statistical significance of the inferred parameters (e.g. Vrugt et al., 2008). A solution is to resort to MCMC search algorithms, which are efficient to explore the multi-dimensional and correlated parameter space. This has been recently tackled in the hydrological literature (e.g. Laloy and Vrugt, 2012) and in the more general statistical literature (e.g. Ter Braak, 2006). Another solution is to reduce the dimension of the parameter space. Alternative methods to Approach 1 (i.e. the case where each time step is an input rainfall parameter to calibrate) exist. For example, Kavetski et al. (2006) use storm-event multipliers under the assumption of perfect dependence of input errors within single storm events. By letting these multipliers vary according to the plausible range of hydrological variation, they correct for systematic error in the rainfall input. The major limitation is the need to define hydrological ranges in which the calibrated multiplier is kept constant.

6.5.4 Implications for rain gauge density

The impact of the rain gauge density on parameter posterior distributions was modest. Parameter posteriors for Approaches 1 and 2 show that there was typically little difference between the rain gauge scenarios, while there was almost no difference in the case of Approach 3. This suggests that parameter estimation was robust to the density of the rain gauges. This is an important finding, as the necessary condition for the regionalization of a rainfall-runoff model is that the parameters are insensitive to the choice of the number and locations of rain gauges (Thyer et al., 2009). We acknowledge that these results may be different in different circumstances, such as in a case where rainfall uncertainty has a larger impact on the parameter posterior distributions. This study, however, contradicts the findings of Zeng et al. (2018), who found that parameter posterior distributions vary considerably under different rain gauge densities. However, Zeng et al. (2018) did not propagate rainfall input uncertainty and simply sampled a large number of possible rain gauge combination for a given density and analysed the differences between rain gauge densities in terms of the model parameter posterior distribution. Thus, they ignored a substantial proportion of the uncertainty, which potentially caused model parameter uncertainty to compensate for the unaccounted rainfall uncertainty (Kavetski et al., 2006).

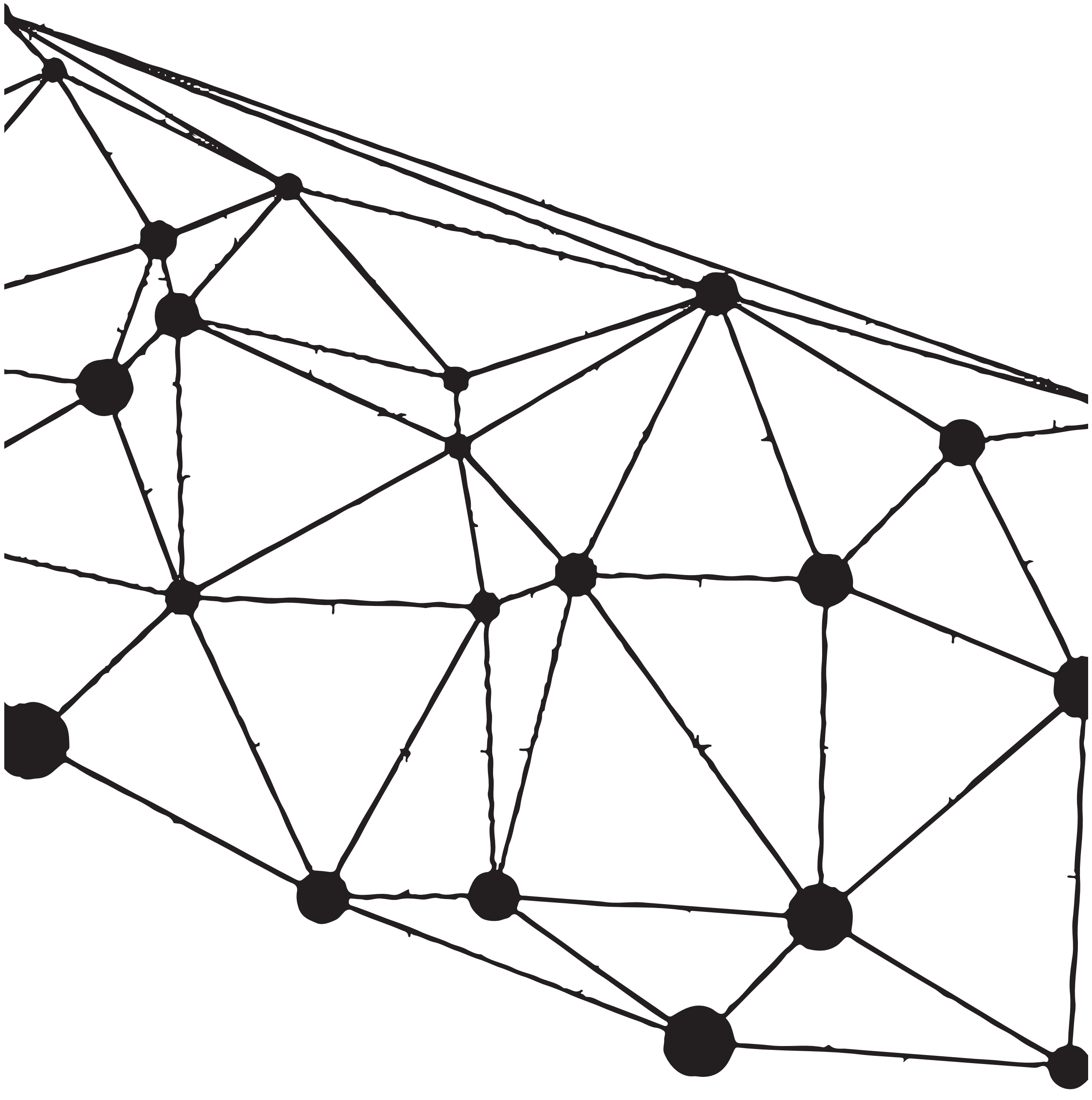
In our experiment, low rain gauge densities already produced accurate model predictions. This particularly applied to Approaches 1 and 2 where only one gauge led to a NSE of 0.43. This threshold is low compared to other studies. Dong et al. (2005) reported that five rain gauges were enough to calibrate the HBV model in a 17,000 km^2 basin in China, using the expected variance of the areal rainfall as a measure of input uncertainty. In a 3234 km^2 catchment in France, Anctil et al. (2006) concluded that ten rain gauges (out of 23) is an absolute minimum to predict discharge using a neural network model. However, the authors did not model the rainfall spatial variation explicitly. Bárdossy and Das (2008) found that the overall model performance worsened radically with an excessive reduction of rain gauges in the upper Neckar catchment of about 4000 km^2. They optimized the rain gauge locations for different rain gauge densities using simulated annealing and kriging with external drift. They showed a significant reduction of the rainfall input variance with increasing density, which paired with a decrease of the discharge prediction error. However, the cited studies did not use Bayesian calibration. In a Bayesian framework, Zeng et al. (2018) found that 10-15 rain gauges were necessary to obtain stable parameter estimates for medium-size sub-basins, but the authors did not propagate input rainfall uncertainty. In our study, the input rainfall uncertainty is estimated using geostatistics and propagated in a Bayesian framework. A fairly accurate estimate of the catchment average 10-day rainfall was obtained using just one or two rain gauges, which explains why a surprisingly small number of rain gauges was enough to calibrate the hydrological model. The rainfall posterior parameters adjust for the missing information using the discharge data. In Approach 3, i.e. when the rainfall is not updated by the discharge data, the model performance was worse than the other approaches in all cases, and particularly for cases with a small number of rain gauges (fewer than 5 rain gauges).

Although using a very small number of rain gauges led to accurate model prediction, using more rain gauges did improve the model predictions. The results of our case study showed that a density larger than five rain gauges led to a marginal improvement of the prediction accuracy. This is equivalent to 1 rain gauge per 340 km^2. It should be noted that this result cannot easily be generalized because it is likely case-dependent. In particular, the surprisingly low rain gauge density is likely related to the 10-day time step that we used in the case study. Aggregating rainfall over 10-day periods automatically leads to a decrease of the rainfall spatial variation. Figure 6.2 shows that the spatial variation of the 10-day average rainfall in the Thur basin is relatively small: the variogram sill was small and the variogram range was large. We emphasize that predicting 10-day average discharge also leads to smoothing and hence will miss peak discharges. For certain applications a smaller time step will be required, although the time step that we used is suitable for many applications, e.g. for total discharge prediction over long time periods.

6.6 Conclusion

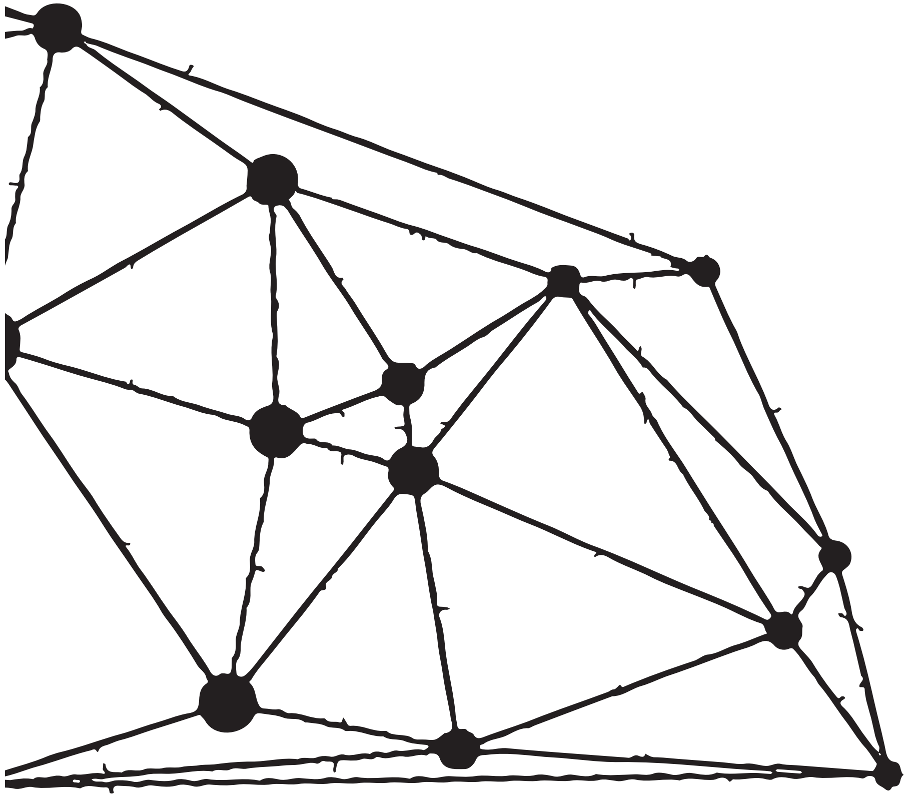
We calibrated the HBV rainfall-runoff model accounting for input, parameter, initial state and model structural uncertainty using a Bayesian framework for a 1700 km^2 basin in Switzerland. Prior input rainfall distributions were derived using a geostatistical approach. We tested several scenarios for incorporating the input uncertainty and assessed the effect of rain gauge density on calibration. The main conclusions are:

- Assumptions regarding the formulation of the model structural uncertainty play a major role in distinguishing between model structural and input uncertainty. Since model structural uncertainty is incorporated explicitly it is unlikely that input uncertainty compensates for deficits in the model structure.
- In our case study, input uncertainty was small compared to model structural and parameter uncertainty.
- Calibrating the rainfall parameters (Approach 1) led to more accurate model performance compared to the case where rainfall uncertainty was not updated using discharge data (Approach 2). The increased dimensionality of the parameter space when calibrating rainfall did not lead to computational intractability.
- Parameter estimates were robust to rain gauge density. This is important, as this enables regionalization of the rainfall-runoff model.
- In our case study, using a single rain gauge did not seriously deteriorate discharge prediction.
- Adding up to five rain gauges improved the model prediction. Adding even more only produced a marginal improvement of the prediction accuracy. For our study area, five rain gauges is equivalent to one rain gauge per 340 km^2.



Chapter 7

Synthesis



7.1 Introduction

In Chapter 1, I argued that there is a discrepancy between research on spatial sampling design optimization and latest developments in geostatistical modelling and mapping. I gave four examples of recent mapping techniques for which little investigation has been dedicated to derive optimal sampling strategies. The sampling design is important, because the map accuracy is partly determined by the number and spatial locations of the units used to calibrate a model for spatial prediction and for the spatial prediction itself. Sampling is also a costly affair. A design can potentially help to save costs if done optimally. In this thesis I made a step towards derivation of optimal designs for novel mapping techniques with case studies on mapping soil and hydrological variables.

In this final Chapter I discuss whether the objectives of this thesis were achieved (Section 7.2). I will also compare the findings of this thesis with existing literature and suggest directions for future research (Section 7.3). Finally, in Section 7.4, I will place this thesis within a more general historical context, provide a personal reflection and give the main conclusion.

7.2 Overview of findings

The overall aim of this thesis was to extend our knowledge on sampling design optimization for recent advances in spatial modelling and prediction. The objective was addressed through four research questions stated in Section 1.3.5, of which the results were presented in Chapters 2 to 6.

Chapter 3 dealt with sampling design optimization using the geostatistical model defined in Chapter 2. For a case study optimizing the locations of rain gauges for mapping rainfall in the North of England, it was shown that spatial prediction benefited from a geostatistical model that includes non-stationarity in the mean and variance, as shown by the likelihood and Akaike Information Criterion statistics. The optimization of the rain gauge network was achieved by spatial simulated annealing. The optimized rain gauge network improved slightly the rainfall mapping accuracy. The accuracy gain was limited because I used a static design (Brus, 2014) for all time steps, while the areas with large prediction uncertainty vary from day to day. The optimized design also showed a specific spatial pattern. It had a fairly uniform spatial distribution but an increased density in areas where the residual variance was generally large. In our case study these areas where in high elevation areas and areas far from the rainfall radar stations. I further tested an optimized design using a reduction of 10% of the total number of rain gauges. The optimized design showed a significant improvement over the original design using all rain gauges. I concluded that at least 10% of the rain gauges could be removed (e.g. to save costs) without loss of mapping accuracy, provided that the rain gauges are placed optimally in the area of interest.

Chapter 4 showed that a spatial coverage design performs poorly for mapping using ordinary kriging because such design lacks information about spatial variation at short distances, which is needed to estimate the variogram parameters. This was tested for a series of variogram parameters of a Matèrn correlation function. An optimized design performed always slightly better, but has several disadvantages. For example, it requires the variogram parameters to be known to define an objective function characterizing the total error, and minimizing this error using optimization algorithms (e.g. Spatial Simulated Annealing). In contrast, a spatial coverage design supplemented by a subset of close-pair units offered accurate results for most combinations of variogram parameters tested. I therefore recommend to use the latter design for designing a geostatistical survey, unless prior knowledge of the variogram is available (e.g. an “average” variogram). If an average variogram is available for the property of interest, it can be used to optimize the design. I further tested the minimum number of sampling units required to make a geostatistical survey worthwhile (i.e. more accurate than a design-based estimate of the mean), and showed that this strongly depends on the degree of spatial correlation of the target variable. I showed that for large values of the variogram effective range and small nugget-to-sill ratios, a sample size of about 20 units is sufficient to make geostatistical analysis more accurate than a design-based estimate. Note that in the latter case, the design-based estimate of the mean was used as the predicted value at any point in the area.

The case study of Chapter 5 demonstrated, for mapping topsoil organic carbon using a machine learning methods (viz. random forest), that an optimized design was up to 10% more accurate in term of MSE than other common sampling designs, but can only be obtained when subsampling an existing dataset with known values of the target variable at all locations. In practice we do not have this information. By comparing the mean square error (MSE) of the maps obtained by an optimized design with those obtained by common designs, it was shown that optimizing a design in terms of MSE is not always worthwhile. When the sample size increases, the maps produced by the different designs converge to similar accuracy values. In our case study on large scale soil organic carbon mapping, a sampling density greater than 1 sampling unit per 4000 km^2 decreased markedly the difference in term of average MSE between designs. A design optimized for the mean squared shortest standardized distance in the feature space presented the closest match with the optimized design in terms of MSE. By analysing the distribution of the sampling locations in both geographic and feature space, I further showed that the optimized design is not spread in the geographic space, but seems to be spread somewhat uniformly in the feature space, and especially in the most important covariates of the machine learning model. It is however difficult to draw further conclusions because of the complex spread of the units in feature space. Further research is needed in this direction.

In Chapter 6 I integrated geostatistics and Bayesian calibration to analyse the effect of the rain gauge density on rainfall-runoff model prediction accuracy, which uses interpolated maps of rainfall as input. The Bayesian calibration enabled to capture model input, initial state, parameter and structural uncertainty, while also taking uncertainties in the output measurements into account. In a case study predicting river discharge using a rainfall-runoff model and maps of rainfall as input, a single rain gauge was sufficient to obtain accurate model parameter calibration and discharge predictions. Adding up to five rain gauges improved the model prediction. Adding even more only produced a marginal improvement of the prediction accuracy. Calibrating the rainfall time series as additional parameters led to more accurate model performance compared to the case where rainfall uncertainty was not updated using measurements of the discharge. Furthermore, it was demonstrated for a case study that model parameter uncertainty was the main contributor to the posterior discharge uncertainty and that input uncertainty had a relatively small contribution. However, the study also showed that Bayesian calibration of rainfall has serious computational disadvantages. In particular, calibrating a large number of rainfall input parameters remains a serious challenge.

7.3 Future research

This thesis has shown the possibility to optimize designs of recent mapping techniques.

Chapter 2 showed a substantial improvement of the non-stationary variance model over the stationary variance model. The non-stationary variance model of Chapter 2 is also an improvement of the model detailed in Lark (2009) since the latter study was limited to the use of the spatial coordinates as explanatory variable for the variance. The strongest point of the non-stationary variance model developed in Chapter 2 is the explicit modelling of the variance by environmental covariates. However, the variance was modelled as a linear combination of covariates (Chapters 2 and 3) which may pose some restrictions. The variance of soil and hydrological variables have often complex, non-linear relationships with environmental covariates. The regression used in Chapters 2 and 3 could surely be improved by increasing the number of covariates and modelling non-linear processes, as in the random forest model (Hengl et al., 2018). The geostatistics community has not yet explored mapping based on non-linear regression of the variance. This might be a valuable extension of Chapters 2 and 3 of this thesis. However, a more complex variance component may conflict with the rigid need to avoid near-singularity when inverting the covariance matrix. Several solutions exist to avoid the near-singularity problem and some have been tested in this thesis. In the simplest case, one may constrain the estimation of the standard deviation parameters so that their combination does not provoke near-singularity of the covariance matrix (Marchant et al., 2009), as evaluated by a criterion, e.g. reciprocal condition number (Golub and Van Loan, 2012). Other solutions that have been mentioned in this thesis are the use of the generalized inverse (Sen and Srivastava, 2012) or the modelling of a log-transformed standard deviation (e.g. as in Pintore and Holmes, 2004). This might solve the problem in some cases but it requires further investigations.

The case study of Chapter 3, mapping rainfall using radar maps as covariates in the spatial trend, showed an example on how the units of the optimized design are dispersed in areas of large radar uncertainty. This can be applied in other fields too, where the use of remote or proximal sensing technologies is becoming increasingly important. For example, soil moisture spatial variability is partly governed by land cover, whose maps are easy to obtain using remote sensing images. It is likely that soil moisture mapping benefits from a model that accounts for deviations from restrictive stationarity and isotropy assumptions. A recent contribution made by Kathuria et al. (2019) shows that this is indeed the case. In this study a land cover map was derived by remote sensing images and used as a covariate in a geostatistical model whose variance and correlation structures are flexible. The improved maps of soil moisture can be used to optimize the locations of ground-based measurements, which in turn are crucial information for calibrating and validating satellite soil moisture retrieval (Reichle et al., 2004).

This thesis has shown the added value of close-pair locations supplementing a spatial coverage when designing a geostatistical survey, and compared this design to an optimized design. The close-pairs were chosen randomly at a fixed short distance from an existing sampling location. We could further test whether this distance relates to the size of the test area and to the variable of interest. We used an arbitrary distance, only imposing that it is short relative to the extent of the study area. This subjective decision is rather common, e.g. Atteia et al. (1994); Cattle et al. (2002). as no specific study presents rules to choose this distance, but this clearly requires further investigation so as to generate rules and guidance for designing efficient geostatistical surveys. Chapter 4 builds on the study made by Lark and Marchant (2018) who showed that it is best to supplement the spatial coverage design with 10% of additional sampling locations to provide some closely spaced pairs. Using a proportion of the total sample size as close-pairs seems to work well on small sample sizes (tested for fewer than 150 sampling units), but has not been tested for larger sample size. This is perhaps not optimal to fix the percentage of close-pair units and we may rather want to find an optimal fixed number instead. The analysis of Lark and Marchant (2018) could be repeated for larger sample size, by testing the threshold in which an absolute number of close-pairs unit does not provide improvement on the prediction error variance. However, when taking a too small number of close-pairs, there is a risk that the close-pairs are located in a-typical conditions, given that we assume second-order stationarity. A large number of close-pairs ensure that this is avoided, for little additional effort in field collection.

There is also room for further research to support sampling for kriging with external drift, in presence of variogram uncertainty. Estimating additional (linear) regression coefficients introduces additional considerations for sampling design. In a linear model, estimation of the coefficients benefit from a sample that is clustered around few areas, corresponding to the extremes of the covariates. We can speculate that a spatial coverage design supplemented by some close-pairs locations is also efficient, albeit necessarily sub-optimal, to support estimation of the additional trend parameters. The spatial coverage sample would ensure reliable estimation of the trend parameters (Brus, 2019) and the close-pairs would ensure estimation of the covariance parameters (Chapter 4).

Sampling optimization for mapping using machine learning techniques has been barely investigated by previous research. In Chapter 5 I tested several sampling strategies for mapping using random forest (RF). RF currently is certainly the most popular machine learning algorithm, which use for soil mapping has been recently formalized by the publication of Hengl et al. (2018). But RF is not the only machine learning technique available for mapping. Support vector machines (Ballabio, 2009), (artificial) neural networks (Behrens et al., 2005; Were et al., 2015), decision trees (Moore et al., 1991) and gradient boosting (Nussbaum et al., 2017) are all linking the variable of interest and the environmental covariates in a non-linear way. In Chapter 5 I assumed that an optimal design for RF would similarly be optimal for other tree-based models. This is a realistic assumption given that RF and other tree based methods share the same basic structure and splitting metrics. Sampling for other machine learning techniques still needs to be explored in further research. I would speculate that the spread of the sampling units in feature space remains important, but that other considerations (e.g. selecting support vectors for mapping with support vector machine) may outweight the uniform spread.

During the course of this thesis, new machine learning techniques, called deep learning (Behrens et al., 2018a), emerged as a valuable tool for spatial analysis (Wadoux et al., 2019b; Padarian et al., 2019). In particular, convolutional neural networks offer attractive features such as contextual mapping and flexibility in modifying the objective function (Wadoux, 2019). Deep learning models are "data-hungry" algorithms. Their use for mapping with scarce data and their optimal sampling require investigation. We are now aware that random forest techniques benefit from a spread of the units in the feature rather than in the geographic space. But the spread in the feature space is complex and not simply uniform on the whole set of covariates. One of the reason of this complexity is that the feature space is multi-dimensional. We are still unsure what makes a good sampling design for mapping using machine learning techniques. To discover it, we must first reveal the characteristics of an optimized design, such as the one derived in Chapter 5, so that future research can generate rules and obtain simple designs able to resemble optimal ones. I believe this will be a major research topic for sampling design in the forthcoming years.

Sampling design optimization becomes more complex when it is used to derive a map used as input for a model whose output is the main interest. This was done by integrating geostatistics for mapping rainfall and Bayesian calibration of a hydrological model for predicting river discharge (Chapter 6). The work presented in Chapter 6 differs in some ways from Renard et al. (2011). The main difference is the use of geostatistics to define a prior distribution for the rainfall, the latter being used for the integration of the rainfall of each time step as an additional parameter to be calibrated. This was not the case in Renard et al. (2011), which used geostatistics to model the uncertainty of the rainfall maps, and propagated these to the discharge output using Bayesian calibration. In Chapter 6 I optimized the rain gauge locations using a criterion related to the rainfall map accuracy, which is used as input in the hydrological model. There is room for further research to optimize the rain gauge locations for a criterion that minimizes directly the hydrological model prediction error Anctil et al. (e.g. as in 2006). This has not yet been investigated in the literature using Bayesian calibration, but will inevitably cause an important increase of the computational load because for each rain gauge network configuration tested, several thousands of Markov Chain Monte Carlo runs have to be computed.

A major limitation of Chapter 6 is the number of parameter to calibrate in Approach 1, when the rainfall is updated by the discharge measurements. Further investigations may show whether it is more efficient to reduce the number of parameters (e.g. using storm event multipliers Kavetski et al., 2006; Vrugt et al., 2008) or to rather find strategies to explore the multi-dimensional and correlated parameter space. In our case study this was not a problem, but Approach 1 will become cumbersome to model hydrological processes at fine temporal resolution (such as daily or finer) or for long time periods (several decades). Finally, I emphasize that in our case study, one single rain gauge was sufficient to obtain reliable prediction of the discharge. This is because rainfall variability was small for the 10-day aggregation period, and that in consequence rainfall uncertainty was modest compared to parameter and model structural uncertainty. This contradicts clearly several previous studies on rainfall input uncertainty in hydrological modelling using Bayesian calibration (e.g. Kavetski et al., 2006; Thyer et al., 2009; Zeng et al., 2018). These studies showed that rainfall input was the main contributor to the joint posterior uncertainty. There is need to bring the study of Chapter 6 one step further and test it on a large number of catchments so as to gain insight into the importance of rainfall uncertainty for different case studies. Recent studies (e.g. Melsen et al., 2018) have made a step forward towards generalization of hydrological modelling to a large number of catchments so as to explore a range of possible scenarios.

7.4 Reflection on sampling design optimization practices

Historical perspectives

Research on sampling design optimization for mapping began in the 1980's and 1990's with numerous studies showing the importance of the distribution of sampling locations in geographic space when designing geostatistical surveys. McBratney et al. (1981) showed that an equilateral, triangular grid provides the minimum estimation variance for an isotropic variogram and ordinary kriging prediction at point locations, and that using a square sampling grid presented a small loss of precision but is more convenient to use. Incorporating a drift into the kriging equations, Olea (1984) tested various sampling strategies for geostatistical modelling of the water table using a known isotropic variogram. This study recommended the use of a regular hexagonal pattern, or alternatively a regular square pattern. Back then, the authors warned against the use of random sampling strategies, which need several times more sampling units than a hexagonal pattern to achieve the same mapping accuracy. Yfantis et al. (1987) promoted the use of an equilateral triangle design, which he proved useful for estimating the variogram and mapping. The studies cited, and others (Villeneuve et al., 1979; Hughes and Lettenmaier, 1981; Flatman and Yfantis, 1984; Odeh et al., 1990), stressed the importance of a uniform spread of the sampling units in the geographic space for mapping using ordinary kriging.

This was further investigated in the 1990's by the work of Van Groenigen et al. (1999), who used simulated annealing, extended for spatial optimization purposes, to optimize sampling designs for mapping using ordinary kriging. The optimized designs using the mean or maximum of the kriging variance as minimization criterion outperformed triangular grid sampling strategies. The authors also showed that the optimized sample has a fairly uniform geographic distribution, with sampling units slightly pushed towards the boundary of the study area. Another contribution was made by Brus et al. (2007), who showed that the mean kriging variance of a spatial coverage sample, obtained by minimizing the mean of the squared shortest distance (MSSD) by the fast k-means algorithm as proposed by Brus et al. (1999), was about equal to the mean kriging variance directly minimized by SSA.

We are now aware that more aspects need to be accounted for when optimizing a survey. Most studies in the 1980's and 1990's investigated the ordinary kriging case, with few exceptions (e.g. Yfantis and Flatman, 1988). With time, models became more complex and mapping techniques became more evolved. We are now making increased use of environmental covariates as auxiliary information to our models and we use machine learning techniques. The complexity increased and so have the sampling strategies for modelling and mapping.

The new millennium: more complex models

A change has been made from 2000 onward, where studies analysed the effect of including a trend into the kriging variance minimization. Hengl et al. (2003) showed that a design optimal to estimate the regression coefficient of a linear model is spatially clustered. Lesch et al. (1995) showed how the calibration of a linear model can lead to strong spatial clustering of the units as it assumes independent residuals. The authors proposed a sampling design (response surface sampling) to avoid spatial clustering and account for possible violations of the residual independence assumption. Later, Brus and Heuvelink (2007) investigated on optimal design for geostatistical mapping using kriging with a linear trend. The study showed the importance of simultaneous optimization of the sample in both geographic and feature space. Additional contributions have accounted for the estimated variogram uncertainty into the sampling design. Lark (2002) and Zhu and Stein (2005) proposed a criterion to model the covariance structure uncertainty, which can be minimized using a spatial optimization algorithm (Van Groenigen et al., 1999). These studies reflect the constant need for adaptations of the input sampling design when models change. In particular, there is an important research effort towards the improvement for modelling the stochastic residuals and the use of non-linear, data-driven machine learning techniques for mapping. For example, Pozdnoukhov and Kanevski (2006) and later Tuia et al. (2013) optimized a network for mapping using support vector machine. They specifically aimed at minimizing the "risk" of selecting new sampling units that do not have a valuable contribution to the model (by becoming support vectors). To the best of my knowledge, they were the first to investigate on sampling designs for mapping with a machine learning technique.

Deriving an optimized design becomes even more complex when the map is not the main goal, but simply one of the inputs of a model whose output is the main interest. This is a frequent exercise in studies in hydrology, soil science, climatology, land surface and reservoir modelling (Mishra and Coulibaly, 2009). Sampling for an input that is used as input in a model has been tackled in the hydrology literature in various forms. Back in the 80's, Troutman (1983) and Krajewski et al. (1991) have shown that the spatial variability of rainfall, and therefore the sampling design, severely affects the storm runoff prediction accuracy. Troutman (1983) used the Green-Ampt soil-water infiltration model and tested the effect of the rainfall-induced bias on the estimation of physical parameters. This has been taken one step further in the last two decades. St-Hilaire et al. (2003) compared two network density scenarios: one dense and another sparse. The network was used to derive daily average rainfall maps, which were used as input into the HSAMI model to simulate runoff. The study showed the importance of a dense network to capture peak flows and summer flood events. Later, Dong et al. (2005) investigated on optimal number of rain gauges on hydrological model prediction using the HBV model. They showed that the discharge prediction accuracy increases hyperbolically by adding more rain gauges, but levels off after only five rain gauges. This is further explored by Anctil et al. (2006) who also analysed the effect of the sampling locations on runoff forecasting. A genetic algorithm was used to optimize the rain gauge sampling configuration so as to improve the forecasting performance. This study showed that if placed optimally, using fewer rain gauges can lead to better forecast than when all available rain gauges are used. Note that research on sampling design for producing maps which are used as input to a model is a common problem which has also been tackled in other fields, such as in climatology. Examples of such studies are found in PaiMazumder and Mölders (2009), Mauger et al. (2013) or Yang et al. (2014).

Is there a single best optimal design?

It was shown that, for a wide range of models and mapping scenarios, using an optimized design always makes a significant difference in terms of mapping accuracy. On the basis of the literature and the various cases treated in this thesis, I can however not conclude that there is a single best optimal design. It is very much case dependent. It depends, among others, on: (i) the assumed model of spatial variation, and therefore the mapping technique because the mapping method depends on the model; (ii) the assumption whether we need or need not estimate the model parameters from the data. For instance, this is the case for the variogram parameters, if we assume these known then we do not need close-pairs units. In the same way, if linear regression model coefficients are known then there is no need to take the feature space into account; (iii) the criterion that is used to optimize the sampling configuration. The most common and perhaps simplest criterion to minimize is the spatially averaged or maximum prediction error variance, but I also discussed in Chapter 4 and 6 that other criteria may be used, such as variogram parameter uncertainty or in a more complex case the prediction error variance based on a model taking as input maps for which the sample is optimized. Recall from the Introduction that in practice, we may also include some additional constraints in the optimization, such as the cost of sampling and accessibility. This may all have an impact on the optimal design. Being conscious that there is no single best optimal design implies that for each case study, investigations must be made given the model, the assumptions and the objectives. The methods developed in this thesis and the literature provide useful information to derive an appropriate design for a given case study. Being aware that there is no single best optimal design also means that it is difficult to tell at the beginning of a project which sampling design one must adopt. This is yet a frequent problem.

Which sampling strategy to adopt at a start of a project?

Chapters 3 to 6 and the literature have shown that optimizing the sampling design is always preferable because it leads to smaller prediction variance (Chapter 3 and 4), prediction error (Chapter 5) or model output variance (Chapter 6). We know that to apply an optimization we must satisfy the three conditions detailed in the previous paragraph, i.e. (i) we know the model of spatial variation, (ii) we know which parameters to estimate and (iii) we know which criterion to optimize. The choice of the criterion is major and it has a serious effect on the optimal design. In practice we may not know the three conditions. This is typically the case at the start of a project when no previous data or expertise are available but we need to design a survey. In is case, it is sensible to use some rules of thumb to design a survey for mapping. I provide some below, based on this PhD-research.

At the higher level, one may wish to decide to use either a design-based or a model-based approach for sampling. A design-based approach is appropriate to estimate global or local quantities (e.g. global or local mean and variance) such as the regional mean or mean values within subregions (Brus and De Gruijter, 1997). In this thesis I was interested in mapping and so I did not analyse design-based sampling strategies. Brus and De Gruijter (1997) and Webster and Lark (2012) showed how the design-based mean (for a region, or subregion) may be treated as a spatial prediction. While it is acknowledged that the design-based approach is appropriate for estimating quantities, there is still debate on the relevance to estimate the value of a specific location using the regional mean or mean value within subregions, and on how to estimate accuracy measures at point. See for example the discussion in Laslett (1997). For mapping, I mentioned in the Introduction that a model-based sampling strategy is preferable. A simple random sampling provides pairs of units at many separation distances, this is favorable for a variogram estimation, but inevitably sub-optimal because a criterion related to the variogram parameter uncertainty is not minimized. For mapping using an (unknown) geostatistical model, a robust strategy is to use a spatial coverage sampling design supplemented by a sample of 10% of the total sample size at short distance from the existing locations (Chapter 4). Alternatively, databases of variogram parameters will help in deriving an average variogram which can be used to optimize a design of a given size. These simple rules apply for mapping using a geostatistical model. Close-pairs units are useful for estimating the variogram parameters while a regular pattern is appropriate for spatial interpolation. Using covariates and multiple linear regression, a rule of thumb is that the sample size must be 20 times larger than the number of covariates (Franklin, 2010), and that the units should be clustered at the minimum and maximum of the covariate values. For other mapping techniques, such as machine learning, it is still difficult to provide guidance on sampling without a reconnaissance survey. A robust recommendation is to avoid conditioned Latin Hypercube sampling (Chapter 5) and to use a feature space coverage sampling design instead.

Conclusion

All the aspects of sampling designs mentioned in this thesis cannot be analysed in separation. The previous sections showed that including covariates, using machine learning, including a non-stationary variance, need for estimation of variogram parameters, all influence what makes an optimal sampling design. I believe that this thesis made a substantial contribution to adjusting spatial sampling design optimization to recent spatial modelling developments, but I also believe that we are just at the beginning of this specific field of science. There is a large increase in complexity of techniques and models used for mapping. We make more use of spatially explicit information, such as remote sensing imagery, and measurements are increasingly inferred rather than measured. In the last decade, techniques for mapping became more data-driven and non-linear, increasing *de facto* the complexity of the sampling designs that should accompany such developments. In fact, common spatial sampling designs seem outdated for such techniques (Chapter 5). While the characteristics of sampling designs for data-driven techniques (e.g. machine learning) are not yet fully discovered, new mapping techniques such as deep learning appeared. Such techniques are known to be "data-hungry", which conflicts with the generally small number of sampling units available for mapping. We must make an efficient use of the available data. New data collection approaches must be justified, in particular to funding bodies. Thus, research on spatial sampling design and optimizing the sampling density is highly relevant in the modern spatial modelling world. It must not be considered as an end in itself but as a tool to help obtaining knowledge about the new mapping techniques and use them optimally. Because sampling is the basis of mapping and has a large impact on cost and accuracy, this research field will remain as important as ever in geostatistics and spatial modelling.

References

Abramowitz, M., Stegun, I. A., 1972. Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables,. Vol. 55. Dover publications, New York, USA.

Ajami, N. K., Duan, Q., Sorooshian, S., 2007. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resources Research 43, W01403.

Akaike, H., 2011. Akaike Information Criterion. In: International Encyclopedia of Statistical Science. Springer-Verlag, Berlin-Heidelberg, Germany, pp. 25–25.

Anctil, F., Lauzon, N., Andréassian, V., Oudin, L., Perrin, C., 2006. Improvement of rainfall-runoff forecasts through mean areal rainfall optimization. Journal of Hydrology 328 (3-4), 717–725.

Andréassian, V., Lerat, J., Loumagne, C., Mathevet, T., Michel, C., Oudin, L., Perrin, C., 2007. What is really undermining hydrologic science today? Hydrological Processes 21 (20), 2819–2822.

Angelini, M. E., Heuvelink, G. B. M., Kempen, B., 2017. Multivariate mapping of soil with structural equation modelling. European Journal of Soil Science 68 (5), 575–591.

Ardia, D., Mullen, K. M., Peterson, B. G., Ulrich, J., 2015. DEoptim: Differential Evolution in R. Version 2.2-3. Accessed 19.10.2015.
URL http://CRAN.R-project.org/package=DEoptim

Arya, S., Mount, D., Kemp, S. E., Jefferis, G., 2017. RANN: Fast Nearest Neighbour Search (Wraps ANN Library) Using L2 Metric. R package version 2.5.1. Accessed 21.10.2016.
URL https://CRAN.R-project.org/package=RANN

Asadollahfardi, G., 2015. Selection of water quality monitoring stations. In: Water Quality Management. Springer, Berlin, Germany, pp. 5–20.

Atkinson, P. M., LLoyd, C. D., 2007. Non-stationary variogram models for geostatistical sampling optimisation: An empirical investigation using elevation data. Computers & Geosciences 33 (10), 1285–1300.

Atteia, O., Dubois, J.-P., Webster, R., 1994. Geostatistical analysis of soil contamination in the Swiss Jura. Environmental Pollution 86 (3), 315–327.

Baddeley, A., Turner, R., 2005. spatstat: An R Package for Analyzing Spatial Point Patterns. Journal of Statistical Software 12 (6), 1–42.

Ballabio, C., 2009. Spatial prediction of soil properties in temperate mountain regions using support vector regression. Geoderma 151 (3-4), 338–350.

Barca, E., Castrignanò, A., Buttafuoco, G., De Benedetto, D., Passarella, G., 2015. Integration of electromagnetic induction sensor data in soil sampling scheme optimization using simulated annealing. Environmental Monitoring and Assessment 187 (7), 422.

Barca, E., Passarella, G., Uricchio, V., 2008. Optimal extension of the rain gauge monitoring network of the Apulian Regional Consortium for Crop Protection. Environmental Monitoring and Assessment 145 (1-3), 375–386.

Bárdossy, A., Das, T., 2008. Influence of rainfall observation network on model calibration and application. Hydrology and Earth System Sciences 12 (1), 77–89.

Bastin, G., Lorent, B., Duque, C., Gevers, M., 1984. Optimal estimation of the average areal rainfall and optimal selection of rain gauge locations. Water Resources Research 20 (4), 463–470.

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 (3), 289–294.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. Journal of Plant Nutrition and Soil Science 168 (1), 21–33.

Behrens, T., Schmidt, K., MacMillan, R. A., Viscarra Rossel, R. A., 2018a. Multi-scale digital soil mapping with deep learning. Scientific reports 8 (1), 15244–15244.

Behrens, T., Schmidt, K., Viscarra Rossel, R. A., Gries, P., Scholten, T., MacMillan, R. A., 2018b. Spatial modelling with Euclidean distance fields and machine learning. European Journal of Soil Science 69 (5), 757–770.

Behzadian, K., Kapelan, Z., Savic, D., Ardeshir, A., 2009. Stochastic sampling design using a multi-objective genetic algorithm and adaptive neural networks. Environmental Modelling & Software 24 (4), 530–541.

Beven, K., 2006. On undermining the science? Hydrological Processes 20 (14), 3141–3146.

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. Journal of Hydrology 249 (1), 11–29.

Beven, K. J., Freer, J., Hankin, B., Schulz, K., 2000. The use of generalised likelihood measures for uncertainty estimation in high order models of environmental systems. In: Nonlinear and Nonstationary Signal Processing. CUP, Cambridge, UK, pp. 115–151.

Bishop, T. F. A., McBratney, A. B., 2001. A comparison of prediction methods for the creation of field-extent soil property maps. Geoderma 103 (1), 149–160.

Breiman, L., 1996. Bagging predictors. Machine Learning 24 (2), 123–140.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breiman, L., 2017. Classification and Regression Trees. Routledge, New York, USA.

Brooks, S., Gelman, A., Jones, G., Meng, X.-L., 2011. Handbook of Markov Chain Monte Carlo. CRC press, Boca Raton, USA.

Brown, J. D., Heuvelink, G. B. M., 2005. Assessing uncertainty propagation through physically based models of soil water flow and solute transport. In: Encyclopedia of hydrological sciences. John Wiley & Sons, New-York, USA, pp. 1181–1195.

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: Digital Soil Mapping. Springer, Berlin, Germany, pp. 67–75.

Brus, D. J., 2014. Statistical sampling approaches for soil monitoring. European Journal of Soil Science 65 (6), 779–791.

Brus, D. J., 2019. Sampling for digital soil mapping: A tutorial supported by R scripts. Geoderma 338, 464–480.

Brus, D. J., De Gruijter, J. J., 1993. Design-based versus model-based estimates of spatial means: Theory and application in environmental soil science. Environmetrics 4 (2), 123–152.

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 (1-2), 1–44.

Brus, D. J., De Gruijter, J. J., Breeuwsma, A., 1992. Strategies for updating soil survey information: a case study to estimate phosphate sorption characteristics. Journal of Soil Science 43 (3), 567–581.

Brus, D. J., De Gruijter, J. J., Van Groenigen, J. W., 2007. Designing spatial coverage samples using the k-means clustering algorithm. Developments in Soil Science 31, 183–192.

Brus, D. J., Heuvelink, G. B. M., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138 (1-2), 86–95.

Brus, D. J., Spätjens, L. E. E. M., De Gruijter, J. J., 1999. A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. Geoderma 89 (1-2), 129–148.

Brus, D. J., Yang, L., Zhu, A.-X., 2019. Accounting for differences in costs among sampling locations in optimal stratification. European Journal of Soil Science 70 (1), 200–212.

Brus, D. J., Yang, R.-M., Zhang, G.-L., 2016. Three-dimensional geostatistical modeling of soil organic carbon: A case study in the Qilian mountains, China. Catena 141, 46–55.

Castro-Franco, M., Costa, J. L., Peralta, N., Aparicio, V., 2015. Prediction of soil properties at farm scale using a model-based soil sampling scheme and random forest. Soil science 180 (2), 74–85.

Cattle, J. A., McBratney, A. B., Minasny, B., 2002. Kriging method evaluation for assessing the spatial distribution of urban soil lead contamination. Journal of Environmental Quality 31 (5), 1576–1588.

Cecinati, F., Rico-Ramirez, M. A., Heuvelink, G. B. M., Han, D., 2017. Representing radar rainfall uncertainty with ensembles based on a time-variant geostatistical error modelling approach. Journal of Hydrology 548, 391–405.

Changnon, S. A., et al., 1980. Rainfall prediction-measurement systems and rainfall design information for urban areas. Tech. rep., Illinois State Water Survey, USA.

Cochran, W. G., 1977. Sampling Techniques, 3rd Edition. John Wiley & Sons, New York, USA.

Cressie, N., 2015. Statistics for spatial data. John Wiley & Sons, New York, USA.

Cressie, N., Hawkins, D. M., 1980. Robust estimation of the variogram: I. Journal of the International Association for Mathematical Geology 12 (2), 115–125.

Cui, H., Stein, A., Myers, D. E., 1995. Extension of spatial information, Bayesian kriging and updating of prior variogram parameters. Environmetrics 6 (4), 373–384.

Davis, M. W., 1987. Production of conditional simulations via the LU triangular decomposition of the covariance matrix. Mathematical Geology 19 (2), 91–98.

De Gruijter, J. J., Brus, D. J., Bierkens, M. F. P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer Science & Business Media, Dordrecht, NL.

Deb, K., Mohan, M., Mishra, S., 2003. Towards a quick computation of well-spread pareto-optimal solutions. In: International Conference on Evolutionary Multi-Criterion Optimization. Springer, New York, USA, pp. 222–236.

Delhomme, J.-P., 1978. Kriging in the hydrosciences. Advances in Water Resources 1 (5), 251–266.

Deutsch, C., 1997. Direct assessment of local accuracy and precision. In: Geostatistics Wollongong'96. E. Y. Baafi & N. A. Schofield (Eds.). Vol. 1. Dordrecht: Kluwer Academic, NL, pp. 115–125.

Di Baldassarre, G., Montanari, A., 2009. Uncertainty in river discharge observations: a quantitative analysis. Hydrology and Earth System Sciences 13 (6), 913.

Diggle, P. J., Ribeiro, P. J., 2007a. Geostatistical design. In: Model-Based Geostatistics. Springer, New York, USA, pp. 199–212.

Diggle, P. J., Ribeiro, P. J., 2007b. Model based Geostatistics. Springer Series in Statistics, New York, USA.

Domenech, M. B., Castro-Franco, M., Costa, J. L., Amiotti, N. M., 2017. Sampling scheme optimization to map soil depth to petrocalcic horizon at field scale. Geoderma 290, 75–82.

Dong, X., Dohmen-Janssen, C. M., Booij, M. J., 2005. Appropriate spatial sampling of rainfall or flow simulation/Échantillonnage spatial de la pluie approprié pour la simulation d'écoulements. Hydrological Sciences Journal 50 (2), 279–298.

Dowd, P. A., 1982. Lognormal kriging-the general case. Journal of the International Association for Mathematical Geology 14 (5), 475–499.

Durbin, J., Koopman, S. J., 2012. Time Series Analysis by State Space Methods. Vol. 38. OUP, Oxford, UK.

Ehrgott, M., 2005. Multicriteria optimization. Vol. 491. Springer Science & Business Media, Berlin, Germany.

Engeland, K., Xu, C.-Y., Gottschalk, L., 2005. Assessing uncertainties in a conceptual water balance model using Bayesian methodology/Estimation bayésienne des incertitudes au sein d'une modélisation conceptuelle de bilan hydrologique. Hydrological Sciences Journal 50 (1), 45–63.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., et al., 2007. The Shuttle Radar Topography Mission. Reviews of Geophysics 45 (2), RG2004.

Flatman, G. T., Yfantis, A. A., 1984. Geostatistical strategy for soil sampling: the survey and the census. Environmental Monitoring and Assessment 4 (4), 335–349.

Fouedjio, F., 2017. Second-order non-stationary modeling approaches for univariate geostatistical data. Stochastic Environmental Research and Risk Assessment 31 (8), 1887–1906.

Franklin, J., 2010. Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge, UK.

Frei, C., 2014. Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances. International Journal of Climatology 34 (5), 1585–1605.

Fuentes, M., Reich, B., Lee, G., 2008. Spatial-temporal mesoscale modeling of rainfall intensity using gage and radar data. The Annals of Applied Statistics 2 (4), 1148–1169.

Gallego, J., Delincé, J., 2010. The European land use and cover area-frame statistical survey. In: Agricultural Survey Methods. John Wiley & Sons, Chichester, UK, pp. 149–168.

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 (1), 92–110.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., 2014. Bayesian Data Analysis. Vol. 2. CRC press, Boca Raton, USA.

Golub, G. H., Van Loan, C. F., 2012. The sensitivity of square systems. In: Matrix Computations. Vol. 3. JHU Press, Baltimore, USA, pp. 87–93.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, Oxford, UK.

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma 103 (1), 3–26.

Goudenhoofdt, E., Delobbe, L., 2009. Evaluation of radar-gauge merging methods for quantitative precipitation estimates. Hydrology and Earth System Sciences 13 (2), 195–203.

Gräler, B., Pebesma, E., Heuvelink, G. B. M., 2016. Spatio-temporal interpolation using gstat. R Journal 8 (1), 204–218.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro

Colorado Island—Digital soil mapping using Random Forests analysis. Geoderma 146 (1-2), 102–113.

Gyasi-Agyei, Y., 2016. Assessment of radar-based locally varying anisotropy on daily rainfall interpolation. Hydrological Sciences Journal 61 (10), 1890–1902.

Habib, E., Krajewski, W. F., Kruger, A., 2001. Sampling errors of tipping-bucket rain gauge measurements. Journal of Hydrologic Engineering 6 (2), 159–166.

Hamm, N. A. S., Atkinson, P. M., Milton, E. J., 2012. A per-pixel, non-stationary mixed model for empirical line atmospheric correction in remote sensing. Remote Sensing of Environment 124, 666–678.

Han, S., Coulibaly, P., 2017. Bayesian flood forecasting methods: A review. Journal of Hydrology 551, 340–351.

Harrison, D. L., Scovell, R. W., Kitchen, M., 2009. High-resolution precipitation estimates for hydrological uses. In: Proceedings of the Institution of Civil Engineers-Water Management. Vol. 162. Thomas Telford Ltd, Scotland, UK, pp. 125–135.

Hartigan, J. A., Wong, M. A., 1979. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28 (1), 100–108.

Haskard, K. A., Lark, R. M., 2009. Modelling non-stationary variance of soil properties by tempering an empirical spectrum. Geoderma 153 (1), 18–28.

Heesterman, A. R. G., 1983. Block-equations and matrix-inversion. In: Matrices and Simplex Algorithms: A Textbook in Mathematical Programming and Its Associated Mathematical Topics. Springer, Dordrecht, NL, Dordrecht, pp. 33–61.

Heistermann, M., Kneis, D., 2011. Benchmarking quantitative precipitation estimation by conceptual rainfall-runoff modeling. Water Resources Research 47 (6), w06514.

Henderson, B. L., Bui, E. N., Moran, C. J., Simon, D. A. P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124 (3-4), 383–398.

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., et al., 2017. Soilgrids250m: Global gridded soil information based on machine learning. PLoS one 12 (2), e0169748.

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., de Jesus, J. M., Tamene, L., et al., 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. PloS one 10 (6), e0125814.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518.

Hengl, T., Rossiter, D. G., Stein, A., 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Soil Research 41 (8), 1403–1422.

Hengl, T., Toomanian, N., 2006. Maps are not what they seem: representing uncertainty in soil-property maps. In: Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 5 – 7 July 2006, Lisboa. pp. 805–813.

Heuvelink, G. B. M., 1998. Error Propagation in Environmental Modelling with GIS. CRC Press, London, UK.

Heuvelink, G. B. M., 2014. Uncertainty quantification of GlobalSoilMap products. In: GlobalSoilMap: Basis of the Global Spatial Soil Information System. Proceedings of 1st GlobalSoilMap Conference. Taylor & Francis Group, London, UK, pp. 335–340.

Heuvelink, G. B. M., Brus, D. J., de Gruijter, J. J., 2006. Optimization of sample configurations for digital mapping of soil properties with universal kriging. In: Developments in Soil Science. Vol. 31. Elsevier, Amsterdam, NL, pp. 137–151.

Heuvelink, G. B. M., Griffith, D. A., Hengl, T., Melles, S. J., 2012. Sampling design optimization for space-time kriging. In: Spatio-Temporal Design. John Wiley & Sons, New-York, USA, pp. 207–230.

Heuvelink, G. B. M., Jiang, Z., De Bruin, S., Twenhöfel, C. J. W., 2010. Optimization of mobile radioactivity monitoring networks. International Journal of Geographical Information Science 24 (3), 365–382.

Heuvelink, G. B. M., Pebesma, E., Gräler, B., 2015. Space-time geostatistics. In: Encyclopedia of GIS. Shekhar, S., Xiong, H., Zhou, X.(Eds.). Springer, Boston, USA, pp. 1–7.

Heuvelink, G. B. M., Pebesma, E. J., 1999. Spatial aggregation and soil process modelling. Geoderma 89 (1), 47–65.

Hoeting, J. A., Davis, R. A., Merton, A. A., Thompson, S. E., 2006. Model selection for geostatistical models. Ecological Applications 16 (1), 87–98.

Højberg, A. L., Refsgaard, J. C., 2005. Model uncertainty–parameter uncertainty versus conceptual models. Water Science and Technology 52 (6), 177–186.

Huard, D., Mailhot, A., 2008. Calibration of hydrological model GR2M using Bayesian uncertainty analysis. Water Resources Research 44 (2), w02424.

Hughes, J. P., Lettenmaier, D. P., 1981. Data requirements for kriging: estimation and network design. Water Resources Research 17 (6), 1641–1650.

Jacques, D., Mouvet, C., Mohanty, B., Vereecken, H., Feyen, J., 1999. Spatial variability of atrazine sorption parameters and other soil properties in a podzoluvisol. Journal of Contaminant Hydrology 36 (1), 31–52.

Janssen, P. H. M., Heuberger, P. S. C., 1995. Calibration of process-oriented models. Ecological Modelling 83 (1-2), 55–66.

Jarboui, B., Cheikh, M., Siarry, P., Rebai, A., 2007. Combinatorial particle swarm optimization (CPSO) for partitional clustering problem. Applied Mathematics and Computation 192 (2), 337–345.

Jewell, S. A., Gaussiat, N., 2015. An assessment of kriging-based rain-gauge radar merging techniques. Quarterly Journal of the Royal Meteorological Society 141 (691), 2300–2313.

Kalman, R. E., 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82 (1), 35–45.

Kass, R. E., Wasserman, L., 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Journal of the American Statistical Association 90 (431), 928–934.

Kathuria, D., Mohanty, Binayak, P., Katzfuss, M., 2019. A non-stationary geostatistical framework for soil moisture prediction in the presence of surface heterogeneity. Water Resources Research 55 (1), 729–753.

Kavetski, D., Kuczera, G., Franks, S. W., 2006. Bayesian analysis of input uncertainty in hydrological modeling: 1. theory. Water Resources Research 42, w03408.

Kennedy, M. C., O'Hagan, A., 2001. Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (3), 425–464.

Kerry, R., Oliver, M. A., 2004. Average variograms to guide soil sampling. International Journal of Applied Earth Observation and Geoinformation 5 (4), 307–325.

Kerry, R., Oliver, M. A., 2007. Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. Geoderma 140 (4), 383–396.

Keum, J., Kaluarachchi, J. J., 2015. Development of a decision-making methodology to design a water quality monitoring network. Environmental Monitoring and Assessment 187 (7), 466.

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P., 1983. Optimization by simulated annealing. Science 220 (4598), 671–680.

Kitanidis, P. K., 1987. Parametric estimation of covariances of regionalized variables. JAWRA Journal of the American Water Resources Association 23 (4), 557–567.

Kovac, M., Lawrie, J. W., 1991. Soil landscapes of the singleton 1:250.000 sheet. Soil Conservation Service of NSW. Sydney.

Krajewski, W. F., Lakshmi, V., Georgakakos, K. P., Jain, S. C., 1991. A Monte Carlo study of rainfall sampling effect on a distributed catchment model. Water Resources Research 27 (1), 119–128.

Krige, D. G., 1951. A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of the Southern African Institute of Mining and Metallurgy 52 (6), 119–139.

Laloy, E., Vrugt, J. A., 2012. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (ZS) and high-performance computing. Water Resources Research 48 (1), w01526.

Lark, R. M., 2000. Estimating variograms of soil properties by the method-of-moments and maximum likelihood. European Journal of Soil Science 51 (4), 717–728.

Lark, R. M., 2002. Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. Geoderma 105 (1), 49–80.

Lark, R. M., 2003. Developing a cost-effective procedure for investigating within-field variation of soil conditions. Tech. rep., Home-Grown Cereals Authority, UK.

Lark, R. M., 2009. Kriging a soil variable with a simple nonstationary variance model. Journal of Agricultural, Biological, and Environmental Statistics 14 (3), 301–321.

Lark, R. M., Cullis, B. R., 2004. Model-based analysis using reml for inference from systematically sampled data on soil. European Journal of Soil Science 55 (4), 799–813.

Lark, R. M., Marchant, B. P., 2018. How should a spatial-coverage sample design for a geostatistical soil survey be supplemented to support estimation of spatial covariance parameters? Geoderma 319, 89–99.

Lark, R. M., Webster, R., 2006. Geostatistical mapping of geomorphic variables in the presence of trend. Earth Surface Processes and Landforms 31 (7), 862–874.

Laslett, G. M., 1997. Discussion of the paper by DJ Brus and JJ de Gruijter. Geoderma 1 (80), 45–59.

Lesch, S. M., Strauss, D. J., Rhoades, J. D., 1995. Spatial prediction of soil salinity using electromagnetic induction techniques: 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation. Water Resources Research 31 (2), 387–398.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., Bergström, S., 1997. Development and test of the distributed HBV-96 hydrological model. Journal of Hydrology 201 (1-4), 272–288.

Lopes, M. E., 2015. Measuring the algorithmic convergence of random forests via bootstrap extrapolation. Tech. rep., Davis CA: Department of Statistics, University of California, USA.

Lopez, M. G., Seibert, J., 2016. Influence of hydro-meteorological data spatial aggregation on streamflow modelling. Journal of Hydrology 541, 1212–1220.

Lophaven, S., 2004. Design and analysis of environmental monitoring programs. Ph.D. thesis, Technical University of Denmark.

Louppe, G., 2014. Understanding random forests: From theory to practice. Ph.D. thesis, University of Liège, Belgium.

Lugg, W. H., Griffiths, J., van Rooyen, A. R., Weeks, A. R., Tingley, R., 2018. Optimal survey designs for environmental DNA sampling. Methods in Ecology and Evolution 9 (4), 1049–1059.

MacKay, D. J. C., 1992. Information-based objective functions for active data selection. Neural Computation 4 (4), 590–604.

Mann, H. B., Whitney, D. R., 1947. On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics, 50–60.

Marchant, B. P., 2018. Model-based soil geostatistics. In: Pedometrics. Springer, Berlin, Germany, pp. 341–371.

Marchant, B. P., Lark, R. M., 2004. Estimating variogram uncertainty. Mathematical Geology 36 (8), 867–898.

Marchant, B. P., Lark, R. M., 2007a. Optimized sample schemes for geostatistical surveys. Mathematical Geology 39 (1), 113–134.

Marchant, B. P., Lark, R. M., 2007b. Robust estimation of the variogram by residual maximum likelihood. Geoderma 140 (1), 62–72.

Marchant, B. P., McBratney, A. B., Lark, R. M., Minasny, B., 2013. Optimized multi-phase sampling for soil remediation surveys. Spatial Statistics 4, 1–13.

Marchant, B. P., Newman, S., Corstanje, R., Reddy, K. R., Osborne, T. Z., Lark, R. M., 2009. Spatial monitoring of a non-stationary soil property: phosphorus in a Florida water conservation area. European Journal of Soil Science 60 (5), 757–769.

Maskey, S., 2004. Modelling Uncertainty in Flood Forecasting Systems. CRC Press, Boca Raton, USA.

Matérn, B., 1986. Spatial Variation. Springer, Berlin, Germany.

Mateu, J., Müller, W. G., 2012. Spatio-Temporal Design: Advances in Efficient Data Acquisition. John Wiley & Sons, New York, USA.

Matheron, G., 1963. Principles of geostatistics. Economic Geology 58 (8), 1246–1266.

Mauger, G. S., Bumbaco, K. A., Hakim, G. J., Mote, P. W., 2013. Optimal design of a climatological network: beyond practical considerations. Geoscientific Instrumentation, Methods and Data Systems 2 (2), 199–212.

McBratney, A. B., Minasny, B., 2013. Spacebender. Spatial Statistics 4, 57–67.

McBratney, A. B., Pringle, M. J., 1999. Estimating average and proportional variograms of soil properties and their potential use in precision agriculture. Precision Agriculture 1 (2), 125–152.

McBratney, A. B., Webster, R., 1981. Spatial dependence and classification of the soil along a transect in northeast scotland. Geoderma 26 (1-2), 63–82.

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

Meinshausen, N., 2006. Quantile Regression Forests. Journal of Machine Learning Research 7, 983–999.

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 (3), 280–288.

Melsen, L. A., Addor, N., Mizukami, N., Newman, A. J., Torfs, P. J. J. F., Clark, M. P., Uijlenhoet, R., Teuling, A. J., 2018. Mapping (dis) agreement in hydrologic projections. Hydrology and Earth System Sciences 22 (3), 1775–1791.

Melsen, L. A., Teuling, A. J., Torfs, P. J. J. F., Zappa, M., Mizukami, N., Clark, M. P., Uijlenhoet, R., 2016. Representation of spatial and temporal variability in large-domain hydrological models: case study for a mesoscale pre-Alpine basin. Hydrology and Earth System Sciences 20 (6), 2207–2226.

Met Office, 2003. 1 km Resolution UK Composite Rainfall Data from the Met Office Nimrod System. NCAS British Atmospheric Data Centre. http://catalogue.ceda.ac.uk/uuid/82adec1f896af6169112d09cc1174499, accessed 01.02.2016.

Minasny, B., McBratney, A. B., 2005. The Matérn function as a general model for soil variograms. Geoderma 128 (3-4), 192–207.

Minasny, B., McBratney, A. B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences 32 (9), 1378–1388.

Mishra, A. K., Coulibaly, P., 2009. Developments in hydrometric network design: A review. Reviews of Geophysics 47 (2), rG2001.

Moore, D. M., Lees, B. G., Davey, S. M., 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. Environmental Management 15 (1), 59–71.

Moore, I. D., Gessler, P., Nielsen, G., Peterson, G., 1993. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal 57 (2), 443–452.

Müller, W. G., Stehlík, M., 2010. Compound optimal spatial designs. Environmetrics 21 (3-4), 354–364.

Müller, W. G., Zimmerman, D. L., 1999. Optimal designs for variogram estimation. Environmetrics 10 (1), 23–37.

Muthusamy, M., Schellart, A., Tait, S., Heuvelink, G. B. M., 2017. Geostatistical upscaling of rain gauge data to support uncertainty analysis of lumped urban hydrological models. Hydrology and Earth System Sciences 21 (2), 1077–1091.

Nanding, N., Rico-Ramirez, M. A., Han, D., 2015. Comparison of different radar-raingauge rainfall merging techniques. Journal of Hydroinformatics 17 (3), 422–445.

Nembrini, S., König, I. R., Wright, M. N., 2018. The revival of the Gini importance? Bioinformatics 34 (21), 3711–3718.

Nunes, L. M., Cunha, M. C., Ribeiro, L., 2004. Optimal space-time coverage and exploration costs in groundwater monitoring networks. Environmental Monitoring and Assessment 93 (1-3), 103–124.

Nussbaum, M., Walthert, L., Fraefel, M., Greiner, L., Papritz, A., 2017. Mapping of soil properties at high resolution in Switzerland using boosted geoadditive models. SOIL 3 (4), 191–210.

Odeh, I. O. A., McBratney, A. B., Chittleborough, D. J., 1990. Design of optimal sample spacings for mapping soil using fuzzy-k-means and regionalized variable theory. Geoderma 47 (1-2), 93–122.

Olea, R. A., 1984. Sampling design optimization for spatial functions. Journal of the International Association for Mathematical Geology 16 (4), 369–392.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2018. LUCAS Soil, the largest expandable soil dataset for Europe: a review. European Journal of Soil Science 69 (1), 140–153.

Ortiz, J., Deutsch, C. V., 2002. Calculation of uncertainty in the variogram. Mathematical Geology 34 (2), 169–183.

Padarian, J., Minasny, B., McBratney, A. B., 2019. Using deep learning for digital soil mapping. SOIL 5 (1), 79–89.

PaiMazumder, D., Mölders, N., 2009. Theoretical assessment of uncertainty in regional averages due to network density and design. Journal of Applied Meteorology and Climatology 48 (8), 1643–1666.

Pappenberger, F., Beven, K. J., 2006. Ignorance is bliss: Or seven reasons not to use uncertainty analysis. Water Resources Research 42 (5), w05302.

Pardo-Igúzquiza, E., 1998. Optimal selection of number and location of rainfall gauges for areal rainfall estimation using geostatistics and simulated annealing. Journal of Hydrology 210 (1), 206–220.

Pardo-Igúzquiza, E., Dowd, P., 2001. Variance–covariance matrix of the experimental variogram: assessing variogram uncertainty. Mathematical Geology 33 (4), 397–419.

Paterson, S., McBratney, A. B., Minasny, B., Pringle, M. J., 2018. Variograms of soil properties for agricultural and environmental applications. In: Pedometrics. Springer, Berlin, Germany, pp. 623–667.

Patterson, H. D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58 (3), 545–554.

Pintore, A., Holmes, C. C., 2004. Spatially Adaptive Non-stationary Covariance Functions via Spatially Adaptive Spectra. Tech. rep., University of Oxford, Oxford, UK.

Pozdnoukhov, A., Kanevski, M., 2006. Monitoring network optimisation for spatial data classification using support vector machines. International Journal of Environment and Pollution 28 (3-4), 465–484.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L., Vanrolleghem, P. A., 2007. Uncertainty in the environmental modelling process–a framework and guidance. Environmental Modelling & Software 22 (11), 1543–1556.

Reichle, R. H., Koster, R. D., Dong, J., Berg, A. A., 2004. Global soil moisture from satellite observations, land surface models, and ground data: Implications for data assimilation. Journal of Hydrometeorology 5 (3), 430–442.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S. W., 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resources Research 46 (5), w05521.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., Franks, S. W., 2011. Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. Water Resources Research 47 (11), w11516.

Rendu, J.-M., 1980. Disjunctive kriging: comparison of theory with actual results. Journal of the International

Association for Mathematical Geology 12 (4), 305–320.

Reusser, D., Buytaert, W., Vitolo, C., 2017. RHydro: Classes and methods for hydrological modelling and analysis. R package version 2014-04.1/r185. Accessed 19.10.2016.
URL https://R-Forge.R-project.org/projects/r-hydro/

Rico-Ramirez, M. A., Gonzalez-Ramirez, E., Cluckie, I., Han, D., 2009. Real-time monitoring of weather radar antenna pointing using digital terrain elevation and a Bayes clutter classifier. Meteorological Applications 16 (2), 227–236.

Rosenthal, J. S., et al., 2011. Optimal proposal distributions and adaptive MCMC. In: Handbook of Markov Chain Monte Carlo. Vol. 4. CRC Press, Boca Raton, USA, pp. 93–111.

Roudier, P., 2018. Package "clhs". R package version 0.7-0. Accessed 01.08.2018.
URL https://CRAN.R-project.org/package=clhs

Roudier, P., Beaudette, D. E., Hewitt, A. E., 2012. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. In: Digital Soil Assessments and Beyond. CRC Press, Sydney, NSW, Australia, pp. 227–231.

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

Russo, D., 1984. Design of an optimal sampling network for estimating the variogram 1. Soil Science Society of America Journal 48 (4), 708–716.

Samuel-Rosa, A., 2017. spsann: Optimization of Sample Configurations using Spatial Simulated Annealing. R package version 2.1-0. Accessed 21.10.2017.
URL https://CRAN.R-project.org/package=spsann

Sawicka, K., Wadoux, A. M. J.-C., Heuvelink, G. B. M., 2017. Sample design optimisation techniques and associated software. Tech. rep., Wageningen University & Research, NL, QUICS project deliverable 3.1.

Schiemann, R., Erdin, R., Willi, M., Frei, C., Berenguer, M., Sempere-Torres, D., 2011. Geostatistical radar-raingauge combination with nonparametric correlograms: methodological considerations and application in Switzerland. Hydrology and Earth System Sciences 15 (5), 1515–1536.

Sen, A., Srivastava, M., 2012. Regression Analysis: Theory, Methods, and Applications. Springer Science & Business Media, New York, USA.

Seo, D.-J., Breidenbach, J. P., 2002. Real-time correction of spatially nonuniform bias in radar rainfall data using rain gauge measurements. Journal of Hydrometeorology 3 (2), 93–111.

Shrestha, D. L., Solomatine, D. P., 2008. Data-driven approaches for estimating uncertainty in rainfall-runoff modelling. International Journal of River Basin Management 6 (2), 109–122.

Sideris, I. V., Gabella, M., Erdin, R., Germann, U., 2014. Real-time radar–rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland. Quarterly Journal of the Royal Meteorological Society 140 (680), 1097–1111.

Sinclair, S., Pegram, G., 2005. Combining radar and rain gauge rainfall estimates using conditional merging. Atmospheric Science Letters 6 (1), 19–22.

Solow, A. R., 1986. Mapping by simple indicator kriging. Mathematical Geology 18 (3), 335–352.

St-Hilaire, A., Ouarda, T. B. M. J., Lachance, M., Bobée, B., Gaudet, J., Gignac, C., 2003. Assessment of the impact of meteorological network density on the estimation of basin precipitation and runoff: a case study. Hydrological Processes 17 (18), 3561–3580.

Starks, T. H., 1986. Determination of support in soil sampling. Mathematical Geology 18 (6), 529–537.

Stein, M. L., 2006. Interpolation of Spatial Data: Some Theory for Kriging. Springer Science & Business Media, Dordrecht, NL.

Stockmann, U., Minasny, B., McBratney, A. B., Hancock, G. R., Willgoose, G. R., 2012. Exploring short-term soil landscape formation in the Hunter Valley, NSW, using gamma ray spectrometry. In: Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping 2012, Sydney, Australia. CRC Press, London, UK, pp. 77–82.

Storn, R., Price, K., 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization 11 (4), 341–359.

Stumpf, F., Schmidt, K., Behrens, T., Schönbrodt-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 (4), 499–509.

Talamba, D. B., Parent, E., Musy, A., 2010. Bayesian multiresponse calibration of TOPMODEL: Application to the Haute-Mentue catchment, Switzerland. Water Resources Research 46 (8), w08524.

Ter Braak, C. J. F., 2006. A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy

Bayesian computing for real parameter spaces. Statistics and Computing 16 (3), 239–249.

Terink, W., Leijnse, H., van den Eertwegh, G., Uijlenhoet, R., 2018. Spatial resolutions in areal rainfall estimation and their impact on hydrological simulations of a lowland catchment. Journal of Hydrology 563, 319–335.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., Srikanthan, S., 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. Water Resources Research 45 (12), w00B14.

Tian, Y., Booij, M. J., Xu, Y.-P., 2014. Uncertainty in high and low flows due to model structure and parameter errors. Stochastic Environmental Research and Risk Assessment 28 (2), 319–332.

Todini, E., 2001. A Bayesian technique for conditioning radar precipitation estimates to rain-gauge measurements. Hydrology and Earth System Sciences 5 (2), 187–199.

Tóth, G., Jones, A., Montanarella, L., 2013. LUCAS Topsoil Survey: Methodology, Data and Results. Tech. rep., JRC, publications Office of the European Union, Luxembourg.

Troutman, B. M., 1983. Runoff prediction errors and bias in parameter estimation induced by spatial variability of precipitation. Water Resources Research 19 (3), 791–810.

Tuia, D., Pozdnoukhov, A., Foresti, L., Kanevski, M., 2013. Active learning for monitoring network optimization. In: Spatio-Temporal Design: Advances in Efficient Data Acquisition. Wiley Online Library, New-York, USA, pp. 285–318.

Van der Keur, P., Henriksen, H.-J., Refsgaard, J. C., Brugnach, M., Pahl-Wostl, C., Dewulf, A. R. P. J., Buiteveld, H., 2008. Identification of major sources of uncertainty in current IWRM practice. Illustrated for the Rhine Basin. Water Resources Management 22 (11), 1677–1708.

Van Groenigen, J. W., Siderius, W., Stein, A., 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87 (3-4), 239–259.

Van Groenigen, J. W., Stein, A., 1998. Constrained optimization of spatial sampling using continuous simulated annealing. Journal of Environmental Quality 27 (5), 1078–1086.

Velasco-Forero, C. A., Sempere-Torres, D., Cassiraga, E. F., Gómez-Hernández, J. J., 2009. A non-parametric automatic blending methodology to estimate rainfall fields from rain gauge and radar data. Advances in Water Resources 32 (7), 986–1002.

Velasco-Forero, C. A., Sempere-Torres, D., Sánchez-Diezma, R., Cassiraga, E. F., Gómez-Hernández, J. J., 2005. Automatic estimation of rainfall fields for hydrological applications: blending radar and rain gauge data in real time. In: 32nd Conference on Radar Meteorology. Vol. 12. American Meteorological Society, Albuquerque, USA, pp. CD–ROM, P13R.

Verworn, A., Haberlandt, U., 2011. Spatial interpolation of hourly rainfall–effect of additional information, variogram inference and storm properties. Hydrology and Earth System Sciences 15 (2), 569–584.

Villeneuve, J.-P., Morin, G., Bobee, B., Leblanc, D., Delhomme, J.-P., 1979. Kriging in the design of streamflow sampling networks. Water Resources Research 15 (6), 1833–1840.

Viscarra Rossel, R. A., Taylor, H. J., McBratney, A. B., 2007. Multivariate calibration of hyperspectral γ-ray energy spectra for proximal soil sensing. European Journal of Soil Science 58 (1), 343–353.

Vogl, S., Laux, P., Qiu, W., Mao, G., Kunstmann, H., 2012. Copula-based assimilation of radar and gauge information to derive bias-corrected precipitation fields. Hydrology and Earth System Sciences 16 (7), 2311–2328.

Voltz, M., Webster, R., 1990. A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. European Journal of Soil Science 41 (3), 473–490.

Vrugt, J. A., Ter Braak, C. J. F., Clark, M. P., Hyman, J. M., Robinson, B. A., 2008. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. Water Resources Research 44 (12), w00B09.

Wadoux, A. M. J.-C., 2019. Using deep learning for multivariate mapping of soil with quantified uncertainty. Geoderma 351, 59–70.

Wadoux, A. M. J.-C., Brus, D. J., Heuvelink, G. B. M., 2018. Accounting for non-stationary variance in geostatistical mapping of soil properties. Geoderma 324, 138–147.

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. Advances in Water Resources 107, 126–138.

Wadoux, A. M. J.-C., Marchant, B. P., Lark, R. M., 2019a. Efficient sampling for geostatistical surveys. European Journal of Soil Science (In Press).

Wadoux, A. M. J. C., Padarian, J., Minasny, B., 2019b. Multi-source data integration for soil mapping using deep learning. SOIL 5 (1), 107–119.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., Sorooshian, S., 2001. A framework for development and application of hydrological models. Hydrology and Earth System Sciences 5 (1), 13–26.

Walvoort, D. J. J., Brus, D. J., De Gruijter, J. J., 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Computers & Geosciences 36 (10), 1261–1267.

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 (2), 539–543.

Webster, R., 2000. Is soil variation random? Geoderma 97 (3-4), 149–163.

Webster, R., Beckett, P. H. T., 1968. Quality and usefulness of soil maps. Nature 219 (5155), 680.

Webster, R., Lark, R. M., 2012. Field sampling for environmental science and management. Routledge, Abingdon-on-Thames, UK.

Webster, R., Oliver, M. A., 1989. Disjunctive kriging in agriculture. In: Proceedings of the Third International Geostatistics Congress September 5–9, 1988, Avignon, France. Springer, Dordrecht, NL, pp. 421–432.

Webster, R., Oliver, M. A., 1992. Sample adequately to estimate variograms of soil properties. Journal of Soil Science 43 (1), 177–192.

Webster, R., Oliver, M. A., 2007. Geostatistics for Environmental Scientists. John Wiley & Sons, Chichester, UK.

Were, K., Bui, D. T., Dick, Ø. B., Singh, B. R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators 52, 394–403.

Wright, M. N., Ziegler, A., et al., 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software 77 (1), 1–17.

Xu, H., Xu, C.-Y., Chen, H., Zhang, Z., Li, L., 2013. Assessing the influence of rain gauge density and distribution on hydrological model performance in a humid region of China. Journal of Hydrology 505, 1–12.

Yang, E.-G., Kim, H. M., Kim, J., Kay, J. K., 2014. Effect of observation network design on meteorological forecasts of Asian dust events. Monthly Weather Review 142 (12), 4679–4695.

Yang, L., Brus, D. J., Zhu, A.-X., Li, X., Shi, J., 2018. Accounting for access costs in validation of soil maps: A comparison of design-based sampling strategies. Geoderma 315, 160–169.

Yfantis, E. A., Flatman, G. T., 1988. On sampling nonstationary spatial autocorrelated data. Computers & Geosciences 14 (5), 667–686.

Yfantis, E. A., Flatman, G. T., Behar, J. V., 1987. Efficiency of kriging estimation for square, triangular, and hexagonal grids. Mathematical Geology 19 (3), 183–205.

Zeng, Q., Chen, H., Xu, C.-Y., Jie, M.-X., Chen, J., Guo, S.-L., Liu, J., 2018. The effect of rain gauge density and distribution on runoff simulation using a lumped hydrological modelling approach. Journal of Hydrology 563, 106–122.

Zevenbergen, L. W., Thorne, C. R., 1987. Quantitative analysis of land surface topography. Earth Surface Processes and Landforms 12 (1), 47–56.

Zhu, Z., Stein, M. L., 2005. Spatial sampling design for parameter estimation of the covariance function. Journal of Statistical Planning and Inference 134 (2), 583–603.

Zhu, Z., Stein, M. L., 2006. Spatial sampling design for prediction with estimated parameters. Journal of Agricultural, Biological, and Environmental Statistics 11 (1), 24.

Zimmerman, D. L., 2006. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. Environmetrics 17 (6), 635–652.

Summary

Space-time monitoring and prediction of environmental variables requires measurements of the environment. But environmental variables cannot be measured everywhere and all the time. Scientists can only collect a fragment, a sample of the property of interest in space and time, with the objective of using this sample to infer the property at unvisited locations and times. Sampling might be a costly and time consuming affair. Consequently, we need efficient strategies to select an optimal design for mapping.

Most studies on sampling design optimization consider the case of predictive mapping using geostatistics. In recent years geostatistical models and associated mapping techniques have advanced, which calls for adaptation of associated sampling designs. The main objective of this thesis is to address the optimal design of four recent advances in mapping.

Chapter 3 explores sampling design optimization for the non-stationary variance geostatistical model defined in **Chapter 2**. Accounting for non-stationarity in the variance of environmental properties in complex landscapes leads to better quantification of the mapping uncertainty. This is applied in a case study mapping daily rainfall in the north of England, and optimizing the rain-gauges for mapping. It is shown that rainfall prediction benefits from a model that includes non-stationarity in the mean and variance, as shown by the likelihood and Akaike Information Criterion statistics. The optimization of the rain gauge network is achieved by spatial simulated annealing. The optimized rain gauge network improves slightly the rainfall mapping accuracy. The accuracy gain is limited because I used a static design for all time steps, while the areas with larger prediction uncertainty vary day-by-day. The optimized design also shows a specific spatial pattern, with a fairly uniform spatial distribution but an increased density in areas where the residual variance is large. I further test an optimized design using a reduction of 10% of the total number of rain-gauges. The optimized design shows a significant improvement over the original design using all rain-gauges. I conclude that 10% of the rain-gauges may be removed (e.g. to save costs) without loss of mapping accuracy, provided that the

rain-gauges are placed optimally.

Chapter 4 investigates the use of simple sampling strategies to account for a criterion that encompasses both prediction error variance and variogram parameter uncertainty in geostatistical mapping of soil properties. I test two sampling designs: spatial coverage and spatial coverage supplemented by a subset of close-pairs units, and compare these to a design optimized for this criterion. I show that a spatial coverage design performs poorly for mapping using ordinary kriging because of the lack of information at short distances to estimate the variogram parameters. This is valid for series of estimated variogram parameters of a Matèrn function. An optimized design performs always slightly better, but has several disadvantages. For example, it requires the variogram parameters to be known. It also involves defining an objective function characterizing the total error, and minimizing this error using optimization algorithms. In contrast, a spatial coverage design supplemented by a subset of close-pair units offers accurate results for most variograms tested. I therefore recommend to use the latter design for designing a geostatistical survey, unless prior knowledge of the variogram is available (e.g. an average variogram). If an average variogram is available for the property of interest, it can be used to optimize the design. I further test the minimum number of units required to estimate the variogram of a geostatistical survey, and show that it strongly depends on the degree of spatial correlation of the target variable. For large values of the variogram effective range and small nugget to sill ratios, it is shown that only 15 units are enough to make geostatistical analysis worthwhile, i.e. more accurate than a design-based estimate.

Mapping is not always performed using geostatistical methods. There is growing interest towards mapping using data-driven, non-linear machine learning techniques. The objective of **Chapter 5** is to extend our knowledge on sampling optimization for mapping using random forest, and to compare it to conventional sampling designs. I tested the methodology in a potential application scenarios, mapping topsoil organic carbon at European scale using measurements of the LUCAS dataset as population of interest. I demonstrate that an optimized design is always more accurate than other common designs, but possible to obtain only when subsampling an existing dataset with known values of the soil property at all locations. By comparing the mean square error (MSE) of the maps obtained by an optimized design with the those obtained by common designs, it is shown that optimizing a design in terms of MSE is not always worthwhile. When the sample size increases, the maps produced by the different designs converge to similar accuracy values. In a case study on large scale soil organic carbon mapping, a sampling density greater than 1 sampling unit per 4000 km^2 decreases markedly the difference in term of average MSE between designs. A design optimized for the mean squared shortest standardized distance in

the feature space has the closest match with the optimized design in terms of MSE. By analysing the distribution of the sampling locations in both geographic and feature space, I further show that the optimized design is not spread in the geographic space, but seems to be spread somewhat uniformly in the feature space, and especially in the most important covariates of the machine learning model. It is however difficult to draw further conclusions because of the complex spread of the units in feature space. Further research is needed in this direction.

Sampling design optimization becomes more complex when the ultimate goal is to provide a map used as input for a model whose output is the main interest. This is done by integrating geostatistics for mapping rainfall and Bayesian calibration of a hydrological model for predicting discharges in **Chapter 6**. The Bayesian calibration enables to capture model input, initial state, parameter and structural uncertainty, while also taking uncertainties in the output measurements into account. In a case study predicting river discharge using a rainfall-runoff model and maps of rainfall as input, a single rain gauge is sufficient to obtain accurate model parameter calibration and discharge predictions. Adding up to five rain gauges improves the model prediction. Adding even more only produces a marginal improvement of the prediction accuracy. Calibrating the rainfall time series as additional parameters leads to more accurate model performance compared to the case where rainfall uncertainty is not updated using discharge measurements. Furthermore, it is demonstrated for the case study that model parameter uncertainty is the main contributor to the posterior discharge uncertainty and that input uncertainty has a relatively small contribution. However, the study also shows that Bayesian calibration of rainfall has serious computational disadvantages. In particular, calibrating a large number of rainfall input parameters remains a serious challenge.

The thesis synthesis is given in **Chapter 7**. It discusses the findings of this thesis, compares these with existing literature, gives directions for future research and provides a personal reflection on sampling design optimization practices. On the basis of this thesis, I conclude that *there is no single best optimal design*. It is very much case dependent. It depends, among others, on: (i) the assumed model of spatial variation, (ii) the assumption whether we need or need not estimate the parameters from the data, and (iii) the criterion that is used to optimize the sampling configuration. This thesis shows that the choice of the criterion has a serious impact on the optimized design. In practice we may not know the three elements listed above. This is typically the case at the start of a project when no previous data or expertise are available but where we need to design a survey. In this case, it is sensible to use some rules of thumb to design a survey for mapping. Chapter 7 provides some basis for this.

This thesis makes a step towards derivation of optimal designs for novel mapping techniques, with case studies on mapping soil and hydrological variables. But it also shows that we are just at the beginning of this specific field of science. In recent years, there has been a large increase in complexity of techniques and models used for mapping. We make more use of spatially explicit covariate information, such as remote sensing imagery, and measurements are increasingly inferred rather than measured. Mapping techniques have become more data-driven and non-linear, increasing *de facto* the complexity of the sampling designs that should accompany such developments. Because sampling is the basis of mapping and has a large impact on cost and accuracy, this research field will remain as important as ever in geostatistics and spatial modelling.

Résumé

La surveillance et la prévision spatio-temporelle des variables environnementales demande d'acquérir des observations qu'il est difficile d'obtenir partout et à tout moment. Les scientifiques ne peuvent ainsi collecter qu'un fragment, un échantillon de la variable étudiée dans l'espace et dans le temps, dans le but d'utiliser cet échantillon pour déduire et cartographier cette variable aux endroits et moments non visités. L'échantillonnage peut représenter une tâche coûteuse et fastidieuse. En conséquence, il est nécessaire de disposer de stratégies efficaces pour produire un plan d'échantillonnage optimal pour la cartographie.

La plupart des études menées sur l'optimisation du plan d'échantillonnage se cantonnent au cas de la cartographie prédictive utilisant la géostatistique. Au cours des dernières années, les modèles géostatistiques et les techniques de cartographie associées ont largement évolué, nécessitant une adaptation des plans d'échantillonnage disponibles. Cette thèse a ainsi pour principal objectif d'aborder le plan d'échantillonnage de quatre avancées récentes en cartographie.

Le **Chapitre 3** explore l'optimisation du plan d'échantillonnage pour un modèle géostatistique de variance non stationnaire défini dans le **Chapitre 2**. La prise en compte de cette non-stationnarité dans la variance des propriétés environnementales des paysages complexes permet de mieux quantifier l'incertitude associée aux cartographies. Cette méthode est appliquée dans une étude de cas de cartographie des précipitations journalières dans le nord de l'Angleterre et d'optimisation spatiale des pluviomètres. Je montre dans ce chapitre que la prévision spatiale et temporelle des précipitations tire avantage d'un modèle qui inclut la non-stationnarité dans la moyenne et la variance, comme le montrent les statistiques de vraisemblance et de critère d'information d'Akaike. L'optimisation du réseau pluviométrique est obtenue par un recuit simulé spatial. Le réseau de pluviomètres ainsi optimisé améliore légèrement la précision de la cartographie des précipitations. Le gain de précision reste limité car j'ai utilisé un échantillon identique pour tous les pas de temps, tandis que les zones avec une plus grande incertitude de prévision varient de jour en jour. Le plan d'échantillonnage optimisée montre également une configuration spé-

cifique, avec une distribution spatiale assez uniforme mais une densité accrue dans les zones où la variance résiduelle est grande. Je teste ensuite un plan d'échantillonnage optimisé en utilisant une réduction de 10% du nombre total de pluviomètres. Je montre alors une amélioration significative par rapport au plan d'échantillonnage original utilisant tous les pluviomètres disponibles. Je conclus que 10% des pluviomètres peuvent être supprimés (par exemple pour des réductions de coûts) sans perte de précision pour la cartographique, à la condition que les pluviomètres soient placés de manière optimale.

Le **Chapitre 4** examine l'utilisation de stratégies d'échantillonnage simples pour prendre en compte un critère englobant à la fois la variance d'erreur de prévision et l'incertitude des paramètres du variogramme dans la cartographie géostatistique des propriétés du sol. Pour cela, je teste deux plans d'échantillonnage : le premier correspond à un répartition homogène des points dans l'espace (couverture spatiale) et le suivant correspondant à la couverture spatiale complétée par un sous-ensemble d'unités proches. Je compare ces plans à un plan optimisé pour les critères retenus. Je montre qu'un plan d'échantillonnage de couverture spatiale donne de piètres résultats pour la cartographie utilisant le krigeage ordinaire en raison du manque d'informations à courte distance pour estimer les paramètres du variogramme. Ceci est valable pour des séries de paramètres estimé sur la base d'un variogramme de Matèrn. Utiliser un échantillonnage optimisé fonctionne toujours légèrement mieux, mais présente plusieurs inconvénients dont notamment celui de devoir connaître les paramètres du variogramme. Cela implique également de définir une fonction objectif caractérisant l'erreur totale et de la minimiser à l'aide d'algorithmes d'optimisation. En revanche, un échantillonnage dit de couverture spatiale complétée par un sous-ensemble d'unités proches offre des résultats précis pour la plupart des variogrammes testés. Je recommande donc d'utiliser cette dernière méthode d'échantillonnage pour la conception d'un échantillonnage géostatistique, à moins que le variogramme ne soit préalablement connu (par exemple, un variogramme moyen). Si un variogramme moyen est disponible pour la propriété d'intérêt, il peut être utilisé pour optimiser l'échantillonnage. Je teste ensuite le nombre minimum d'unités nécessaires pour estimer le variogramme d'une étude géostatistique et montre que cela dépend fortement du degré de corrélation spatiale de la variable étudiée. Pour les grandes valeurs de portée effective du variogramme et petit ratios de pépite sur palier, il est montré que seulement 15 unités suffisent pour rendre l'analyse géostatistique intéressante, c'est-à-dire plus précise qu'une estimation fondée sur un échantillonnage non probabiliste.

La cartographie n'est pas toujours effectuée à l'aide de méthodes géostatistiques. Il existe un intérêt croissant pour la cartographie utilisant des techniques d'apprentissage automatique non linéaires basé sur les données. L'objectif du **Chapitre 5** est d'étendre nos connaissances sur l'optimisation de l'échantillonnage pour la cartographie à l'aide de forêt d'arbres décisionnels et de la comparer aux plans d'échantillonnage conventionnels. J'ai testé la méthodologie dans le cadre de scénarios d'application potentiels, en cartographiant le carbone organique de la couche superficielle du sol à l'échelle européenne à l'aide des données LUCAS en tant que population d'intérêt. Dans ce chapitre, je démontre qu'un échantillonnage optimisé est toujours plus précis que d'autres plans d'échantillonnages couramment utilisés. Cependant, cette approche n'est possible que dans un cas restreint où on procède à un sous-échantillonnage d'un jeu de données existant avec des valeurs connues de la propriété du sol. En comparant l'erreur quadratique moyenne (EQM) des cartes obtenues par un échantillonnage optimisé à celles obtenues par des plans échantillonnages communs, il est montré que l'optimisation d'un échantillonnage en termes d'EQM n'est pas toujours intéressant. Lorsque la taille de l'échantillon augmente, les précisions des cartes produites par les différents type échantillons convergent vers des valeurs similaires. Dans une étude de cas sur la cartographie du carbone organique du sol à grande échelle, une densité d'échantillonnage supérieure à 1 unité d'échantillonnage par 4000 km^2 réduit considérablement la différence en terme de EQM moyen entre les types d'échantillons. Un échantillon optimisé pour la distance normalisée quadratique moyenne la plus courte dans l'espace des covariables correspond le mieux à un échantillon optimisé en termes d'EQM. En analysant la distribution des localisations des échantillons dans l'espace géographique et dans l'espace des covariables, je montre également qu'un échantillon optimisé n'est pas distribué uniformément dans l'espace géographique, mais semble être répartie de manière assez uniforme dans l'espace des covariables, et en particulier en considérant les variables les plus importantes pour le modèle d'apprentissage automatique. Il est toutefois difficile de tirer des conclusions supplémentaires en raison de la dispersion complexe des unités dans l'espace des covariables. Des recherches complémentaires sont nécessaires dans cette direction.

L'optimisation du plan d'échantillonnage devient plus complexe lorsque le but ultime est de fournir une carte utilisée comme entrée pour un modèle dont la prévision est le principal objectif. J'aborde cette question dans le cas où la géostatistique est utilisée pour la cartographie des précipitations et la calibration bayésienne d'un modèle hydrologique pour la prévision des débits au **Chapitre 6**. La calibration bayésienne permet de capturer les incertitudes d'entrée, d'état initial, de paramètre et de structure du modèle, tout en tenant compte des incertitudes des mesures de sortie. Dans une étude de cas de prévision du débits d'une rivière à l'aide d'un modèle pluie-ruissellement et de cartes des précipitations, un seul pluviomètre peut suffire pour obtenir un étalonnage précis des paramètres du modèle et des prévisions de débits. L'ajout de cinq pluviomètres améliore cependant la prévision du modèle. En ajouter davantage ne produit qu'une amélioration marginale de la précision des prévisions. Le calibrage de la série chronologique des précipitations en tant que paramètres supplémentaires permet d'obtenir des performances de modèle plus précises que dans le cas où l'incertitude des précipitations n'est pas actualisée à l'aide des mesures de débit. En outre, il est démontré pour l'étude de cas que l'incertitude des paramètres du modèle est le principal facteur d'incertitude du la loi postérieure du débit et que l'incertitude des intrants a une contribution relativement faible. Cependant, l'étude montre également que l'étalonnage bayésien de la pluviométrie présente de graves inconvénients de calcul. En particulier, la calibration dans un temps raisonnable d'un grand nombre de paramètres d'entrée de pluie reste un défi majeur.

La synthèse de la thèse est présentée au **Chapitre 7**. Elle rappelle les résultats de cette thèse, les compare à la littérature existante, donne des orientations pour les recherches futures et fournit une réflexion personnelle sur les pratiques d'optimisation des plans d'échantillonnages. Sur la base de cette thèse, je conclue qu'*il n'existe pas un unique plan d'échantillonnage optimal.* Cela dépend beaucoup de la finalité de ce plan et, entre autres : (i) du modèle supposé de variation spatiale, (ii) de l'hypothèse selon laquelle nous avons besoin d'estimer ou non les paramètres à partir des données, et (iii) du critère utilisé pour optimiser la configuration d'échantillonnage. Cette thèse montre que le choix du critère a un impact important sur l'échantillon optimisée. En pratique, nous pouvons ne pas connaître les trois éléments énumérés ci-dessus. C'est généralement le cas au début d'un projet lorsqu'aucune donnée ou expertise antérieure n'est disponible, mais qu'il est nécessaire de concevoir un plan d'échantillonnage. Dans ce cas, il est judicieux d'utiliser certaines règles empiriques pour concevoir une plan d'échantillonnage à des fins de cartographie. Le Chapitre 7 fournit une base pour cela.

Cette thèse a pour ambition de constituer un pas en avant vers la dérivation de plans d'échantillonnage optimaux pour de nouvelles techniques de cartographie, avec des études de cas sur la cartographie des variables hydrologiques et pédologiques. Mais elle montre aussi que nous ne sommes qu'au début de ce domaine scientifique spécifique. Ces dernières années, la complexité des techniques et des modèles utilisés pour la cartographie a considérablement augmenté. Nous utilisons davantage les informations sur des covariables spatialement explicites, telles que les images de télédétection, et les mesures sont de plus en plus déduites plutôt que mesurées. Les techniques de cartographie sont devenues davantage axées sur les données et modèlent des processus non linéaires, augmentant de fait la complexité des plans d'échantillonnage devant accompagner de tels développements. Parce que l'échantillonnage est la base de la cartographie et a un impact important sur les coûts et la précision des prévisions, ce domaine de recherche restera aussi important que jamais en géostatistique et en modélisation spatiale.

Acknowledgements

It was an early autumn morning of 2017, the air was cool and sky was grey. The heavy night rain had slowed down to a faint drizzle. There I found myself struggling in a field between the villages of Cotgrave and Keyworth, Southeast of Nottingham. The night rain had turned the ground into thick mud. The farmer had ploughed the field a day earlier, and left the path with a layer of what the British scientists elegantly classified as “Slowly permeable, seasonally wet, slightly acid but base-rich loamy and clayey soils”. My landlord in Cotgrave had lent me a red kid’s bike for the time of my stay. The bike was too small for me, in particular to cross the hilly landscape between my apartment and the British Geological Survey in Keyworth. That day of November, the sticky clayey soils were stronger than what my legs could take, pedalling on a small bike. I continued walking through the mud when the drizzle turned once again into rain. The situation must have been somewhat funny to the observer’s eye: A Frenchman pushing that tiny bike through the mud under the rain, trying hard his way to work in the English countryside but employed in the Netherlands. How did I get there?

Obviously, this was because of my position as a PhD candidate in Wageningen. When I sent an application to Gerard and Dick in April 2015, as I realize now, I had no idea of what kind of adventure I was actually about to apply to. The PhD topic was hydrology and applied mathematics oriented and supported by a Marie-Curie ITN. I had intuitively applied at the project at the time, not with great excitement for the topic itself, rather because I knew by names the two PhD advisors: Gerard Heuvelink and Dick Brus. I was neither a hydrologist nor a mathematician, but they decided to give the motivated young scientist a chance to work on the project. With hindsight, I dare to say that working on a PhD is an easy task when the supervision is excellent.

Gerard, as my daily supervisor, you were always patient with me and available to answer my questions. I have always admired how you could explain me complicated matters and share your knowledge in such a way that even my grandmother would understand. You gave me the necessary freedom to conduct my research while being always there to provide constructive and positive comments, at all steps of my PhD. Dick, as my co-supervisor, I greatly appreciated your directness and your attention to detail. In 2016, when I sent you a draft of my first article, you answered very frankly that the work was "sloppy", and you explained me how to improve it. This was very valuable, as I constantly read and improved my manuscript afterwards before sending you a draft to comment. Moreover, I realized the importance of choosing the right terms in the right context.

The successful completion of my PhD was made possible with the trust I had in my supervisors, and my appreciation to both their personal and scientific qualities. I will continue to build upon with what I have learned from them during my career. You made me enjoy my time as a PhD candidate, even though I was based in Wageningen (*sic*).

Mais revenons à nos moutons ![1]

With further thought, the actual reason why I had soaked clothes and muddy shoes is because Murray Lark and Ben Marchant kindly accepted to host and supervise me for three months at the British Geological Survey in Keyworth. But I would not pass them the buck of my misadventure. Actually, I made the PhD article that I like most with them. Ben provided daily supervision and was open for discussions at all times, while Murray made time to meet me and provided ideas even though he was busy with his move to the University of Nottingham and his office was full of boxes. I also made three other secondments within the QUICS project: in Bristol, Delft and Sydney. In Bristol, I was advised by Miguel Rico-Ramirez and I worked together with Francesca Cecinati. I also appreciated sharing my office with Omar Wani, who demonstrated me that it is possible to be efficient and work late at night without drinking coffee. This was an eye opening experience to me. Late 2016, I went to Delft as Francois Clemens and Jeroen Langeveld kindly accepted to host me for three months. There, I spent time with Antonio Moreno Ródenas, another QUICS fellow, who is passionate about the PhD topic of his girlfriend and builds amazing robots to detect where insects lay their eggs. My series of secondments ended up in Sydney, where I was welcomed by Budiman Minasny and Alex McBratney. I had a great time in Sydney, thanks to the enthusiasm of Budiman and the company of Mario, Vanessa, José, Yuxin and all the others. I look forward to meeting and working together again soon.

Before my secondments, my infancy in research started in Angers in 2009, where I made my Bachelor degree and received teaching in geomorphology by Grégoire Maillet, soil science by Aziz Ballouche and petrology by Fabrice Redois. You three

[1] But let us come back to the topic!

piqued my interested for these fields and at the end of my first year, I knew that I wanted to work on a PhD. Also, I spent four years in Tübingen for my Master's degree. I arrived there by coincidence during an Erasmus, but it was no serendipity when I asked to stay there. I learned a lot from Thomas Scholten, Karsten Schmidt, Leonardo Ramirez-Lopez and Thorsten Behrens. Thomas, you also gave me incredible opportunities; I remember visiting you in October 2012, after a course on soil mapping, to ask with my imperfect German whether I could contribute to the "YangtzeGeo" project. You accepted immediately and introduced me to your PhD student Felix Stumpf. With Felix, I went to China several times, I learned English by talking with him during the endless field campaigns and I met a friend. Your decision to accept me in this project guided my path to this PhD, but also, surprisingly, to the mud on that drizzly morning in November.

I now continue my walk in the muddy path and think, with a hint of irony, about the funding that I received to make this moment possible. I was funded by a Marie-Curie ITN. The project was organized on a daily basis by Will Shepherd and Alma Schellart. This was a lot of work and you allowed me and all other QUICS fellows to have an unforgettable and enjoyable PhD experience. You also made it possible (with all the other organizers) for me to meet the other QUICS fellows: Nazmul, Vasilis, Vivian, Carla, Francesca, Sanda, Mathieu, Mahmood, Antonio, Manoranjan, Kasia, Ambuj, Arturo and Omar. One cannot feel lonely with so many (remote) office mates. Actually, I also happened to have real office mates on my occasional moments in Wageningen: Simona, Arturo, Marcos, Kasia, Marijn, Jasper, Selçuk, Luc and the others. I specially thank Titia for sharing some beers and funny stories and to make the office more lively than a cemetery. I am also grateful for receiving additional funds from the LEB and the Huub and Julienne Spiertz funds.

On a more amusing note, I also see myself telling that story to my friends from the village. They think I spend my days on a chair, the *fonctionnaire* as they call me. I do not expect that they believe me, but to share a few beers, funny stories and stupid jokes, as we always do. Adrien, Julien, Romain, Alexis, and their girlfriends (or boyfriends) are always of great company. But I hope you will behave on my defence.

If I tell the exact same story to my friends from my Tübingen time: David, Jordi, Marie, Nerea, Andrea, Manuel, Marie-Léonie, Gerard and all the others, I expect them not to be surprised. They think I am a geologist always in the field. I can explain you again what I do around a *HeffeWeizen* in *Tangente*.

When I am tired, I know that I can always go back to my village, to the family house. When moving from place to place all around the world, it is important to have a base to rest and find stability. I left home when I was 17 because I wanted to move fast.

Years later, I come back when I need to slow down.

I kept the very best for the end. Anna, like it meant to be, I met you the day I arrived in the Netherlands to start my PhD. You supported me in every aspect of the life in the Netherlands. You listened to my complains about Wageningen and helped me to escape whenever I needed it. This whole adventure would not have been the same without you.

Arriving on the tarred road, at the intersection of Nicker hill and Willow brook. I realize that the loose mud on my shoes will soon vanish as I walk down the street. The clouds on the horizon seem to be clearing out. What an engaging and fresh start to the day.

Publications

Refereed articles

Wadoux, A.M.J-C., Samuel-Rosa, A., Poggio, L. and Mulder, T.V. (2019). A note on knowledge discovery and machine learning in digital soil mapping. *European Journal of Soil Science*. Under review.

Wadoux, A.M.J-C., Brus, D.J. and Heuvelink, G.B.M. (2019). Sampling design optimization for soil mapping with random forest. *Geoderma*. Accepted with revision.

Wadoux, A.M.J-C., Heuvelink, G.B.M., Uijlenhoet, R. and de Bruin, S. (2019). Optimization of rain gauge sampling density for discharge prediction using Bayesian calibration. *Water Resources Research*. Accepted with revision.

Wadoux, A.M.J-C. (2019). Using deep learning for multivariate mapping of soil with quantified uncertainty. *Geoderma*, 351, 59-70.

Wadoux, A.M.J-C., Padarian, J. and Minasny, B. (2019). Multi-source data integration for soil mapping using deep learning. *SOIL*, 5, 107-119.

Wadoux, A.M.J-C., Marchant, B.P. and Lark, R.M. (2019). Efficient sampling for geostatistical surveys. *European Journal of Soil Science*. In Press.

Ramirez-Lopez, L., **Wadoux, A.M.J-C.**, Franceschini, M.H.D., Terra, F.S., Marques, K.P.P., Sayão V.M. and Demattê J.A.M., (2019). Robust soil mapping at farm-scale using vis-NIR spectroscopy. *European Journal of Soil Science*, 70, 378-393.

Wadoux, A.M.J-C., Brus, D.J. and Heuvelink, G.B.M. (2018). Accounting for non-stationary variance in geostatistical mapping of soil properties. *Geoderma*, 324, 138-147.

Wadoux, A.M.J-C., Brus, D.J., Rico-Ramirez, M.A. and Heuvelink, G.B.M. (2017). Sampling design optimisation for rainfall prediction using a non-stationary geostatistical model. *Advances in Water Resources*, 107, 126-138.

Stumpf, F., Schmidt, K., Goebes, P., Behrens, T., Schönbrodt-Stitt, S., **Wadoux, A.M.J-C.**, Xiang, W. and Scholten, T. (2017). Uncertainty-guided sampling to improve digital soil maps. *Catena*, 153, 30-38.

Stumpf, F., Goebes, P., Schmidt, K., Schindewolf, M., Schönbrodt-Stitt, S., **Wadoux, A.M.J-C.**, Xiang, W. and Scholten, T. (2017). Sediment reallocations due to erosive rainfall events in the Three Gorges Reservoir Area, Central China. *Land Degradation & Development*, 28(4), 1212-1227.

Stumpf, F., Schmidt, K., Behrens, T., Schönbrodt-Stitt, S., Buzzo, G., Dumperth, C., **Wadoux, A.M.J-C.**, Xiang, W. and 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(4), 499-509.

Conference contributions

Wadoux, A.M.J-C., Brus, D.J. and Heuvelink, G.B.M. (2019). Sampling design optimization for soil mapping with machine learning. Pedometrics, June 2-6, Guelph, Canada.

Mulder, T.V. and **Wadoux, A.M.J-C.** (2019). The importance of soil for global ecosystem modelling. Pedometrics, June 2-6, Guelph, Canada.

Wadoux, A.M.J-C. (2019). Using deep learning for multivariate mapping of soil with quantified uncertainty. Joint workshop for Digital Soil Mapping and GlobalSoilMap, March 12-16, Santiago, Chile.

Mulder, T.V. and **Wadoux, A.M.J-C.** (2018). A global-scale assessment of soil geography and soil-landscape functioning for modelling natural resources in a changing world. 21st World Congress of Soil Science, August 12-17, Rio de Janeiro, Brazil.

Wadoux, A.M.J-C., Heuvelink, G.B.M., Uijlenhoet, R. and de Bruin, S. (2018). Optimization of rain gauge sampling density for discharge prediction using Bayesian calibration. ERAD-10th European Conference on Radar in Meteorology, July 1-6, Ede, the Netherlands.

Wadoux, A.M.J-C., Heuvelink, G.B.M., Uijlenhoet, R. and de Bruin, S. (2018). Optimization of rain gauge sampling density for discharge prediction using Bayesian calibration. European Geosciences Union General Assembly, April 8-13, Vienna, Austria.

Wadoux, A.M.J-C., Marchant, B., and Lark, R.M. (2018). Optimal spatial coverage to support estimation of covariance parameters and minimization of prediction error for soil mapping. European Geosciences Union General Assembly, April 8-13, Vienna, Austria.

Wadoux, A.M.J-C., Brus, D.J., Rico-Ramirez, M.A. and Heuvelink, G.B.M. (2017). Sampling design optimisation for rainfall prediction using a non-stationary geostatistical model. Spatial Statistics, July 4-7, Lancaster, UK.

Wadoux, A.M.J-C., Ramirez-Lopez, L., Stumpf, F., Scholten, T. and Schmidt, K. (2017). Modelling the mid-infrared information content of European soils. Pedometrics, June 26-30, Wageningen, the Netherlands.

Wadoux, A.M.J-C., Brus, D.J. and Heuvelink, G.B.M. (2017). Accounting for non-stationary variance in geostatistical mapping of soil properties. Pedometrics, June 26-30, Wageningen, the Netherlands.

Cecinati, F., **Wadoux, A.M.J-C.**, Rico-Ramirez, M.A. and Heuvelink, G.B.M. (2017). Rainfall estimation using a non-stationary geostatistical model and uncertain measurements. Weather Radar and Hydrology-WRaH, April 10-13, Seoul, Korea.

Wadoux, A.M.J-C., Brus, D.J., Rico-Ramirez, M.A. and Heuvelink, G.B.M. (2016). Sampling design optimisation for rainfall prediction using a non-stationary geostatistical model. ERAD-9th European Conference on Radar in Meteorology, October 10-14, Antalya, Turkey.

Wadoux, A.M.J-C., Brus, D.J., Rico-Ramirez, M.A. and Heuvelink, G.B.M. (2016). Sampling design optimisation for rainfall prediction using a non-stationary geostatistical model. GeoENV, July 6-8, Lisboa, Portugal.

Other contributions

Wadoux, A.M.J-C. (2018). Epistemological aspects of soil science in the late nineteenth century, (Master thesis). Centre Francois Viète, Nantes, France.

Sawicka, K., **Wadoux, A.M.J-C.** and Heuvelink, G.B.M. (2017). Sample design optimisation techniques and associated software, (Report). QUICS project deliverable 3.1.

Wadoux, A.M.J-C. (2015). Mid-infrared spectroscopy for soil and terrain analysis, (Master thesis). Eberhard Karls Universität Tübingen, Germany.

About the author

Alexandre Wadoux was born on 19 October 1991 in Mont-Saint-Aignan, France. When he was five years old, his family moved to the village of le Puy-Notre-Dame in the South Saumurois, where he grew up surrounded by millenary castles, rivers, troglodytes and colorful vineyards. He attended primary school in Montreuil-Bellay and secondary school in Saumur at the lycée Duplessis-Mornay where he obtained his Baccalauréat in June 2009. The same year he started to study Geography in Angers, the closest university city, where he got acquainted with a large variety of subjects ranging from geomorphology and geology to economy and history. Alexandre soon realized that he wanted an academic career, and benefited from an Erasmus program at the Eberhard Karls Universität Tübingen (Germany) to finish his Bachelor degree in Environmental Geography. He pursued his study at the same university with a MSc in Physical Geography and Soil Science. In Tübingen, he learned about soil processes, GIS and applied mathematics, worked as intern at SERTIT (Strasbourg) mapping temporal forest changes using remote sensing and was employed as assistant in the soil analysis laboratory. Alexandre was involved in the Yangtze-GEO project which aimed at quantifying the soil erosion in the reservoir area of the newly built Three-Georges Dam in central China. For this reason, he went several times to remote areas of the Hubei Province to collect soil samples. Within this framework, he realized a Master thesis on soil spectroscopy under the supervision of Dr Ramirez-Lopez (ETH Zurich). After obtaining his MSc in 2015, Alexandre started this PhD as part of an International Training Network project funded by the European Union. Within this project, he worked with several of the twelve other project fellows and in total spent about ten months on secondments with project partners. At the same time, he obtained in 2018 a second Master's degree in Epistemology of Sciences at the University of Nantes. After obtaining his PhD, Alexandre continues his academic career as Research Associate in predictive soil mapping at the Sydney Institute of Agriculture.

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review of literature (4.5 ECTS)

- Spatial sampling design optimization

Writing of project proposal (4.5 ECTS)

- Sampling design optimization for uncertainty propagation

Post-graduate courses (7.5 ECTS)

- QUICS Training: technical and general skills; Sheffield, UK (2015)
- Geostatistics; PE&RC, NL (2015)
- Bayesian statistics; PE&RC, NL (2015)
- Hydrocourse: model building, inference and hypothesis testing in hydrology; LIST, Luxembourg (2016)
- Spatial and spatio-temporal Bayesian models with R-INLA; Bergamo, Italy (2016)
- Collecting spatial data; Southampton, UK (2016)
- Modelling techniques and uncertainty analysis framework; Delft, NL (2016)

Laboratory training and working visits (4.5 ECTS)

- Rainfall mapping; department of Civil Engineering, University of Bristol (2016)
- Bayesian calibration; department of Civil Engineering, Delft Institute of Technology (2016)
- Sampling for geostatistical mapping; GeoAnalytics department, British Geological Survey (2017)
- Proximal sensing for soil quality; Sydney institute of Agriculture, University of Sydney (2018)

Invited review of (unpublished) journal manuscript (2 ECTS)

- International Journal of Geographical Information Science (2018)
- Journal of Geophysical Research: Earth Surface (2018)
- European Journal of Soil Science (2018-2019)
- Geoderma (2018-2019)

Competence strengthening / skills courses (2.1 ECTS)

- Scientific writing; WGS, NL (2017)
- Presentation skills; QUICS training, Sheffield, UK (2015)
- Transferable skills training; QUICS training, Sheffield, UK (2017)

PE&RC Annual meetings, seminars and the PE&RC weekend (1.2 ECTS)

- PE&RC Weekend (2016)
- PE&RC Carrousel (2017)

D**iscussion groups / local seminars / other scientific meetings (4.9 ECTS)**

- Sheffield training event outreach (2015)
- Delft training event outreach (2016)

- Aquafin / Antwerp training event outreach (2016)
- Three day event at IIT Mumbay (2017)
- Amsterdam international water week (2017)
- Master courses on philosophy of sciences/ethics and epistemology
- Editor newsletter of the IUSS Pedometrics commission
- Discussion group modelling and statistics/R users

International symposia, workshops and conferences (8 ECTS)

- Geostatistics for environmental applications; Lisbon, Portugal (2016)
- European conference on radar in meteorology; Antalya, Turkey (2016)
- Pedometrics; Wageningen, NL (2017)
- European geosciences union; Vienna, Autria (2018)

Lecturing / supervision of practicals / tutorials (1.8 ECTS)

- Environmental data collection and analysis (2016)
- Advanced GIS (2018)

Supervision of a MSc student (3 ECTS)

- Sampling design optimization for mapping using machine learning

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

Colophon
Cover and chapter figures: Alexandre M.J.-C. Wadoux
Cover design: Design Your Thesis | www.designyourthesis.com
Printed by: ProefschriftMaken Utrecht