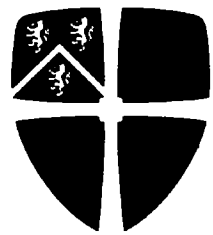
Bayesian sampling design for contaminated land investigation

Rebecca O'Neil

Bayesian sampling design for contaminated land investigation

Rebecca O'Neil

A Thesis presented for the degree of
Doctor of Philosophy



Statistics and Probability Group
Department of Mathematical Sciences
Durham University
England

June 2009

Bayesian sampling design for contaminated land investigation

Rebecca O'Neil

Submitted for the degree of Doctor of Philosophy
June 2009

Abstract

The problem of sampling design for contaminated land investigation is approached using Bayesian methods. We develop a decision tool designed to aid site investigators and decision makers in the process of site investigation. Current legislation and guidance is considered, and used to drive the development of a spatial model to describe the contamination levels over a site.

This model is updated using a full Bayes approach and combined with a detailed loss structure in order to calculate the expected losses associated with the possible decisions. A sampling search algorithm looks for good designs with which we can further update beliefs and improve decision making ability through reduced uncertainty and therefore increased confidence. We also offer an MCMC approach to learn about multiple contaminants which are believed to be related.

The decision tool provided offers a flexible environment in which multiple decisions, outcomes and contaminants may be considered simultaneously in order to assist the site investigator in implementing a cost effective sampling strategy.

Declaration

The work in this thesis is based on research carried out at the Statistics and Probability Group, the Department of Mathematical Sciences, the University of Durham, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

I need to thank a number of individuals who have been involved with this work. Firstly I am extremely grateful for the guidance and financial support of my CASE sponsor Ian Farmer (Ian Farmer Associates), along with the support of EPSRC. I have had the good fortune to have the expert guidance of my supervisor Michael Goldstein, as well as Karen Johnson and David Toll of the Engineering department, without whom my thesis would be about two pages long. They have all been patient and supportive throughout this project, even when I was completely missing the point and going off on a tangent. I have also had input from Bob Barnes at the Environment Agency, as well as contacts at the HPA and at local council level.

So many other people have helped me with this thesis I could write another 20 pages of thanks, but the following sweeping remarks will have to do ... Everyone who steered me back on track when I was having one of my *many* stupid moments (Jonathan and Danny in particular), to those who reminded me how to write in English as opposed to Beckyspeak (mainly my mother and Ric), and those of you that forced me to see this through, whether you helped me by distracting me with a beer and football (or Flirtini and a boogie Ben!), or by putting up with my sleep deprived stressed out rants and making me mashed potato and meatballs (Ariane!). I appreciate it all and could not have done it without you. To my other friends, family and colleagues I haven't mentioned by name, your friendship and support hasn't gone unnoticed and I will endeavour to buy each and every one of you a drink asap. Finally my parents, who made it possible for me to complete this with a roof over my head and a smile on my face (most of the time), I hope this begins to repay all you've done for me until I can buy you the cruise and car and retirement mansion you deserve.

Contents

Chapter 1

Introduction

Contaminated land comes under the remit of the Department for Environment Food and Rural Affairs (DEFRA) [1]. The Environment Agency (EA) and the Health Protection Agency (HPA) work to protect human health, through assessment of the pollution of controlled waters, and directly through contact with contaminated soil. Local Planning Authorities (LPA's) must consider the classification of contaminated land as part of every planning application. If the land is determined as contaminated under the legal definition outlined in the Environmental Protection Act of 1995 [28, 35], then the developer is responsible for remediation of the land. The determination is made under Part IIA of the act, if this occurs then the site must be placed on the contaminated land register. In practice contamination is usually dealt with at the planning stage, and monitored to ensure no future potential for harm, and so in reality the contaminated land register is very short.

1.1 Background

This thesis aims to consider the views of all relevant stakeholders involved in the investigation and remediation of a site in order to provide an appropriate decision tool for this purpose. This decision tool will aid the selection of sampling locations at the site by using a risk based approach. The stakeholders will include the site

[1] A glossary of acronyms is included on Page 238

owners, developers, consultants, remediation companies, end users of the site and government bodies; essentially all those who stand to gain or lose in some way depending on the outcome of the investigation or long term state of the land.

The most recent guidance from DEFRA suggests the use of Soil Guideline Values (SGVs), derived using the Contaminated Land Exposure Assessment (CLEA) model, to guide the assessment of an existence of significant possibility of significant harm (SPOSH) on the site. We will look in more detail at the regulatory guidance and the CLEA framework to highlight the main issues that need addressing.

1.1.1 Types of contaminated land, and sources

Generally the source of contamination problems come from human activity, although gaseous emissions or increased concentrations of certain elements can occur in soil naturally. The sources can usually be split into four groups: industrial, commercial, municipal, mineral extraction; all of which potentially present different problems.

Since the industrial revolution there has been a spread in the range of industry in the UK, which leads to a wide range of potential contaminants. These have a varying effect on redevelopment depending on how the site will be used and what pathways for movement of contaminants have been discovered.

There are several possible hazards related to the presence of these contaminants, all of which need to be considered. Some are relevant only in certain situations, but many will need to be addressed on all sites. Of key concern among these hazards are human health, ecological receptors and water. The EU Water Framework Directive (WFD) [39], now requires all waters (rather than just water for consumption), to be of a "good" status. The possibility of risk to groundwater should also be assessed. Once a contaminant reaches the water table it cannot easily be controlled and is more likely to enter surrounding water bodies. Other potential hazards include phytotoxicity, chemical attack (on buildings), fires (underground) and explosions from flammable gases, asphyxiation from exclusion of oxygen through presence of other gases, odours and radioactivity.

The occurence of contamination in the land may be extremely variable, and it is difficult to resolve all uncertainty regarding contamination levels due to this inherent

variability. We may find a surface of contamination that we can map through the site, or a number of isolated hotspots. A hotspot of contamination is a localised area of high contamination which does not belong to the underlying contamination surface.

As we will discuss, the guidance is changing on a regular basis; the authorities are aware of many problems associated with the classification of contaminated land and so the best we can hope to do here is present a tool which is as flexible as possible and which can be adapted as the information from the EA and DEFRA is updated. We do not propose to comment on or suggest alternatives to the scientific work behind the SGVs and CLEA model.

1.1.2 Source Pathway Receptor linkage

When a contaminant is present in the land, it can only present a risk to a receptor if there exists a viable Source-Pathway-Receptor (S-P-R) linkage, which must be identified by the site investigator and/or the LPA. In the pathway, the concentration of the contaminant may be affected by one or more potential chemical processes. There are two ways in which this can occur:

1. Mixing processes, such as diffusion and mechanical dispersion
2. Chemical reactions, which include sorption processes and chemical precipitation

Soil chemistry is the study of the composition, properties and reactions of soils. Environmental soil chemistry is the specific understanding of the reactions between soils and environmental receptors which are affected. Soil is made up of air, water, inorganic solids, organic solids and microorganisms. Reactions which may be relevant to the presence and spread of contamination include dissolution, precipitation, polymerization, adsorption/desorption and oxidation-reduction. These processes may affect the solubility, mobility, form and toxicity of contaminants [44, 58].

To be a potential danger to human health a chemical must be toxic, mobile and in a bioavailable form. Bioavailability is the degree to which, or rate at which a substance is absorbed or becomes available for use within the human body. These

considerations are taken into account in the EA documents outlining the derivation and selection of SGVs [23].

The number of potential contaminants in the land is large, and the effects of contact with humans/environment/buildings vary widely. When deciding on the potential risk to a receptor, the distribution of the contaminant of interest in the soil should be considered in order to assess the existence of an S-P-R linkage.

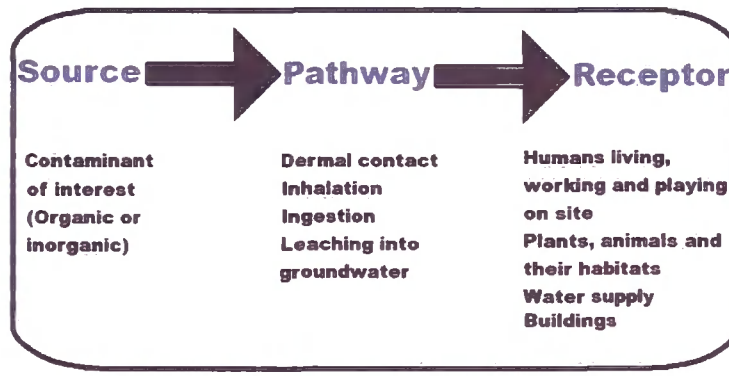


Figure 1.1: The S-P-R chain

Contamination is considered a "material planning consideration", and it is the responsibility of the developer and the LPA through planning permission to ensure the contamination status of a site is determined and dealt with through remediation where necessary. The main concern for the UK Government is the protection of human health, and we must find a way to quantify this concern and include it in our model. In this study we will only consider the human health risk through direct inhalation, ingestion or dermal contact at the surface and to a depth of approximately 100cm. However, there are other potential receptors to consider, as shown in Figure 1.1.

1.1.3 Regulatory guidance

There is no specific technical guidance which must be used in order to meet the requirements of the statutory documentation. Rather, a wide range of literature is available. The UK Government backed technical guidance is described below, and will form the main basis for any comparisons we make within this thesis; other

documents will be referenced as necessary. Local authorities can consult government agencies for specific issues, for example the HPA, EA and FSA (Food Standards Agency).

At the time of writing this thesis, DEFRA was having a major overhaul of its approach to contaminated land risk assessment, and has published the outcome of the "way forward" exercise on soil guideline values [30]. There is some concern within the industrial community that the new guidance may not fully address the problems associated with contaminated land investigation, and that there is still a need for even more detailed information and guidance. Developers and local authorities are finding it difficult to pass consistent judgements on the status of sites and often find it hard to meet the strict guidelines which have been set (many think with little scientific backing, due to the scientific information not being available).

The main guidance for determining whether land must be classified as contaminated is covered in Contaminated Land Reports (CLR) 7 through 10 [21–24], while CLR 11 describes the effective management of the land, and gives a large number of references to aid the decision maker at every stage of the process [25]. An older report, CLR 4 [20] discusses the selection of samples, and we will look at this in further detail when looking at sampling methodology in Section 2.1.2. Table 1.1 gives a brief summary of the content of the CLR series.

DEFRA has announced the withdrawal of documents CLR 7-10 and the CLEA model which will be replaced in due course and new guidance is being prepared. The "current practice" described here will relate to the CLEA UK software based on the CLR series 7-10, as this was current for the main duration of this work. However, this is not a major issue as the decision tool provided within this thesis is flexible enough to deal with changing goalposts and can be adapted as new generic and/or site specific information becomes available.

In [36], the EA gives an overview of the steps required in the investigation of potentially contaminated land. These are shown in Figure 1.2. Some of these steps involve an element of overlap. The document gives a checklist for each stage, highlighting the steps to be taken. The stages of site investigation that we consider are those of initial sampling, and subsequent decision-making with respect to further

Report	Description
CLR 4	Sampling guidance, including information on : the number of sampling points, their location, and the depth of samples.
CLR 7	Basically an introductory report which outlines the legal framework and the development of the SGVs
CLR 8	Identifies the priority contaminants. That is, those that are likely to be present on many current or former industrial sites in the UK, and at concentrations sufficient to cause harm
CLR 9	Gives the scientific background to the approach used to select tolerable daily intakes (TDIs) and index doses (IDs) which are used in the derivation of the SGVs
TOX reports	Detailed derivation of the TDI and ID values for specific contaminants
CLR 10	The Contaminated Land Exposure Assessment Model. See Section 1.1.4
SGV reports	These give the derivation of the individual SGVs
CLR 11	This report is designed to assist with the process of site investigation. Provides a hierarchy of information, looking at the risk management framework, technical detail and and many references to other useful guidance

Table 1.1: Summary of the contaminated land reports (partially taken from CLR 7 [21]), and related documents

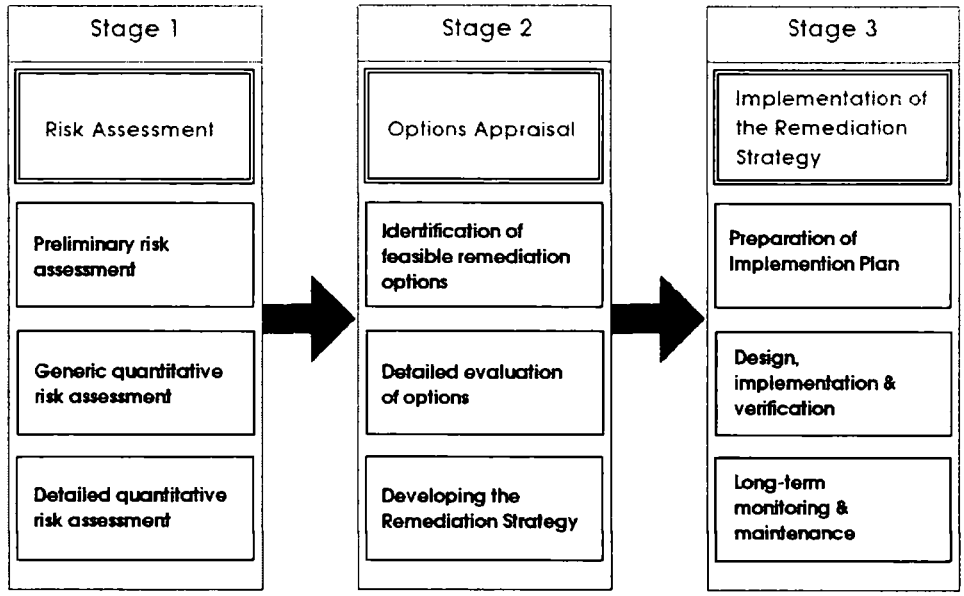


Figure 1.2: Overview of Model Procedures, reproduced from [36]

stages of sampling and choices of whether to remediate. These are vital procedures and should aid the efficient compilation of the site report at all three of the stages outlined in Figure 1.2. We propose an alternative to the current approach for assessing contamination levels and risk, by modelling the contamination levels over the site, and by developing a Bayesian approach with which we can update our beliefs in order to direct the next stage of sampling.

1.1.4 CLEA framework

The Contaminated Land Exposure Assessment (CLEA) framework was jointly issued in 2002 by the EA and DEFRA. It aims to assist the developer in assessing the potential exposure to potentially harmful levels of a contaminant on a site. The CLEA model takes into account many factors on the site and returns a value which can be compared with representative site measurements. This is the point at which we enter the process. If an SGV is exceeded, the assessor will need to decide whether a significant possibility of significant harm ("SPOSH") exists. The SGVs at present are very conservative; in some cases, exceeding them even by 10 times may not lead to a conclusion of SPOSH. However, their purpose is to alert the investigator

to a potential risk, and to tell us that a particular contaminant warrants further investigation.

The EA has outlined the level of information required within site investigation reports, and in particular the potential risk must be clearly addressed within the report. Most site investigations are carried out in order to begin redevelopment on sites, and, if the site investigation report does not contain the necessary detail and information, planning permission may be refused by the LPA. The EA guidance follows a risk-based framework and CLR11 lists the requirements for every report [15]. As well as assessing the risk to human health, the EA has published guidance relating to ecological risk assessment [38]. This introduces soil screening values (SSVs) as an indicator of potential harm to ecological receptors.

1.2 Key issues

The problem of site investigation is multi-faceted, and requires a thorough set of tools for analysis. We divide this problem into three key areas and consider each individually, then combine our methods to form a decision support tool for site investigation. Every site encountered will have different features. We must ensure that the analysis that our methodology provides is as site specific as possible. The three key areas are:

1. EA guidance and legislation requires that site investigators consider the amount of any contaminants present in the soil, and the impact on human health and ecological receptors of these contaminants; so we require a probabilistic model with which we may describe uncertainty about the levels of contamination.

2. Having modelled the spatial distribution of contaminants, we apply Bayesian decision theory to determine an effective method for quantifying and valuing information gain. This will involve the expert specifying the possible decisions we can take, and the associated costs which arise from taking each decision and observing a particular outcome (and therefore a related consequence).

3. Finally, the problem of "optimal" sampling selection is integral to effective

decision-making, and we consider efficient approaches for sample selection. As well as the selection of sampling locations, we will need to make decisions regarding the number of stages of sampling required. In this thesis, we refer to each stage of sampling as a sample, and each individual sampling point as an observation.

The process we will follow is summarised by the flow chart in Figure 1.3.

When looking to make decisions regarding sampling schemes and remediation alternatives, several factors must be taken into account, as outlined below.

- The opinions and expectations of all relevant stakeholders must be considered. That is, we need to elicit beliefs regarding the contamination levels over the site as well as information required for the construction of cost and consequence functions. All involved parties would not necessarily be approached regarding every aspect of the investigation. Rather, the relevant "experts" would be consulted for each aspect of the analysis. All stakeholders should be involved at some level in order to obtain as much information as possible from the site.

- A detailed site history, and any hard data collected from the current investigation (or previously collated information), may be combined with expert belief statements to develop a prior probability distribution for contamination over the site before any (more) sampling takes place.

- The type of sampling we undertake should be carefully considered to meet the requirements of the study. A combination of intrusive and non-intrusive methods is likely to be used to ensure a detailed description of the land may be taken without excessive intrusive investigation.

- There are several types of uncertainty present in the investigation of contaminated land. These need to be separated and analysed, and results communicated effectively to the decision makers [81]. We will look in detail at the different levels of uncertainty in Section 2.1.5.

- Whilst a model which is as realistic as possible is desirable, a balance must be struck between complexity and ease of specification and implementation. An

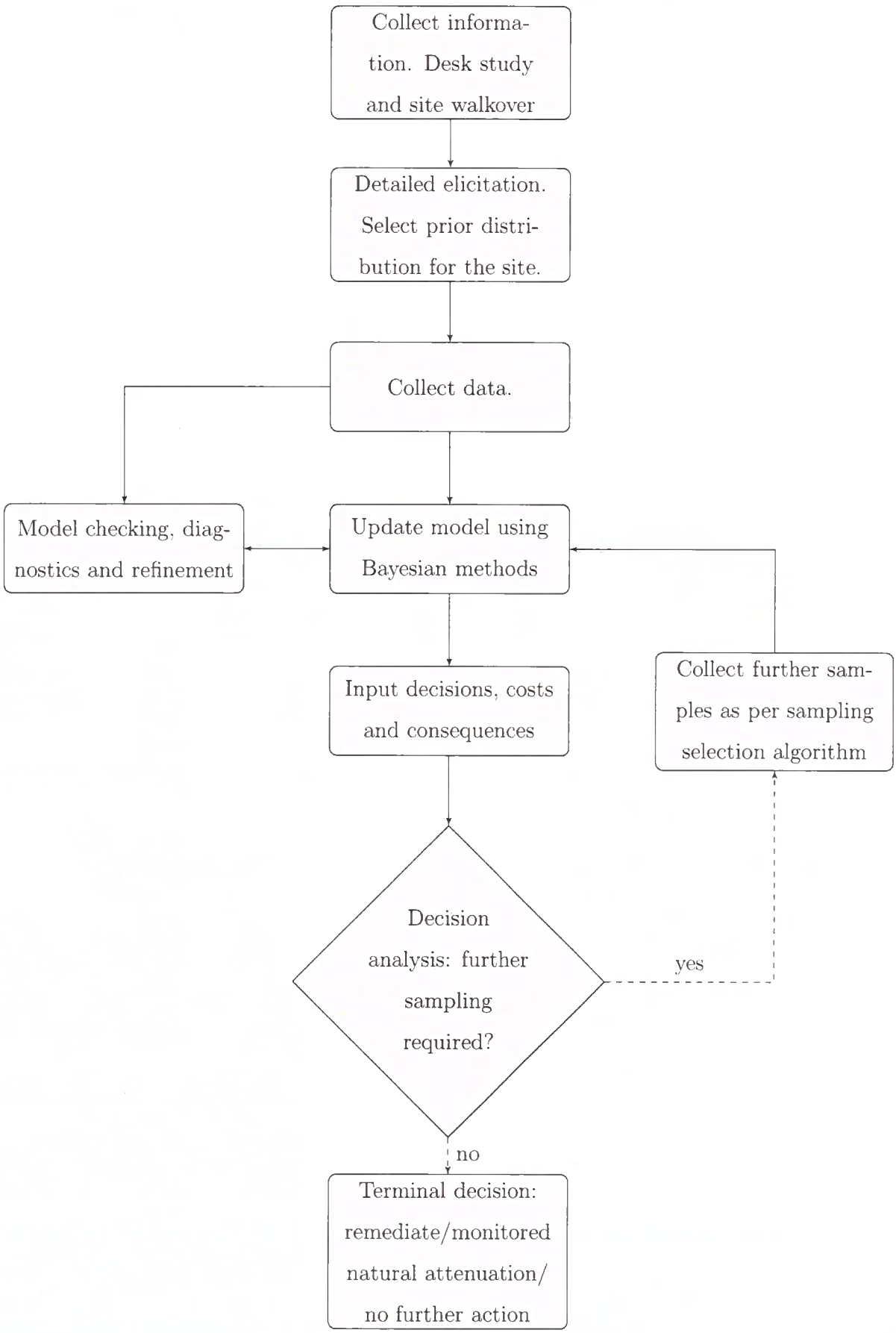


Figure 1.3: Flow chart of steps required, remedial options outlined in Section 2.3

algorithm for the selection of sampling points which takes many days to run may give theoretically excellent results and pinpoint an exact set of locations for maximum information gain. However, on site, it may not be feasible or worthwhile to place a borehole this precisely.

- The total financial resource for site investigation is usually quite limited, this places practical limitations on the sampling search algorithm.

It is important to note that the analytical methods developed in this document are based on subjective assessment of a number of parameters and therefore require a detailed sensitivity analysis to consider how changing beliefs would affect the outcome. We will perform a sensitivity analysis, with both hypothetical and real examples, to help understand how effectively our model is performing. For example, the opinions regarding expected contamination levels over the site may come from a single "expert", and could vary markedly from the beliefs of another "expert" asked the same questions. At the other end of the scale, when collecting information for the construction of utility functions (see Chapter 5) to quantify the preference for a particular outcome, the group as a whole may not be able to agree on a choice of function, and so methods of compromise will need to be considered.

Every site is unique, and as such, investigation will always be site specific in nature. This has always been the case with site investigation, as subjective assessment has been taking place in the form of the desk study. The methods introduced here are fairly general, and may be applied to a wide variety of case studies.

To validate and demonstrate the ideas and methods developed in this thesis, we will introduce several datasets. Again, sensitivity analyses will be an important feature of any case studies discussed, to demonstrate the advantages of subjective probability assessment, and to point out any problems or shortcomings of the method which need addressing.

1.3 Thesis outline

In Chapter 2 the current statistical guidance will be discussed, and the problem we will consider is identified and important features are highlighted. A model with

which we will describe the contamination problem will be introduced in Chapter 3, as well as Bayesian methods for the updating of beliefs through data collection. This chapter will consider updates for single contaminants, and a multiple approach when related contaminants are to be considered. A closed form approach will be introduced, and then we look to a conditional conjugate form which will be necessary for the multiple contaminant update. We will illustrate the methodology with a hypothetical example, before going on to perform analyses for the real case study in Chapter 4.

Combining this model with statistical decision theory will allow the implementation of sampling strategies, for both initial investigation and further stages of sampling; along with a tool for effective decision making under uncertainty. This will be introduced and developed in Chapter 5, for single contaminant decision making with simple, fixed costs, and then expanded to allow for more complex loss functions, and multiple contaminant decision making. Once we have demonstrated the way in which we shall calculate the expected value of a single sampling design, Chapter 6 will consider a strategy for searching among designs to select one that is optimal in some sense. Finally, we shall bring together the modelling of Chapter 4 and the methodology of Chapters 5 and 6 to perform a decision analysis and sample selection for the case study in Chapter 7. In Chapter 8 we conclude with a summary of the work undertaken in the thesis and an indication of further research arising from the issues raised.

Chapter 2

Site investigation and current methods

In order to understand how we can best analyse the contamination over a site in order to make sampling and remediation decisions, it is important to consider the current sampling and analysis procedures used in the industry. We will look at the problems associated with the current statistical approach, and discuss the different forms of uncertainty which must be quantified in order to make decisions with an informed view of the associated risks. We then go on to introduce the idea of Bayesian decision theory and the structure of the problem.

2.1 Site investigation methodology

The investigation of contaminated land involves a combination of qualitative data coming from experts, and quantitative data from sampling and statistical analysis. It can be broken down into three broad steps, risk assessment, risk management and risk reduction, as in [58, 82]. The mitigation of risk is included in both the management and reduction stages. In particular the risk assessment stage contains a further three stages: hazard identification, risk estimation and risk evaluation. The risk assessment stage is the one to which we pay most attention in the statistical sense as it influences all the future decisions. The aim is to discover and describe accurately the possible pathways for movement of contaminants. In order for there

to be a risk and therefore a need for remediation or some other form of management of the situation, there must be a complete S-P-R chain. This implies that even if a contaminant value is in breach of a government guideline of acceptability, if there is no pathway for movement, or indeed no sensitive receptors, then there is no risk and no statutory demand for remediation. The existence of a valid S-P-R chain, or indeed the non-existence, will determine the possible consequences, to which we require the assignment of loss values in order to carry out a decision analysis for the site. These consequences need not be the same throughout the site, and so we can equivalently assign loss values as a function of location. For example the top half of a site may be accessible and available for a child to play on, whereas the bottom half may be completely inaccessible. We would be more concerned about contamination levels in the top zone, and thus could assign higher loss values to locations in this area. However, this is a complex issue, and the expert would have to take into account many factors, such as the potential for the contamination to spread, alternative pathways (such as dust inhalation) or alternative receptors (such as ecological receptors in the bottom half of the site).

At the time of writing, the EU was pushing for a Soil Framework Directive [40], in order to have a EU wide legislative stance for the treatment of soil. It argues that soil is an important resource, and any contamination should be cleaned up, regardless of whether a complete S-P-R chain exists. If this is implemented, it would mean that loss at all points on the site would be the same, and we wouldn't have to worry about assigning complex loss functions dependent on location.

A natural way to approach the problem of site investigation is by carrying out a decision analysis to assess how many samples to take and their optimal location, and, given the observed contamination values, to make a decision on whether to continue sampling, take no action, or remediate. There are costs associated with all of these choices, and a sensitivity analysis will allow us to study how varying any of the costs will affect our decisions.

We can divide the investigation into the collection of both qualitative and quantitative data, as well as the collection of prior information off-site, which helps lead to better sampling methods and design. We look particularly at risk assessment to

begin with, to get an idea of the methods currently in place. We need to be able to collect enough reliable information about the site, as well as being able to identify all possible problems which may arise, without incurring huge costs at the planning stage.

2.1.1 The desk study and conceptual site model

We aim to make prior judgements which utilise as much information as possible, in order to improve our decision-making. As such, we can take advantage of the fact that, as part of every site investigation, a comprehensive study of the site history and previous investigations is compiled. These prior judgements may be made informally within the desk study, and we look to use these judgements to make a formal prior statement of belief. Within the guidance literature there is information relating to the compilation of the desk study, and what information should be included. In CLR 11 [25], this stage is referred to as the preliminary risk assessment.

History of land use

The majority of land contamination has occured since the industrial revolution, although some can be dated as far back as Roman times. There are sites which have only been contaminated relatively recently, and those which have seen a range of different contaminating activities over a long period of time. As such, a large part of developing site knowledge has to be involved in the desk study, looking at available historical data for the past uses of the site, which will enable the creation of a prior distribution with which a Bayesian approach can be implemented. This stage aims to establish what is already known about the site, and a set of conditions that can reasonably be expected.

As well as documentary sources, local knowledge, for example that of previous site workers, may provide insight that a map cannot. This will lead to the development of a conceptual model which will be used to specify a sampling strategy for data collection in order to verify/refine this model as a basis for making decisions about remediation. This stage of the investigation will also include site walkovers and possibly a small amount of exploratory work on site in order to rule out the

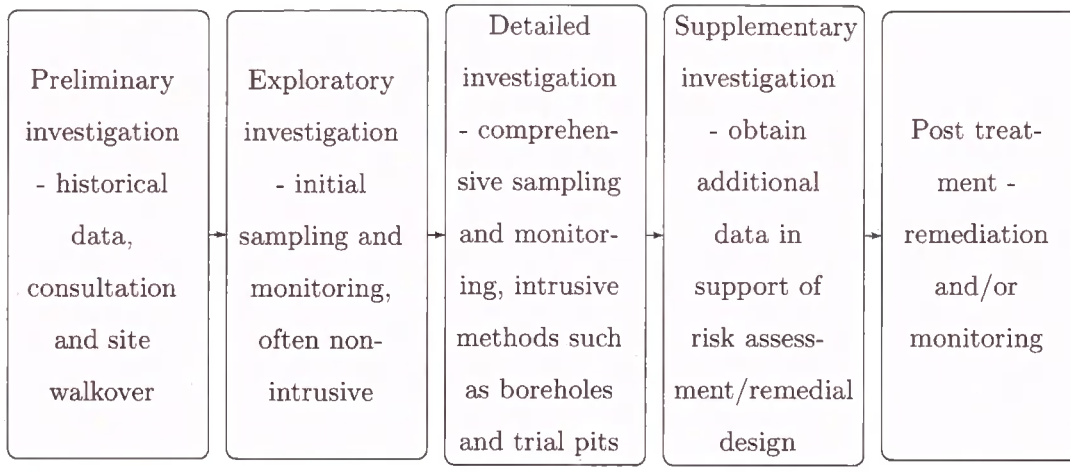


Figure 2.1: Five main stages of a site investigation

possibility of any hazards which may need urgent attention and to determine the geology and topography of the site.

2.1.2 Sampling practice

A good sampling strategy should aim to identify high levels of contamination and reduce the uncertainty concerning the spatial distribution of contamination. The "current" method for assessing contamination in the guidance does not allow for a quantification of the uncertainty at specific locations, and so it becomes difficult without a spatial model to compare the relative worth of candidate sampling locations. If we cannot describe the uncertainty associated with observations or predictions, it becomes very difficult to define criteria for decision-making. As a rule of thumb, to reduce the uncertainty in a sample by a factor of n, we must increase the sample size by n^2 which can become very costly very quickly [60]. Sampling choice will depend on how much the client is willing to spend on investigation, which may well depend on how much the client is willing to spend on "failure". "Failure" encompasses the two possible erroneous outcomes: cost of remediating when the site is clean and the cost of not remediating when the site was contaminated. The relationship between these two extremes drives the whole decision problem as we shall see in Chapter 5.

At the most basic level, there are two ways to sample: targeted and non-targeted.

Most site investigations will use a combination of these; targeted, in order to assess the levels of a suspected hotspot, whilst non-targeted sampling can be used to confirm the distribution of contamination on the site, and potentially hit any undiscovered hotspots.

Five main stages of investigation are suggested, as shown in Figure 2.1. The stages concerned with sampling design are the second to fourth, and to an extent these are influenced by the first stage, and as such are likely to be different for each site. Comparing this flow chart to Figure 1.3, which describes the steps we will follow in the procedure developed in this thesis, we see that the two are similar, and so the tool we aim to provide should naturally complement and enhance the current practice of decision making.

First it has to be decided what methods of data collection are to be used. The best approach is likely to be some combination of non-intrusive and intrusive methods:

- **Non-Intrusive** - Surface gas emission testing, geophysical testing, false colour infra-red technology, thermography, tracer gas testing, XRF. Ground penetrating radars can be used to map the subsurface and locate any anomalies that require investigation. Non-intrusive methods are generally implemented first as they are cheaper.

- **Intrusive** - Boreholes, trial pits and trenches, probing techniques, window sampling and gas and water monitoring wells. Intrusive methods do not cause an issue in terms of destructive sampling as small quantities of soil are removed, and much is put back (for example in trial pitting). The only potential issues may be cross contamination, or the introduction of a new pathway for movement of contaminants (for example a borehole may be drilled which redirects a water flow direction).

Non-intrusive techniques can be used to look for patterns in the ground or vegetation which indicate contamination, location of buried hazards, and investigating soil conditions. Generally intrusive methods will be used to back up the findings of non-intrusive methods, along with the collection of sub-surface data. The use of

boreholes tends to be the most common intrusive method, along with trenches and trial pits. Trial pits are a cheaper option but can only give information about the top few metres of the soil, whereas boreholes are more expensive but give information about groundwater and contamination at depth. The selection of sampling locations and methodologies are discussed further in [6, 10, 58].

2.1.3 Sampling strategy

An optimum sampling strategy should consider the following factors:

- Number of stages of sampling. This may be predetermined, or an iterative process, repeated until it is determined that a decision can be made at an acceptable level of confidence.
- Number of observations in each sample. We may have a fixed budget, and we search for the optimal sample design to fit within this budget, or we continue selecting observation locations until a stopping rule is met.
- Choice of sampling pattern, and position of observation locations. It may be decided that a predefined grid is sufficient for the investigation, or extra points may be required within the grid to further determine quantities such as correlation length (the distance beyond which two locations will be deemed unrelated).
- Whether replicates are required. We can take replicates at a location to attempt to determine the measurement error and natural variability present in the soil.
- The planned end use for the site, as mentioned previously this may direct sampling to an area of the site where potential receptors may be particularly vulnerable to adverse levels of contamination.

In many cases contaminants are located in hotspots, and we would like to know how confident we are that we have found all hotspots in the site. Possible sampling patterns include a regular grid, simple random and stratified random sampling, and are

described by Ripley [76]. Stratified random sampling means that the site is divided into equal subzones and then an equal number of observation locations are selected at random within each subzone. CLR 4 [20] gives four criteria for the most effective sampling; the optimal sampling scheme should be stratified, each stratum should contain one sampling point, points should not be aligned and sampling should be systematic (i.e. the sample selected should follow some design and not be random). This may be the case when we do not have a detailed spatial description of the site. However, we plan to introduce a sampling search algorithm which takes into account several factors and selects locations based on their ability to reduce uncertainty and potentially change decisions. The report also gives a discussion behind the reasons for these particular criteria. The herringbone sampling pattern (Figure 2.2) fits these while the other three listed in CLR4 do not. To construct the herringbone grid, a regular square grid is taken and points are offset by a quarter of a unit. The square grid is nearly as effective as the herringbone at finding a hotspot of size 5% of the site at a certain off-set distance proportional to the size of the site, according to CLR 4. In practice the square grid tends to be favoured for its ease of implementation. While sampling is the key way in which we learn about the levels of contamination, it is possible, due to cost, time and accessibility issues, that a sampling scheme is poorly implemented. We need to address this problem if we are to offer a decision tool for sampling strategies, by effectively communicating the benefits of carefully planning and undertaking sample collection. A discussion of sampling locations and intensity can also be found in [88].

2.1.4 Initial analytical methods

The data obtained from sampling can be analysed with one or more of the available methods from spatial statistics [19,76], and more specifically geostatistics [14,64], in order to determine the distribution and nature of contaminants at a site. Often the results of these analyses are taken to be illustrative of the underlying distribution of the site. However, there is an imperative need to understand that there will always be uncertainties involved and we need to be able to quantify these in order to best advise non-specialists on decisions regarding remediation of a contaminated land

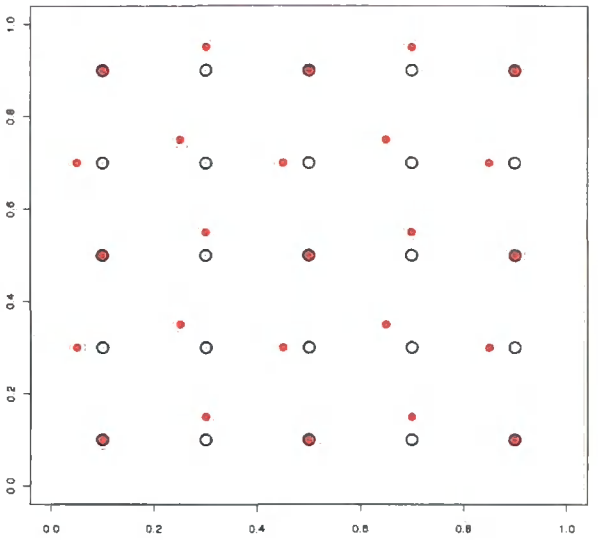


Figure 2.2: Structure of a herringbone grid, and relationship to a regular square grid. We see that the herringbone figuration is created by offsetting each point of the square grid by a quarter of a unit.

site. We will introduce our approach for analysing the data in Chapter 3.

Risk evaluation forms the latter part of risk assessment and the initial stage of risk reduction. The validity of any statistical findings need to be judged, accounting for uncertainty, current policy/guidance (i.e. critical contaminant levels in particular) and variability of risk in terms of the costs of remediation/failure to remediate when it is necessary. The final stage of the process (Figure 1.2), risk reduction, involves collation of the data and presentation of the findings and recommendations to the decision maker. There are three methods of assessing risk:

- **Qualitative**, an assessor ranks risk as high, medium or low according to personal judgements based on desk study, or other qualitative information

- **Semi-quantitative**, formal protocol is followed, and critical levels are used to judge risk. This is the situation suggested at present where a Risk=Hazard × Likelihood approach is applied, where the hazard is the contaminant of concern and the likelihood represents the probability of the hazard existing at a dangerous level. SGVs are used as screening values to determine a need for further investigation. Detailed quantitative risk assessment involves the derivation of

site specific levels of contaminants for assessment. The DEFRA website covers the topic of risk assessment [29]. They give three primary factors to consider when estimating the probability of an outcome: will the outcome occur; will exposure to the hazard occur; and will harm result following exposure.

- **Quantitative**, a statistical approach, which is much more comprehensive and complete although seen by some as too costly and time consuming.

2.1.5 Sources of uncertainty

Applying statistics to the results of a fairly basic sampling scheme, and comparing analysed data with guideline values (as suggested by the regulatory guidance), can potentially reduce the costs involved with site investigation and remediation and lead to better decision making. However, to best determine the probability of making a wrong decision, we need to recognise limitations of the sampling process. If we do this we can further consider ways in which we may present a more effective sampling methodology. The contamination values we observe will be different from the true level of contamination present at that location on the site, for several reasons. The basic concepts are introduced in [13], while a comprehensive description of the sources of uncertainty may be found in [73].

1. When an observation is taken it is subject to "natural variability", or small scale heterogeneity. This means that measurements are not repeatable as fluctuations occur on a microscopic scale. This is a significant problem associated with site investigation as there is no way around this source of variation. However, we can attempt to put a figure on the levels of fluctuation occuring.

2. There is a large-scale variation which relates to the site as a whole. This will depend in part on the strength of the relationship between locations.

3. Human and equipment errors must be accounted for. These again can be separated into error on site (i.e. faulty equipment leading to observations being made at the wrong location, or cross contamination of samples), or error at

the analysis stage in the laboratory. Certified reference materials can be used to determine the bias of the method, see [13].

In the next chapter we will incorporate these different types of uncertainty into the model we develop to describe contamination levels. The EA and others recognise the need for a clear explanation of the sampling strategy implemented:

> "Every site investigation should have a clear sampling strategy which gives specific and stated reasons for each sample collected." [65]

Initial sampling should be sufficient to determine the presence of a significant S-P-R linkage. However, additional sampling may be required in order to determine the extent of the contamination and the levels of uncertainty. Desk study and site walkover drives the selection of which substances to test for.

2.2 Statistics used in current methods

The statistical test offered by the EA is the Mean Value Test (MVT), and is detailed in CLR 7 [21]. To account for the fact that contamination concentrations vary across a site, the purpose of the test is to state that the population mean is less than the SGV with a specified level of confidence. If this cannot be shown, and the test shows that the population mean may be above the SGV, further considerations are required to make a decision regarding classification of the site as clean or contaminated.

The MVT takes the arithmetic mean of a set of observations, and calculates an upper confidence limit in order to take account of the uncertainty regarding the true mean levels of contamination in the site. The confidence limit is calculated using a t-value chosen dependent on the number of observations. This value is then compared with the relevant guideline value to determine whether a SPOSH may exist. The example below shows the calculation of this confidence limit for two cases.

The test treats all values as independent, rather than spatially dependent as we would expect them to be. When the calculated value exceeds the SGV, it is suggested that the Maximum Value Test provides a method to determine whether the large values in the dataset may be statistical outliers. If the test indicates

that a value is to be treated as an outlier, the area may be investigated as an area of localised contamination. If the test indicates that the value belongs to the underlying population this also suggests the need for further investigation, as the MVT has failed and therefore there is a potential risk.

Several problems arise when implementing the MVT; these include the spatial nature of the problem, how to zone the land, and the difficulty of communication of results with decision makers, who tend to work in financial terms (loss). These problems have been noted both by the Environment Agency and others; Nathanail [65] comments, "CLR 7 is however silent on the spatial distribution of samples or on how to deal with heterogeneous ground conditions". A discussion of sampling strategy is given in CLR 4, but is now somewhat outdated and basic [20]. A more recent document uses a similar approach of hypothesis testing based on a t-test [11].

The main issue is that the test does not allow for a spatial relationship between sampled locations when making statements about the mean value and related confidence bounds. Spatial correlation is relevant particularly in the case of clustered locations, as we can no longer treat each sample as contributing equally to estimates. To account for the uncertainty relating to the unknown mean and variance parameters we can incorporate Bayesian inference; this is the key difference between just kriging [1] the data, and learning about the underlying parameters which generate the surface of interest.

Example of the MVT

Figure 2.3 shows two potential sampling configurations for the same site, with 10 locations selected in each. This is an extreme example that highlights why the MVT fails to account for the spatial distribution of the sample locations. The values indicate contamination levels observed for some contaminant of interest. The sampling locations have been arranged in order to demonstrate that we obtain less information from a cluster of samples, which is something the MVT does not consider. The

[1] We will describe kriging in Section 3.3.3, it is a prediction method based on a weighted linear estimator.

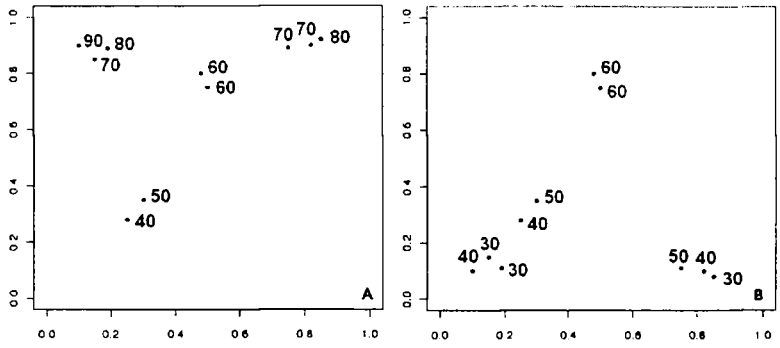


Figure 2.3: Demonstration of the problems arising when using the CLR7 test. Two possible sampling configurations for the same site, both with 10 sampling locations. Numbers represent some "units of contamination". See calculation of MVT for each configuration in text.

values obtained from the application of the MVT to the two sets of sampling results shows how disregarding the spatial location affects our analysis;

- Calculate arithmetic means, 67 and 43 for configuration A and B respectively
- Calculate sample standard deviations, 14.944 for A and 11.595 for B
- Select the appropriate t value, 10 samples in both cases, so 9 degrees of freedom, t=1.833
- Calculate the upper 95th percentile bound of the sample as $US_{95} = \bar{x} + \frac{t.s}{\sqrt{n}}$
- Obtain US_{95} values of 75.66242 and 49.72102 respectively

These two configurations would be seen as equally informative by the MVT, as spatial location is disregarded. However, we see considerably different upper confidence limits obtained. If the SGV was 60 for example, we would reach different conclusions about the state of the land, and potentially make costly erroneous decisions based on this test. Also, there is potentially a linear trend in this site from north to south, which we should be considering when calculating upper bounds, as delineation of zones will depend highly on these trends. We will come back to these values when a model has been introduced with which to consider trends, correlation and uncertainty.

As the updating of the guidance continues, the EA has commented on its website [37], that "The statistical guidance contained within CLR7 represented a starting point for the interpretation of site data... However, there is not one single approach applicable to all sites and circumstances and the wider range of robust statistical techniques developed by organisations including the United States Environmental Protection Agency are also important tools." It also suggests the use of the guidance published by CL:AIRE and CIEH, [11], which expands on the guidance from CLR 7, and stresses the importance of ensuring a sample is representative of the site in order to use critical values (i.e. SGVs in the UK, when available, or some other relevant value as determined by the expert). The US EPA documents [83, 84], give a more detailed quantitative approach to the initial and further examination of potentially contaminated sites, through the use of a scoring system based on several factors.

2.3 Remediation

There are three possible ways to deal with a contaminant in the soil once it has been decided that a SPOSH exists. The remediation approach will be dependent on which of metals, nonmetals or both are present on the site. In general the two types of contamination will have to be dealt with using separate methods. However, there are some approaches which may deal with both. We can decide on one of these options in order to remove the risk present:

1. Source control. In this case the contaminated soil could be removed and sent to landfill. This option is becoming increasingly expensive as the government attempts to achieve the targets set in the EU Landfill Directive [31, 41]. In-situ methods of source removal are also available, such as bioremedition or soil washing.

2. Pathway interruption. Methods of blocking or removing the pathway include a grout curtain or capping of the land to prevent further migration of contamination. Solidification/Stabilisation is also a method of pathway interruption which involves the addition of cement or other additives to limit the solubility or mobility of the contaminants [58].

3. Receptor relocation. For example if the site was residential, the residents could move house.

The first would be the most desirable although this can be expensive, intrusive and time consuming. That is, future risk may not be ruled out entirely, as we cannot be sure all contamination has been removed. Some methods of in-situ remediation are becoming more widely available and cost effective. Organisations such as CL:AIRE work on raising awareness of, and confidence in, practical and sustainable remediation technologies. Pathway interruption would be the next most preferable option, as relocation of the receptor is generally not feasible, and could result in a site being left empty, with no real resolution of the problem. So, pathway interruption and source removal would be the two strategies most likely to be considered in an options appraisal.

2.3.1 Remediation methods

Once one of the options above have been selected, a methodology must be decided upon. There are two main methods of remediation, in-situ and ex-situ. [5] gives an in depth discussion of both in-situ and ex-situ methods, we give a few options here, as discussed in [55]. Some of the methods listed may be performed either in-situ or ex-situ.

In-situ remediation

This method looks to remediate the soil without removing it from the site, and includes options such as inspection wells for pump and treat technology, traditional covers, geochemical covers, slurry walls, grout curtains, chemical oxidation, chemical reduction, stabilisation/solidification, vitrification, soil vapour extraction (SVE) and Monitored Natural Attenuation (MNA).

Ex-situ remediation

In this case the soil is removed from the ground and treated either on site or taken away. Options include landfill, neutralisation, solvent extraction, soil washing, ther-

mal desorption stabilisation/solidification, vitrification, bioreactors and incineration.

Monitored natural attenuation

Monitored natural attenuation, (MNA) [85], is effectively a method of in-situ remediation, but does not require any action other than continued observation of contamination levels over time. Natural processes such as bioremediation reduce contamination levels over time. Natural attenuation often occurs to some level at sites where contamination is present. If used as a method of clean up, this would be a long term solution as these processes must be monitored and verified to ensure they are succeeding in reducing levels below the dangerous values in a reasonable time frame.

The method selected will depend on the type and concentration of contaminants present on site. CLR 11 [25] gives a matrix of feasible options to show which methods can be applied for which contaminants. We can include the option of several viable methods in the decision set-up, which we will discuss in more detail in Chapter 5. We will need to find a way to quantify factors such as the effectiveness, time required and availability of the method chosen in order to compare options.

2.4 Problem description

Now we have considered the current industry practice, we can outline the main features of the problem we shall consider in the remainder of this thesis. It falls into two main stages, that of constructing a probabilistic model to describe the spatial distribution of contamination on the site, and combining the model constructed with a decision analysis and search algorithm to determine the optimal course of action. Deciding if and where to remediate is a complex problem; there are many possible outcomes to consider. The consequences associated with making a "bad/wrong" decision may be far more damaging than a monetary penalty.

2.4.1 Spatial aspect

We have outlined the problems associated with using the MVT and shown that calculating the mean cannot fully describe the contamination levels and associated uncertainty over a site. We need to develop a method which will take into account the observation locations, and give us a framework in which we can learn about the mean and variance on site and update our beliefs. This will involve combining Bayesian methods with spatial statistics, whilst bearing in mind practical computational constraints, and is covered in Chapter 3.

2.4.2 Decision making and sampling design

Decision theory is concerned with how people (decision makers, or DMs) make decisions, and with how optimal decisions can be reached. A decision analysis is not intended to replace the DM's (hopefully) rational and coherent process, rather it provides a method to subject their personal preferences and beliefs to explicit tests of coherence [7]. We will use this theory in Chapter 5 to develop a way of measuring the performance of sampling designs in terms of the expected benefit or loss, in order to compare them.

We use expert judgment and information from the desk study to build a "good" prior probability distribution for the parameters of interest (i.e. the mean and variance of each contaminant), and then use site data to update this to a posterior distribution using Bayesian analysis. The posterior distribution that is produced can be used to predict values, and methods have been developed to allow sampling from increasingly complicated distributions (as we discuss and apply in Chapter 3). We would like to utilise the wide range of prior information in the building of prior probability distributions. However, these can be made as cautious as necessary, depending on the expert's confidence in the quality of the prior information. Each analysis will be unique to the expert consulted, as they will express their own subjective probabilities which drive the whole procedure. We will monitor the sensitivity of the results to the prior information supplied. We want to use the prior information as effectively as possible, without slightly varied expert opinions leading to hugely

varied outcomes.

The analysis has two main stages, formulating a decision model, and calculating probabilities in order to make decisions. This is a site specific process for both stages, and relies on the construction of a sensible and effective prior distribution for the required parameters, and specification of suitable cost structures. Also, the statistical analysis may consist of two problems, ascertaining the level of contamination over the site as a whole; and then characterising localised areas of contamination and determining pathways. This is an important problem to be addressed both by the statistician and the site investigator. If localised areas are deemed to belong to a different population to the rest of the site, then they must be dealt with as such in the statistical analysis. It will also affect the way the sampling design is selected, as targeted sampling may be specifically required.

Clearly the DM must have the final say in this process. The statistical model cannot determine whether there is a valid S-P-R linkage present, and therefore whether elevated contaminant levels at a certain location actually pose a SPOSH. Much of the determination of contaminated land involves the judgement of one or more individuals. However, we hope that the decision tool produced here will aid the DM in coming to a conclusion, regarding contamination levels and adequacy of the sampling undertaken.

2.5 Examples and computation

Throughout the thesis we will demonstrate the methods introduced with two datasets. The first is a small hypothetical dataset on a unit grid with 3 contaminants of interest. The locations and observed values for the example are shown in Figure 2.4. We will refer to this as site H. We will use it to introduce the main ideas and procedures, and to highlight any potential problems that we should be aware of for a real site.

We also have access to a real case study, Site R, shown in Figure 2.5. We will investigate this site and discuss the background of the investigation in Chapter 4. For this site we have data available for soil samples taken at a number of locations over the anonymised site, with several contaminants observed. We will pick a selection

of these after consulting the desk study and a site expert in order to demonstrate the modelling and sample selection methodology. The samples selected will be those where information is available for all designated contaminants of interest. We will be able to use cross validation methods on this data set to determine the ability of our model to learn about "unsampled" locations.

This case study will also allow us to look at the elicitation process, namely the method by which the beliefs of the expert are collected and then quantified in order to use in the Bayesian framework.

All implementation of the methods introduced in this thesis will be performed[2] using the freely available statistical programming language R [72]. While R is a powerful tool for statistical programming and graphics, it is envisioned that in the future this methodology could be packaged into a piece of software more accessible to the end user.

[2]for the reader interested in the code used in the application of the methods in this thesis, contact the author at beccasthesis@hotmail.co.uk

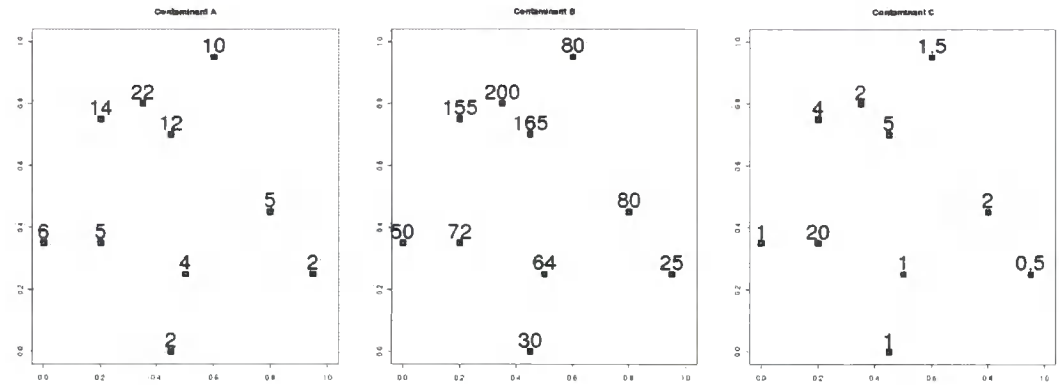


Figure 2.4: Location of hypothetical datapoints and observed values for the three contaminants. The "site" is a unit square, and 10 locations were chosen to enable investigation of a site with a cluster of locatoins, as well as some undersampled areas. We gave hypothetical values for three contaminants, two which are correlated and one which is unrelated. This will enable us to look at a variety of examples, for single and multiple contaminant investigations.

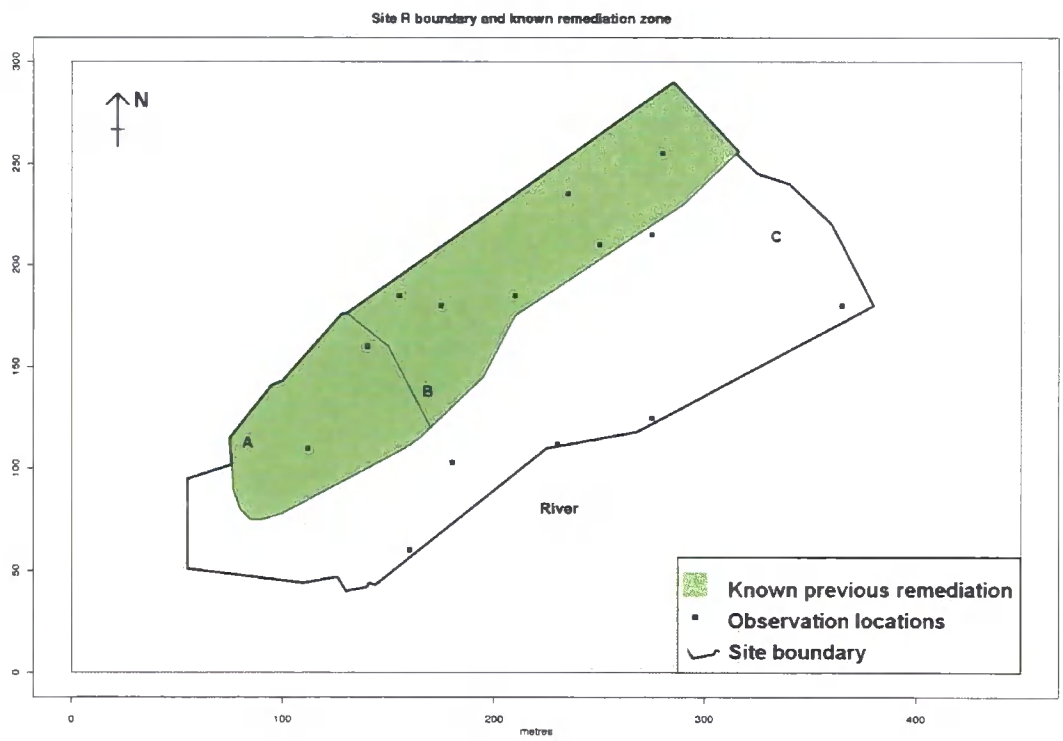


Figure 2.5: Location of observation locations and site boundary of the real site. The key features of the site are shown here, and will be further described in Chapter 4. The green area shows the region of the site where a previous remediation attempt is known to have been undertaken. A,B and C refer to areas of the site with different topographical features, which wll be necessary when considering elicitation and zoning.

Chapter 3

Modelling site contamination

Site investigation lends itself well to a Bayesian modelling approach. The compilation of a thorough desk study and availability of a wide range of expert knowledge, combined with the collection of contamination data, intuitively suggests a method whereby we select a model for the site and update our beliefs regarding model parameters as new data becomes available. We can also incorporate historical data into prior belief statements. The Bayesian method allows us to deal with uncertainty and treats model parameters as random variables. This way we can learn about them by obtaining new data. We expect that this will lead to a reduction in our uncertainty, so that we can make decisions with improved levels of confidence.

In terms of site investigation and the EA guidance, this will allow us to decide whether a site poses a risk to human health by calculating probabilities of SGV exceedance, or some other relevant criteria. We can generate predictive distributions for any required target criterion by repeatedly simulating realisations of the updated model. This tool will be used in later chapters when developing a sample selection methodology, by providing a way to compare expected performance of different sampling designs.

In this chapter, we will initially give a brief review of the Bayesian method. We shall then define the model with which we will describe and learn about contamination levels over a site. The basic model will be introduced in Section 3.3 and then expanded to include several desirable properties, in order to develop as realistic a description as possible. We need to make a compromise between computational

tractability and level of detail, while keeping in mind that the ultimate goal of modelling the site is to help design sampling schemes; and so we must be able to carry this methodology through to a decision analysis and sample search algorithm.

We require a model which is flexible enough to cover a varying range of complexity; from a single contaminant analysis with constant mean, to a multiple contaminant analysis with related polynomial trends. We would like to include as much detail from the desk study and expert's belief specification as possible, and will consider ways to do this. However, the approach is intended for routine use, so that our model choices must reflect the practical limitation of time and resource which will be available for constructing the model in any particular application.

Once the model and its extensions have been covered, we shall look at updating the model when data becomes available using standard Bayesian results, and procedures to simulate from the predictive distribution. We consider two updating approaches, as we will incorporate both in the sampling methodology. A closed form update is available when we are dealing with one contaminant only, using the corresponding conjugate prior distribution, this is computationally convenient. However, as we include multiple contaminants in our model, a closed form update is no longer available and we will introduce MCMC methods.

3.1 Bayesian inference

Before we consider the model with which we shall investigate the contamination levels on a site, we will first consider the methodology with which we update our beliefs regarding the contamination levels. Bayesian inference is the method by which statements of prior belief are made regarding a parameter of interest, and then updated using Bayes theorem in the light of new evidence (data) to obtain adjusted (posterior) beliefs. Figure 3.1 displays the form that the analysis takes, and the terms are defined as:

- Prior belief - this is a formal statement quantifying the subjective belief regarding the unknown parameter or parameters of interest. The prior probability distribution expresses the uncertainty about the unknown before any data is

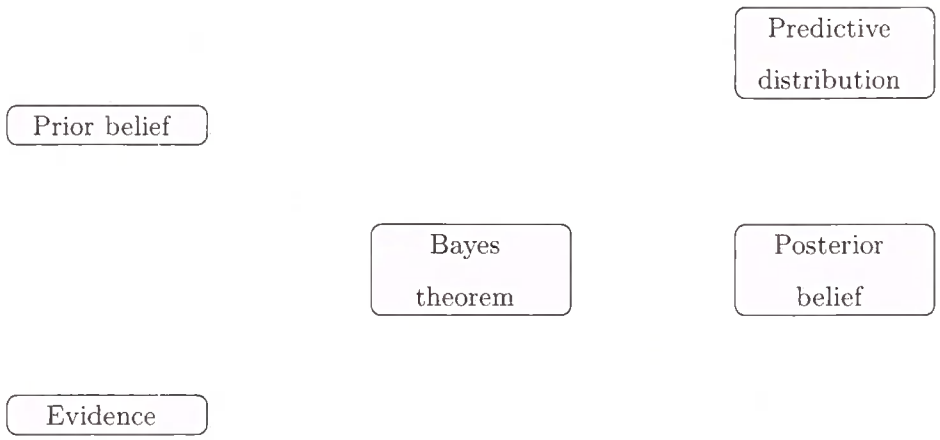


Figure 3.1: The stages of the Bayesian update.

obtained.

- Probability model, or likelihood - is the model with which we predict performance in terms of the unknown parameters of interest.
- Evidence - in order to update our beliefs we require the observation of evidence, through the collection of data.
- Posterior belief - the evidence we observe will modify our prior beliefs and give us a posterior probability distribution.
- Predictive distribution - as well as learning about the unknown parameters we wish to make inferential statements about future observations.

These objects are combined as in Figure 3.1 using Bayes theorem (shown in Appendix A.1.1). We will discuss the individual elements of this process in detail as we progress through this chapter.

3.2 Requirements of a contamination model

The basic aim of developing a model for contamination is to improve on the current statistical approach as discussed in Chapter 2, and so introducing a Bayesian spatial element is an obvious first step. Whilst we do not claim to describe the complex underlying geological and hydrogeological features of specific sites, we can

look at general features which will always be present. We can then incorporate extra structure in our model to account for site specific features as appropriate.

Clearly, we wish to exploit the fact that observations taken at locations in close proximity to each other will usually be more strongly related than those at distance. However, there are more complicated factors to consider. For example, a large regular grid may appear to cover the site, but does not enable us to learn about the relationship between locations closer together than the grid spacing. However, clustered sampling will leave portions of the site uninvestigated. Also, if the site contains any long, thin areas of contamination, we may not pick this up with a regular grid. Furthermore, the strength of association may not be the same in every direction and we need to find ways to account for this. We then have to consider how two locations are related not only at the surface, but according to their depths. We have decided to work with the top 100cm of soil in this modelling approach, as discussed in Section 1.1.4, and so make the assumption that locations at the surface will exhibit some level of correlation throughout the site, allowing this correlation to decay, as a function of distance, to zero as rapidly as required. There may also be discontinuities in the distribution of the contamination, which we should bear in mind when developing a model.

A couple of warnings are attached to this assumption. First we must be aware of any physical obstacles present on site which may interfere with the spatial correlation, such as impermeable clay lenses, or a river. Figure 3.2 gives a simple visual depiction of these issues. If we fix the correlation length and assume it is the same in every direction, then we would expect the relationship between observations at locations A and B to be the same as that of B and C or C and F etc, though clearly this is not likely to be the case in reality. To account for this fact we will allow for the correlation length to vary with direction (although at this stage only by N-S or E-W) and also we can split the site into different zones as suggested in CLR7 [21] in order to account for the river "interrupting" the correlation. To the west of the river the surface and ground water will flow down and in a south-easterly direction accounting for the direction of the river, and equivalently to the east the flow will be south-westerly and so we require different treatment of these zones.

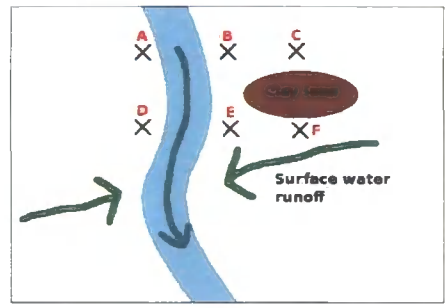


Figure 3.2: Basic features which affect spatial correlation

Secondly, while groundwater will flow at equivalent mAOD [1] levels through a site, we are assuming a relationship between locations at the top metre of the site regardless of the site topography. This may raise concern, but after discussion with geotechnical experts we are happy that the model follows the general behaviour of contaminant spread, given appropriate zoning of the site.

It is unlikely that a site has only one contaminant of concern associated with it, and so the model needs to be able to deal with multiple contaminants. We split contaminants into two broad groups in order to take advantage of between contaminant correlation. We expect that organic compounds may exhibit some level of correlation with each other, as will the inorganics. Therefore, our model shall incorporate expert judgements regarding how the presence of metal A will affect judgements about metal B etc. This enables us to use more information to update our beliefs about each individual contaminant as we can include data from all related contaminants.

3.3 The linear model

In order to use observed data for the prediction of contamination levels throughout a site, we require a spatially continuous model which will enable prediction as a function of location. The methods in this chapter expand the ideas of ordinary kriging [19,49] by introducing parameter uncertainty to reflect our limited knowledge

[1] This stands for metres Above Ordnance survey Datum and is a universal UK zero point (Env. Ag)

regarding the underlying mean and variance of the contamination surface. These results follow and build on those of Diggle and Ribeiro [33] as well as [1, 70, 71].

The Normal Linear model is proposed, with a zero mean, correlated error structure. This has been well documented in the literature as a reasonable model with which to consider spatially continuous problems [1, 33, 69]. We initially consider observing one contaminant, and one location, which is represented by $\underline{x}$, the vector containing the location coordinates. Our judgements about the level of contamination at this location, $y(\underline{x})$ are modelled as [2]

$$y(\underline{x}) = \mathbf{X}\underline{\beta} + \sigma\epsilon(\underline{x}) \tag{3.1}$$

where throughout this thesis, $\epsilon(\underline{x})$ is normal with zero mean and variance 1, and we specify a covariance function for $\epsilon(\underline{x})$ with correlation parameters $\underline{\theta}$ and κ. As we are describing one location only, $\epsilon(\underline{x})$ is a scalar, describing the variation of the observation $y(\underline{x})$ around the mean surface $\mathbf{X}\underline{\beta}$.

We introduce subscripts on the $\underline{x} = (x_e, x_n)$ coordinates to refer to the location in terms of distance east x_e and north x_n from a chosen origin. If we decide to add depth to the model, then we can include an x_d term accordingly. We label the $\underline{\beta}$ parameters in the same way, and introduce an o for the intercept.

The matrix $\mathbf{X}$ contains the coordinates of the locations $\underline{x}_1, \ldots, \underline{x}_m$. As we deal with only three possible polynomial mean specifications, $\mathbf{X}$ is one of the following

- Constant: $\mathbf{X} = \mathbf{1}$ which is $m \times 1$
- Linear: $\mathbf{X} = \begin{pmatrix} 1 & x_e[1] & x_n[1] \\ 1 & x_e[2] & x_n[2] \\ \vdots & \vdots & \vdots \\ 1 & x_e[m] & x_n[m] \end{pmatrix}$
- Quadratic: $\mathbf{X} = \begin{pmatrix} 1 & x_e[1] & x_n[1] & (x_e[1])^2 & (x_n[1])^2 & x_e[1]x_n[1] \\ 1 & x_e[2] & x_n[2] & (x_e[2])^2 & (x_n[2])^2 & x_e[2]x_n[2] \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_e[m] & x_n[m] & (x_e[m])^2 & (x_n[m])^2 & x_e[m]x_n[m] \end{pmatrix}$

[2] A list of notation introduced is included, see Page 235

If only one location is being considered (as in the case of Equation 3.1), $\mathbf{X}$ will just be a single row of one of the above three matrices.

This allows us to model observations at a point $y(\underline{x})$ as two elements which describe global and local effects respectively. $\mathbf{X}\beta$ relates to the underlying mean contaminant surface and ϵ will describe the small scale variations inherent in any heterogeneous process.

This is ann intuitive way to consider the distribution of contamination over a site, as we know contamination is likely to spread from some source(s) and so may exhibit a trend highlighting the transport of contaminant towards the groundwater flow direction or some other factor. The number of elements in the mean parameter $\underline{\beta}$ will depend on the degree of the polynomial assumed for the trend surface, and whether interaction terms are included. We allow the specification of up to a second order polynomial for the mean, as any complicated departures should be picked up by the stochastic part of the model. The other parameters of interest are σ^2 and $\underline{\theta}$, which relate to the variance and correlation respectively.

To understand the model, consider the following example, with a linear trend specification, treating $\underline{\beta}, \sigma^2, \underline{\theta}, \kappa$ as fixed, we have

$$\mathrm{E}\left[y(\underline{x})\right] = \beta_o + \beta_e x_e + \beta_n x_n \tag{3.2}$$

$$\mathrm{Var}(y(\underline{x})) = \sigma^2 \tag{3.3}$$

$$\mathrm{Corr}(y(\underline{x}), y(\underline{x}')) = \mathrm{Corr}(\epsilon(\underline{x}), \epsilon(\underline{x}')) = \exp\left(-\psi(\underline{x}, \underline{x}', \underline{\theta}, \kappa)\right) \tag{3.4}$$

$$\psi(\underline{x}, \underline{x}', \underline{\theta}, \kappa) = \left[\left(\frac{x_e - x_e'}{\theta_e}\right)^{\kappa} + \left(\frac{x_n - x_n'}{\theta_n}\right)^{\kappa}\right]^{\frac{1}{\kappa}} \quad 0 < \kappa \leq 2 \tag{3.5}$$

The $\underline{\theta}$ parameter allows us to quantify the distance at which two locations will no longer exhibit a relationship with each other. We allow $\underline{\theta}$ to consist of differing directional values if the expert determines this to be the case and so we represent $\underline{\theta}$ as a vector. The expert should be able to give an indication of a strength of relationship between locations, and we can perform a sensitivity analysis on this parameter to assess model performance.

The correlation structure describes the relationship between sampled locations and $\underline{\theta}$ and κ are the scale and shape parameters respectively. While we could include a prior on $\underline{\theta}$, for now a fixed correlation structure is used. We choose to use the

powered exponential family of covariance functions, which is the two parameter model defined in Equations (3.4) and (3.5) above. This choice need not be fixed, if on discussion with the expert it was deemed that another valid family of correlation functions would be more appropriate, then it could be used instead.

As we have introduced the correlation structure describing the relationship between observations at different locations, we shall now describe the model in terms of several locations $y(\underline{\mathbf{x}})$ rather than a single point $y(\underline{x})$

When we refer to locations for which we have already taken observations, we shall use the subscript Δ, while for prediction locations we shall use a superscript p. Given the expectation and variance structure of the model and assuming that the residuals, $\underline{\epsilon}(\underline{\mathbf{x}})$, are jointly normally distributed, the distribution of a vector of observations, $y_\Delta(\underline{\mathbf{x}}) = (y_\Delta(\underline{x}_1), \ldots, y_\Delta(\underline{x}_m))$ at locations $\underline{\mathbf{x}} = \underline{x}_1, \ldots, \underline{x}_m$, given the parameters $\underline{\beta}, \sigma^2, \underline{\theta}$ and κ is multivariate normal.

$$y_\Delta(\underline{\mathbf{x}}) \mid (\underline{\beta}, \sigma^2) \sim \mathrm{N}(\mathbf{X}\underline{\beta}, \sigma^2\mathbf{D}) \tag{3.6}$$

As $\underline{\theta}$ and κ are treated as fixed in the model, we condition on the unknown parameters $\underline{\beta}$ and σ^2 so that the pdf of $y_\Delta(\underline{\mathbf{x}})$, for given $\underline{\mathbf{x}}$, is

$$y_\Delta(\underline{\mathbf{x}}) \mid (\underline{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{m/2}|\mathbf{D}|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(y_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta})'\mathbf{D}^{-1}(y_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta})\right) \tag{3.7}$$

$\mathbf{D}$ is the correlation matrix with entries representing the correlation between y values at each pair of locations as defined in Equation (3.4). The $\mathbf{D}$ matrix must be a positive semi-definite matrix to ensure a valid covariance structure. The entry in row i and column j, $\mathbf{D}[i,j] = \exp\left\{-\left[\left(\frac{x_e[i]-x_e[j]}{\theta_e}\right)^\kappa + \left(\frac{x_n[i]-x_n[j]}{\theta_n}\right)^\kappa\right]^{\frac{1}{\kappa}}\right\}$

$\mathbf{X}\underline{\beta}$ relates to the underlying trend surface for a particular contaminant as mentioned above, $\mathbf{X}$ is an $m \times q$ coefficient matrix where q is the number of terms in the vector $\underline{\beta}$ (i.e. $q = 3$ in the setup of Equation (3.2)). Initially only the mean and variance parameters $\underline{\beta}$ and σ^2 will be considered unknown. The fixed correlation parameters which determine $\mathbf{D}$ may be specified a priori by the expert and then varied over several candidate values to observe sensitivity to this specification.

3.3.1 Choice of covariance function

The use of the powered exponential covariance function as introduced above is just one permissible function with which we could describe the spatial dependency of the site. Several other functions are defined and used in the geostatistical literature, the restriction being that the family of functions must satisfy the condition of positive definiteness. The powered exponential family was chosen as it meets both desirable criteria for modelling spatial relationships. The correlation between two locations decreases with distance, and we can control the smoothness of this relationship with a secondary parameter.

As well as the parameter $\underline{\theta}$, the specification of the powered exponential function requires the choice of a shape parameter κ. Permissible values of κ are $0 < \kappa \leq 2$. The case where $\kappa = 2$ is also known as the Gaussian correlation function, and generates very smooth Gaussian processes. Figure 3.3 shows examples of this correlation function with $\underline{\theta} = 1$ for varying κ, showing how the function decreases monotonically with increasing distance. In terms of modelling contamination, this feature is an acceptable way to quantify contamination spread and decreasing association at distance. As mentioned, we could use any of the other functions often used in the geostatistical literature such as the spherical and Matérn models, which are discussed further in [19].

3.3.2 Simulating realisations of this model

In order to learn about the effect of model parameters, we would like a simple method to simulate from this model. We will go into more detail when looking at simulating from the predictive distributions, once we have updated the model after seeing new data. However, we briefly look at simulating from the model with all parameters fixed.

For example, we have a site represented by a unit square, and have known values for the mean, variance and correlation parameter, and would like to simulate a realisation of the site. We can do this by covering the area with a unit grid and making a draw from the multivariate Normal distribution defined in Equation (3.7).

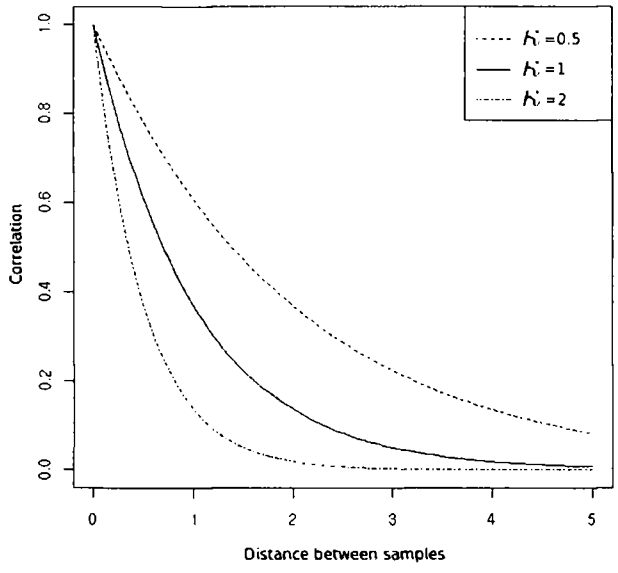


Figure 3.3: Exponential correlation structure for various κ values

As the observation consists of a deterministic and random part, we make the draw to obtain the random part, and then add on the deterministic element. There are several ways to carry out the simulation. We use the Cholesky decomposition of the variance matrix as in [33], by carrying out the following steps.

- Cover the site with a regular $m \times m$ grid of locations $\underline{\mathbf{x}}^p = \underline{x}^p_1 \ldots \underline{x}^p_{m^2}$
- Take an independent random sample of size m^2 from the standard Normal (0,1) distribution, $\mathbf{z}^p = (\mathbf{z}(\underline{x}^p_1), \ldots, \mathbf{z}(\underline{x}^p_{m^2}))$
- Compute the Cholesky decomposition of $\sigma^2\mathbf{D}$ as LL' with L lower triangular
- Now the distribution of $\underline{\epsilon}(\underline{\mathbf{x}}^p)$ is the same as the distribution of $L\mathbf{z}^p$. Adding the deterministic part gives us the simulated value $y^p = \mathbf{X}^p\underline{\beta} + \underline{\epsilon}(\underline{\mathbf{x}}^p)$ where $\mathbf{X}^p$ is the matrix containing the prediction locations.

We can then plot the values of $y(\underline{\mathbf{x}}^p)$ over the grid, where the point locations are at the centre of each square on the image. In Figure 3.4 we demonstrate this for a 10×10 grid over a unit square, and use a constant mean value of 10 (so $\underline{\beta} = \beta_o$ only). The only thing we want to change in each image is the correlation parameter $\underline{\theta}$. We use the same random draw $\mathbf{z}^p$ for each image in order to make sure this is the case.

The value of $\underline{\theta}$ is interpreted in the literature as the range, which is the distance

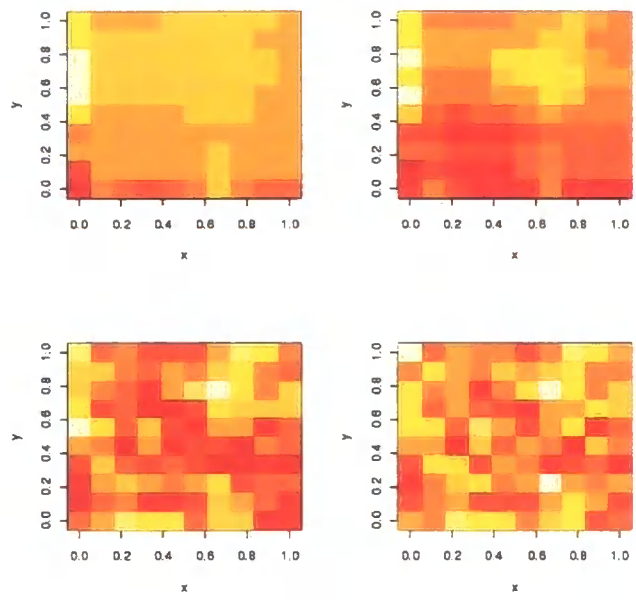


Figure 3.4: Simulation of the model for correlation lengths $\theta_e = \theta_n$, taking values 10,1,0.1 and 0.01 respectively. Red areas indicate high values, through to white for low. $\underline{\beta}$= 10, $\sigma^2 = 3$ and $\kappa = 1$

at which spatial correlation between two locations is effectively zero. This means we will assume that a relationship exists only between points closer to each other than the range. Sometimes the "practical range" is used, this being the distance at which the correlation is 0.05. In the upper left image, the value of $\theta_e = \theta_n = 10$, and so values over the entire grid exhibit a relationship. This becomes weaker as $\underline{\theta}$ becomes smaller. In the lower right hand image, we see little to no relationship between neighbouring locations, as the correlation parameter approaches zero.

3.3.3 Kriging formulation

Prediction using the model described in Section 3.3 is known as kriging in the geostatistical literature. Matheron [64] gave the method this name in honour of D.G. Krige [57]. There are several types of kriging, which we briefly outline below.

The kriging method produces an estimator for each prediction location $\underline{x}^p$ as a weighted linear combination of the sampled values. The weights are selected to give

the best linear unbiased estimator (BLUE) under mean square error minimisation [80] and an assumption of stationarity. The constraint of knowing the covariance parameters is unlikely, and several methods of variogram estimation are available. These methods are covered in detail in books including [1, 14, 19]. We are assuming the correlation parameters $\underline{\theta}$ and κ are known from expert assessment, but will vary them to assess the sensitivity of the model to these parameters.

- Simple kriging assumes a known, constant mean term β_o and the prediction values are $y(\underline{x}^p) = \beta_o + \underline{\mathbf{b}}'\mathbf{D}^{-1}(y(\underline{x}) - \beta_o\mathbf{1})$ where $\underline{\mathbf{b}}$ is the correlation between the sampled locations and the prediction location as calculated using Equations (3.4, 3.5).

- Ordinary kriging assumes a constant but unknown mean. In this case the constant mean term in the simple kriging equation is replaced by the generalised least square estimator, $\hat{\beta} = (\mathbf{1}'\mathbf{D}\mathbf{1})^{-1}\mathbf{1}'\mathbf{D}y$

- Universal kriging allows a polynomial trend for the unknown mean part of the model, and estimates this from the data using $\underline{\hat{\beta}} = (\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}y$

- "Bayesian kriging" has been introduced in the literature [19, 33] in order to account for parameter uncertainty, and we will expand the model and methods used in Section 3.5.

3.4 Incorporating structural information

While the model discussed so far enables us to capture aspects of the spatial relationship on site, we would like to be able to incorporate some other factors as well. After talking to several experts and considering how they chose to describe an example site presented to them, we found several elements which we would like to include, but cannot using the basic model. We will discuss these one by one and add them to the model introduced in the previous sections.

3.4.1 Nugget variance and measurement error

As well as the spatially correlated error structure, we may expect to see variation arising from two other sources.

- Firstly the possibility of human/equipment error means that if we were to repeat a measurement at a location several times we could get different values of contamination after laboratory analysis. This is defined as measurement error and we would like to account for this in our model.
- Secondly there may exist a microscale variation causing a discontinuity in the correlation function at locations with almost zero separation, i.e. repeated measurements at effectively the same location.
- Matheron [64] coined the term "nugget effect" to represent the fact that we cannot learn about the variation between any two locations closer together than the closest sampling points.

We take account of both these features jointly, by writing the model as

$$y(\underline{\mathbf{x}}) = \mathbf{X}\underline{\beta} + \underline{\epsilon}(\underline{\mathbf{x}}) + e(\underline{\mathbf{x}}) \tag{3.8}$$

The term $e(\underline{\mathbf{x}})$ is the sum of two independent terms, one expressing measurement error and one expressing microscale variation. $e(\underline{\mathbf{x}})$ has an associated variance parameter τ^2, and is independently Normal at each location and so $e(\underline{\mathbf{x}}) \sim \mathrm{N}(0, \tau^2 \mathrm{I}_m)$. However, while we may expect to be able to put a fixed value on the measurement error variance, τ^2_{ME}, microscale variation, τ^2_{MS} is much harder to judge. We can separate the two terms, as in Cressie [19], in order to learn about the microscale variation. For each location,

$$\mathrm{Var}(e(\underline{x})) = \tau^2_{ME} + \tau^2_{MS} = \tau^2 \tag{3.9}$$

We also introduce a parameter ν, which we call the relative nugget variance, as in Diggle [33]. In effect we will specify this value as fixed rather than τ^2 itself. This allows us to rewrite the variance as

$$\sigma^2\mathbf{D} + \tau^2\mathbf{I}_m = \sigma^2(\mathbf{D} + \nu\mathbf{I}_m) \quad (3.10)$$

$$\nu = \frac{\tau^2}{\sigma^2} \quad (3.11)$$

3.4.2 Expert uncertainty judgements

We know we will acquire detailed prior beliefs from a relevant site investigation expert (or experts), and the desk study, as discussed in Section 2.1.1. These beliefs are unlikely to allow us to form a simple polynomial mean surface with a single variance parameter. We need to make allowance for this in the model.

Figure 3.5 shows an example of an expert's belief specification. We asked for an opinion regarding the levels of Zinc on the site, and got a fairly detailed map of contaminant levels. The different colours represent high, medium and low levels, but also represent differing levels of uncertainty and correlation lengths. In the green zone, the expert expected to see little to no contamination, and as such did not expect green locations to exhibit much of a spatial relationship. The light green zone covers an area which was known to have been previously remediated, and so the expert chose to call this area very low, and was more sure about this zone. The red zones were selected as areas the expert was fairly sure would contain high levels of contamination, as the desk study mentioned previous contaminating activities in this area of the land, also signs of contamination were observed in the site walkover. It was deemed that the correlation length in these zones would also be short, as there may be several hotspots of contamination within these areas. The blue areas describe where the expert thought contamination would be moving from more contaminated areas to those less so. The correlation length here was determined to be longer than in the red or green, as contaminant transport mechanisms would be affecting the levels observed here.

When the expert is unsure about the extent of a zone, we may specify smoothing functions from zone to zone to remove the presence of a "hard" boundary. How we specify this smoothing function is up to the individual, as with the selection of the mean functions per zone. Two simple functions are linear and exponential as shown

in Figure 3.6. This does not increase the complexity of the method, as we simply have a function on location within the mean function. We will introduce these smoothing functions, and consider other important factors when we undertake the real elicitation process in Chapter 4.

Overall mean surface specification

We cannot build detailed zonal judgements into a simple prior mean specification with up to a 2nd order polynomial trend. Therefore we introduce an explicit expert mean specification, and build our uncertainty model for $\hat{y}(\underline{\mathbf{x}}) = y(\underline{\mathbf{x}}) - \mathrm{E}[y(\underline{\mathbf{x}})]$. This way we will be learning about the difference between the actual surface and the prior specified mean surface. As such, our prior mean for $\hat{y}(\underline{\mathbf{x}})$ should be zero. So now the expert is not restricted to making simple statements about an overall trend for the entire site and can give detailed descriptions of the expected contamination levels within each zone that has been specified.

Zonal variance specification

As we will discuss in the elicitation section of Chapter 4, it is not expected that the expert will be able to make very detailed statements regarding uncertainty over the whole site. They are more likely to feel comfortable describing how uncertain they are in a zone relative to some baseline value. To account for a prior variance that changes over the site we multiply the residual variance by a fixed function, which we will call $\mathbf{G}(\underline{\mathbf{x}})$. The values will be expressed in relation to the variance rather than the standard deviation, and so the model becomes

$$\hat{y}(\underline{\mathbf{x}}) = \mathbf{X}\underline{\beta} + \sigma\sqrt{\mathbf{G}(\underline{\mathbf{x}})}\underline{\epsilon}(\underline{\mathbf{x}}) + e(\underline{\mathbf{x}}) \tag{3.12}$$

We will give an example of $\mathbf{G}(\underline{\mathbf{x}})$ selection later in this chapter.

Zonal correlation structures

Rather than expressing a single correlation length for the whole site, it is more realistic to assume that the expert would like to specify a zonal correlation structure.

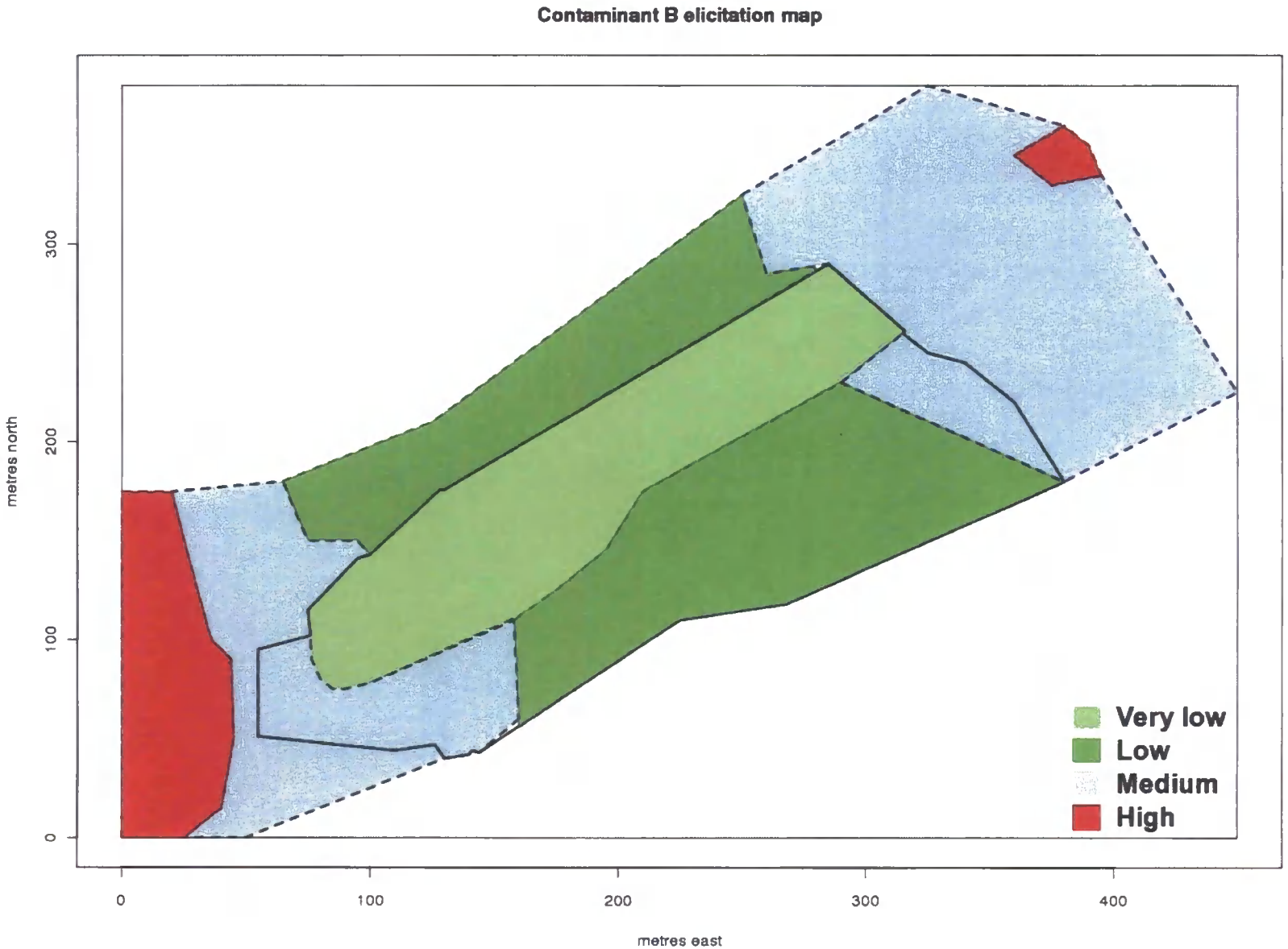


Figure 3.5: Expert elicitation image showing how we can translate the expert's zonal specifications into the computer software, in order to run the model update including zonal information.

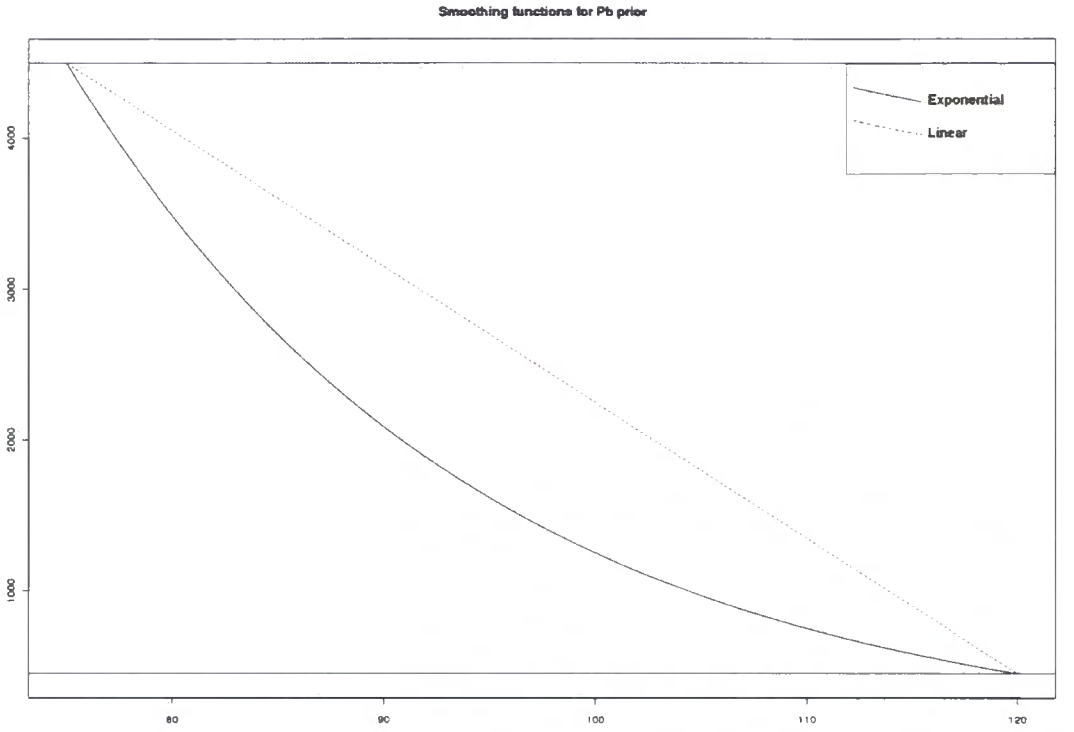


Figure 3.6: Two possible smoothing functions for transitional zones. Exponential shown in solid, allowing for the relationship to decay with distance. A linear smoothing function is shown by the dashed line, here the relationship also decays with distance, but at a constant rate.

This will allow for the inclusion of a number of sources of contamination in the model, each with a different expected level and direction of spread.

Accounting for varying correlation lengths is more challenging, as we must retain the overall positive definiteness required for a valid covariance structure. If we give each zone a suitable powered exponential correlation structure, we can build up the overall correlation as a linear combination of each zone's contribution. We introduce an individual residual variance term $\epsilon_r(\underline{\mathbf{x}})$ for each specified zone, where each $\epsilon_{r_i}(\underline{\mathbf{x}})$ is uncorrelated with $\epsilon_{r_j}(\underline{\mathbf{x}})$ when $r_i \neq r_j$ and different zones may have different associated correlation parameters. To build up the correlation structure across the site we will mix these terms to give a smooth change in correlation across zone boundaries. If we assign known weights $a_r(\underline{x})$, and we have r_l zones, the overall

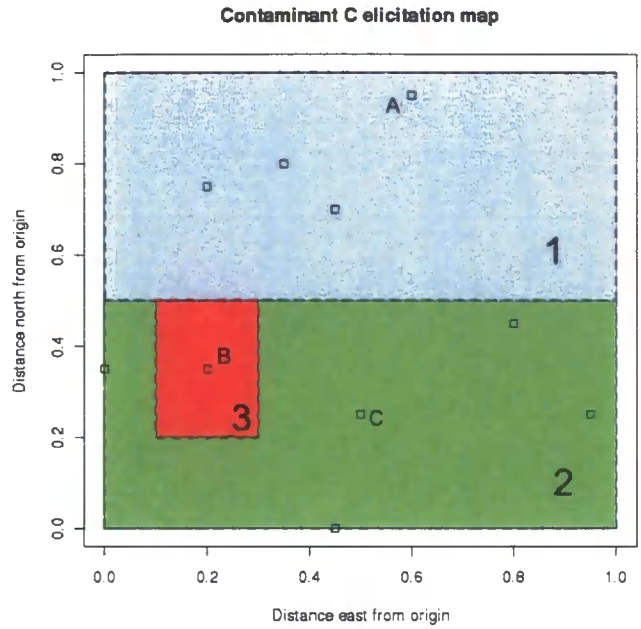


Figure 3.7: Contaminant C elicitation map for hypothetical site, showing observation locations and breakdown of zones

$\underline{x} \in$ Zone	1	2	3
$a_1(\underline{x})$	0.975	0.2	0.15
$a_2(\underline{x})$	0.2	0.975	0.15
$a_3(\underline{x})$	0.1	0.1	0.977

Table 3.1: Example of correlation weights for hypothetical contaminant C, in site of Figure 3.7, with weights rounded to 3dp

variance expression $\epsilon(\underline{x})$ is

$$\epsilon(\underline{x}) = \sum_{k=1}^{l} a_{r_k}(\underline{x})\epsilon_{r_k}(\underline{\mathbf{x}}) \tag{3.13}$$

We will constrain the specification of the weights by requiring the squared weights to sum to 1 for each $\underline{x}$; that is $\sum_{k=1}^{l} a_{r_k}(\underline{x})^2 = 1$. Table 3.1 gives an example of these weights, treated as constant over each zone for a three zone set up which is depicted in Figure 3.7. Each column will square and sum to 1 as required. By specifying two of the weights, the third weight is automatically fixed by the summing to 1 constraint. The highest weight in each case is given to the same zone. That is, if

two points lie in the same zone then most of the correlation contribution will come from that zonal specification, with a small amount from the others. We assign a slightly higher value to the weight for the zone 2 contribution if the point is in zone 1 than for the zone 3 contribution. Clearly this is a subjective judgement, but the prior specification is that zone 3 is a hotspot, and while contamination may spread, we treat the area as an isolated hotspot and so decrease its influence. The "within zone" weight does not have to be the same for every zone, as there may be several surrounding zones which require more influence in a particular case, whereas another zone may only bound one other, and so the weights would only be split between the two.

We give a constant term for the weights in this example to demonstrate the methodology. However, it may be more reasonable to think of a weight which decreases with distance from the boundary, and we have two options to deal with this. First we may add "subzones" around the zone boundaries and attach a larger constant weight here. Or, more realistically, we can use functions of location for the weights in each zone. These functions would satisfy the summing to one constraint, and we might add an extra overall term to fulfil this requirement. When considering the real site in Chapter 4 we shall look at the selection of weight functions.

With these weights specified the correlation between two locations, $\epsilon(\underline{x})$, is now calculated as

$$\text{Corr}(\epsilon(\underline{x}), \epsilon(\underline{x}')) = \sum_{k=1}^{l} a_{r_k}(\underline{x}) a_{r_k}(\underline{x}') \text{Corr}(\epsilon_{r_k}(\underline{x}), \epsilon_{r_k}(\underline{x}')) \tag{3.14}$$

So this formulation uses Equation (3.4) to calculate each $\text{Corr}(\epsilon_{r_k}(\underline{x}), \epsilon_{r_k}(\underline{x}'))$ term along with the specified weights to form the matrix $\mathbf{D}$ in Equation 3.6. We have the final single contaminant model

$$\hat{y}(\underline{x}) = \mathbf{X}\underline{\beta} + \sigma\sqrt{\mathbf{G}(\underline{x})}\epsilon(\underline{x}) + e(\underline{x}) \tag{3.15}$$

where $\epsilon(\underline{x})$ is determined by Equation (3.13). While we could add more structure to the model, we feel we have included a level of detail sufficient enough to aid site investigators in using much of the information available to them to model and learn about a specific site effectively. We add all the features introduced in Section 3.4 and our model is as defined in Equation (3.15).

3.5 Bayesian update of the model

Having introduced the model we will use to describe contamination levels, we now introduce the objects required for the Bayesian update. We will label the parameters we wish to learn about as $\underline{\eta} = (\underline{\beta}, \sigma^2)$, whilst $\underline{\theta}_k$, κ_k and τ^2 are selected by the expert, but remain fixed. As we introduce an MCMC approach, we can learn about τ^2 as well. $\mathrm{E}[y(\underline{\mathbf{x}})]$, $\mathbf{G}(\underline{x})$ and $a_k(\underline{x})$ are all fixed functions specified by the expert during the elicitation process.

Following Bayes' theorem, our posterior probability density function for the parameters of interest $\underline{\eta}$ given the observed data $\hat{y}_\Delta(\underline{\mathbf{x}})$ is

$$p(\underline{\eta} \mid \hat{y}_\Delta(\underline{\mathbf{x}})) \propto p(\hat{y}_\Delta(\underline{\mathbf{x}}) \mid \underline{\eta}) p(\underline{\eta}) \tag{3.16}$$

The Bayesian analysis of the linear model is covered in detail in several places. Here we follow the development of O'Hagan [69] and expand the theory to include the model features introduced in the previous section.

3.5.1 The likelihood in the single contaminant case

In order to calculate the likelihood, we first need to introduce a new piece of notation for the model, $\mathbf{D}_\mathbf{G}$. This combines the (combined zonal) correlation matrix $\mathbf{D}$ with the expert specified zonal variance weighting $\mathbf{G}(\underline{\mathbf{x}})$, as introduced in Section 3.4.2.

We now revise the definition of $\mathbf{D}$ to incorporate zonal specifications. We build up the $\mathbf{D}$ matrix by combining the l individual correlation matrices for each region, which are defined as:

$$\mathbf{D}_{r_k}[i,j] = \exp\left\{ -\left[\left(\frac{x_e[i] - x_e[j]}{\theta_{r_k e}} \right)^{\kappa_{r_k}} + \left(\frac{x_n[i] - x_n[j]}{\theta_{r_k n}} \right)^{\kappa_{r_k}} \right]^{\frac{1}{\kappa_{r_k}}} \right\} \tag{3.17}$$

where the correlation parameters $(\theta_{r_k}, \kappa_{r_k})$ are specified per zone. Then we can combine these zonal correlation matrices, using the zonal correlation weightings. The entry of the matrix in row i and column j, denoted $\mathbf{D}[i,j]$ is now

$$\mathbf{D}[i,j] = \sum_{k=1}^{l} a_{r_k}(\underline{x}_i) a_{r_k}(\underline{x}_j) \mathbf{D}_{r_k}[i,j] \tag{3.18}$$

Once the expert has specified the vector of relative zonal variances at the observation locations as $\mathbf{G}$ (see Section 3.4.2), we construct each element $[i, j]$ of $\mathbf{D_G}$ as

$$\mathbf{D_G}[i,j] = \mathbf{D}[i,j] \times \mathbf{G}[i]^{\frac{1}{2}} \times \mathbf{G}[j]^{\frac{1}{2}} \tag{3.19}$$

Figure 3.7 gives a simple example of this. The expert may be 2 times as uncertain in relation to some "baseline" variance in the red area, with correlation parameters $(\underline{\theta}_3 = 0.1, \kappa_3 = 2)$. In the baseline green zone they have correlation parameters $(\underline{\theta}_2 = 0.5, \kappa_2 = 2)$, and in the blue, where their uncertainty is half of the baseline, they give a correlation specification of $(\underline{\theta}_1 = 1, \kappa_1 = 2)$.

Zone	$\underline{\theta}$	κ	$\mathbf{G}$
1	1	2	0.5
2	0.5	2	1
3	0.1	2	2

Table 3.2: Summary of prior information for hypothetical site, correlation and uncertainty specification

We take observations at locations A,B and C as shown on Figure 3.7. To construct the matrix $\mathbf{D_G}$ we need $\mathbf{D}$ as defined by Equation (3.18), using weights given in Table 3.1, and $\mathbf{G}$, which in this case is (0.5,2,1) as locations A,B and C lie in these respective "uncertainty zones".

Using these values we obtain the following $\mathbf{D}$ and $\mathbf{D_G}$ (for a full calculation of the $\mathbf{D}$ see Appendix(A.2))

$$\mathbf{D} = \begin{pmatrix} 1 & 0.078 & 0.144 \\ 0.078 & 1 & 0.104 \\ 0.144 & 0.104 & 1 \end{pmatrix} \tag{3.20}$$

$$\mathbf{D_G} = \begin{pmatrix} \frac{1}{2} \times 1 & \sqrt{\frac{1}{2}} \times \sqrt{2} \times 0.078 & \sqrt{\frac{1}{2}} \times \sqrt{1} \times 0.144 \\ \sqrt{\frac{1}{2}} \times \sqrt{2} \times 0.078 & 2 \times 1 & \sqrt{2} \times \sqrt{1} \times 0.104 \\ \sqrt{\frac{1}{2}} \times \sqrt{1} \times 0.144 & \sqrt{2} \times \sqrt{1} \times 0.104 & 1 \times 1 \end{pmatrix} \tag{3.21}$$

The distribution of $\hat{y}(\underline{\mathbf{x}})$ given all the parameters is Normal:

$$p(\hat{y}_\Delta(\underline{\mathbf{x}})) \mid \underline{\beta}, \sigma^2, \tau^2, \underline{\theta}, \kappa) \sim \mathrm{N}(\mathbf{X}\underline{\beta}, \sigma^2\mathbf{D_G} + \tau^2\mathrm{I}_m) \tag{3.22}$$

3.5.2 Choosing a conjugate prior family

Having specified the form of our likelihood above, a prior distribution has to be chosen with which the beliefs of the expert(s) may be described. A conjugate prior family is defined [69] as a family of distributions which, when combined with a particular class of likelihood function, will produce a posterior distribution within the same family for any sample size and observation values. This enables computational convenience, and allows for easy comparison of prior and posterior distributions. In the case of the multivariate normal likelihood, the Normal Inverse Gamma (NIG) distribution is a conjugate family. It is possible to parametrise this by a Normal Inverse Scaled-χ^2 distribution, e.g. [33]. The joint distribution of $\underline{\beta}$ and σ^2 is

$$f(\underline{\beta}, \sigma^2) \propto (\sigma^2)^{-(d+q+2)/2} \exp\left[-\{(\underline{\beta} - \mathbf{m})'\mathbf{V}^{-1}(\underline{\beta} - \mathbf{m}) + a\}/(2\sigma^2)\right] \tag{3.23}$$

In order to represent current beliefs regarding the parameters $\underline{\beta}$ and σ^2, we require the specification of 4 hyperparameters $(a, d, \mathbf{m}, \mathbf{V})$. To understand what these hyperparameters relate to, and to learn about parameters of interest, we consider the following distributions which describe $\underline{\beta}$ and σ^2. The marginal distribution for σ^2 is Inverse Gamma (IG, with parameters $\frac{a}{2}$, $\frac{d}{2}$), the conditional distribution of $\underline{\beta} \mid \sigma^2$ is Normal (with parameters $\mathbf{m}$, $\sigma^2\mathbf{V}$) and the marginal distribution for $\underline{\beta}$ is multivariate t with parameters (d,$\mathbf{m}$,$a\mathbf{V}$) representing degrees of freedom, location and scale respectively.

$$\sigma^2 \propto (\sigma^2)^{-(d+2)/2} \exp(-a/2\sigma^2) \tag{3.24}$$

$$\underline{\beta} \mid \sigma^2 \propto (\sigma^2)^{-\frac{1}{2}} |\mathbf{V}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(\underline{\beta} - \mathbf{m})'\mathbf{V}^{-1}(\underline{\beta} - \mathbf{m})\right) \tag{3.25}$$

$$\underline{\beta} \propto \left(1 + (\underline{\beta} - \mathbf{m})'(a\mathbf{V})^{-1}(\underline{\beta} - \mathbf{m})\right)^{-(d+q)/2} \tag{3.26}$$

So, in terms of choosing values, $\mathbf{m}$ and $\mathbf{V}$ can be thought of as the prior beliefs about the trend coefficients and related uncertainty, while a and d describe the beliefs regarding the variance parameter, and uncertainty in that belief. These hyperparameters would not usually be directly elicited from an expert, but could be constructed from several values given as answers to relevant questions about expected contamination on the site (see Section 4.2). As we are subtracting the expert's contamination surface $\mathrm{E}[y(\underline{\mathbf{x}})]$, our prior mean value for $\underline{\beta}$ should be $\mathbf{m} = \underline{\mathbf{0}}$.

While this set up is convenient, it requires the specification of a particular form for the joint prior distribution for $\underline{\beta}$ and σ^2. This may not be a desirable model feature, if the expert believes that learning about σ^2 should not affect the value of $\underline{\beta}$, in which case we require independent prior distributions. We will consider this issue further in Section 3.5.5.

3.5.3 Posterior distribution and predictive distribution

Now that the form of the prior and likelihood have been specified in terms of general hyperparameters, the Bayesian update results in the calculations below. See Appendix A.1.2 for detailed calculations. The posterior distribution given $\underline{\theta}, \kappa, \nu$ is $NIG(\frac{a^*}{2}, \frac{d^*}{2}, \mathbf{m}^*, \mathbf{V}^*)$, where

$$d^* = d + m \tag{3.27}$$

$$\mathbf{m}^* = (\mathbf{V}^{-1} + \mathbf{X}'(\mathbf{D_G} + \nu \mathrm{I}_m)^{-1}\mathbf{X})^{-1}(\mathbf{V}^{-1}\mathbf{m} + \mathbf{X}'(\mathbf{D_G} + \nu \mathrm{I}_m)^{-1}\hat{y}_\Delta(\underline{\mathbf{x}})) \tag{3.28}$$

$$\mathbf{V}^* = (\mathbf{V}^{-1} + \mathbf{X}'(\mathbf{D_G} + \nu \mathrm{I}_m)^{-1}\mathbf{X})^{-1} \tag{3.29}$$

$$a^* = a + \mathbf{m}'\mathbf{V}^{-1}\mathbf{m} + \hat{y}_\Delta(\underline{\mathbf{x}})'(\mathbf{D_G} + \nu \mathrm{I}_m)^{-1}\hat{y}_\Delta(\underline{\mathbf{x}}) - (\mathbf{m}^*)'(\mathbf{V}^*)^{-1}\mathbf{m}^* \tag{3.30}$$

$\hat{y}_\Delta(\underline{\mathbf{x}})$ is the vector of observed values with the expert mean specification subtracted ($\hat{y}_\Delta(\underline{\mathbf{x}}) = y_\Delta(\underline{\mathbf{x}}) - \mathrm{E}(y(\underline{\mathbf{x}}))$), and $(\mathbf{D_G} + \nu \mathrm{I}_m)$ is as defined in Equation (3.10). The equations (3.27-3.30) are standard results, (see [69] and Appendix A.1) with the extra model features of Section 3.4.2 accounted for in the update.

While updating our beliefs about the parameters is important, we are more interested in predicting contamination levels at unsampled locations, $\underline{\mathbf{x}}^p$ as we will be using the predictive distribution to calculate the expected loss associated with the implementation of a sampling design. To compute the Bayesian predictive distribution for any number of points, we need to evaluate the following integral, with fixed $\underline{\theta}, \kappa$ and ν.

$$p(\hat{y}(\underline{\mathbf{x}}^p) \mid \hat{y}_\Delta(\underline{\mathbf{x}})) = \iint p(\hat{y}(\underline{\mathbf{x}}^p) \mid \hat{y}_\Delta(\underline{\mathbf{x}}), \underline{\eta}) p(\underline{\eta} \mid \hat{y}_\Delta(\underline{\mathbf{x}})) \mathrm{d}\underline{\eta} \tag{3.31}$$

where

$$p(\hat{y}(\underline{\mathbf{x}}^p) \mid \hat{y}_\Delta(\underline{\mathbf{x}}), \underline{\eta}, \underline{\theta}, \kappa, \nu) \sim$$
$$\mathrm{N}\left(\mathbf{X}^p\underline{\beta} + \underline{\mathbf{b}}'\left(\mathbf{D_G} + \nu\mathrm{I}_m\right)^{-1}\left(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}\right), \sigma^2(1 - \underline{\mathbf{b}}'\left(\mathbf{D_G} + \nu\mathrm{I}_m\right)^{-1}\underline{\mathbf{b}})\right) \quad (3.32)$$

where $\mathbf{X}^p$ is constructed using the coordinates of the prediction locations. The matrix $\underline{\mathbf{b}}$ contains the correlation between the prediction and sampling locations and each element of the matrix is calculated using Equations (3.4,3.5). So, for example if we have 10 observation locations and 5 prediction locations, the matrix $\underline{\mathbf{b}}$ will be 10×5 and the $[i,j]^{th}$ entry will be the correlation between the i^{th} observation location and the j^{th} prediction location.

Using standard results for conditioning and marginalising the multivariate Normal (see Appendix A.1.3 and [62]), we see that given the parameters, the distribution of $\hat{y}(\underline{\mathbf{x}}^p) \mid \hat{y}_\Delta(\underline{\mathbf{x}})$ is multivariate Normal, and the posterior distribution for $\underline{\eta}$ is Normal Inverse Gamma. Integrating out $\underline{\beta}$ and σ^2 leads to a multivariate-t distribution (see [33] for the standard results):

$$p(\hat{y}(\underline{\mathbf{x}}^p) \mid \hat{y}_\Delta(\underline{\mathbf{x}})) \sim t_{d^*}(\lambda^*, \frac{a^*}{d^*}\Lambda^*) \quad (3.33)$$
$$\mathrm{E}[\hat{y}(\underline{\mathbf{x}}^p) \mid \hat{y}_\Delta(\underline{\mathbf{x}})] = \lambda^* \quad (3.34)$$
$$\mathrm{Var}(\hat{y}(\underline{\mathbf{x}}^p) \mid \hat{y}_\Delta(\underline{\mathbf{x}})) = \frac{a^*}{d^* - 2}(1 - \underline{\mathbf{b}}'\mathbf{D}^{-1}\underline{\mathbf{b}} \quad (3.35)$$
$$+ (\mathbf{X}^p - \underline{\mathbf{b}}'\mathbf{D}^{-1}\mathbf{X})(\mathbf{V}^{-1} + \mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}(\mathbf{X}^p - \underline{\mathbf{b}}'\mathbf{D}^{-1}\mathbf{X})')$$

$$\begin{aligned}\lambda^* &= (\mathbf{X}^* - \underline{\mathbf{b}}'\left(\mathbf{D_G} + \nu\mathrm{I}_m\right)\mathbf{X})\mathbf{V}^{-1}\mathbf{V}^*\mathbf{m} \\ &+ \left(\underline{\mathbf{b}}'\left(\mathbf{D_G} + \nu\mathrm{I}_m\right)^{-1} + (\mathbf{X}^* - \underline{\mathbf{b}}'\left(\mathbf{D_G} + \nu\mathrm{I}_m\right)^{-1}\mathbf{X})\mathbf{V}^*\mathbf{X}'\left(\mathbf{D_G} + \nu\mathrm{I}_m\right)^{-1}\right)\hat{y}_\Delta(\underline{\mathbf{x}}) \\ \Lambda^* &= (\mathbf{D}^p_\mathbf{G} + \nu\mathrm{I}) - \underline{\mathbf{b}}'\left(\mathbf{D_G} + \nu\mathrm{I}_m\right)^{-1}\underline{\mathbf{b}} \\ &+ (\mathbf{X}^* - \underline{\mathbf{b}}'\left(\mathbf{D_G} + \nu\mathrm{I}_m\right)\mathbf{X})\mathbf{V}^*(\mathbf{X}^* - \underline{\mathbf{b}}'\left(\mathbf{D_G} + \nu\mathrm{I}_m\right)\mathbf{X})' \end{aligned} \quad (3.36)$$

The term $(\mathbf{D}^p_\mathbf{G} + \nu\mathrm{I})$ in Equation (3.36) is constructed in the same way as $(\mathbf{D_G} + \nu\mathrm{I}_m)$, but for the prediction locations and is the prior predictive variance matrix. The second term represents the amount we learn about by using the observed data, and we have to add on the third term to account for the parameter uncertainty. The predictive mean is a weighted combination of the prior mean and the observed values.

3.5.4 Difference between Ordinary and "Bayesian" kriging

The key difference between the two methods is the way the parameters are treated. In Ordinary Kriging (OK), the mean value is treated as unknown, but the covariance parameters (σ^2 and $\underline{\theta}$) are fixed and known. The method gives a generalised least squares estimator for $\underline{\beta}$ and uses this in the prediction equation as if it were the true value. The prediction equation gives an estimate for a new location (or vector of locations).

$$\hat{\underline{\beta}} = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1}\hat{y}_{\Delta}(\underline{\mathbf{x}}) \tag{3.37}$$

$$\hat{y}(\underline{\mathbf{x}}^p) = \hat{\underline{\beta}} + \underline{\mathbf{b}}'\mathbf{D}^{-1}(\hat{y}_{\Delta}(\underline{\mathbf{x}}) - \mathbf{X}\hat{\underline{\beta}}) \tag{3.38}$$

$$\mathrm{Var}(\hat{y}(\underline{\mathbf{x}}^p)) \mid \hat{y}_{\Delta}(\underline{\mathbf{x}})) = \sigma^2(1 - \underline{\mathbf{b}}'\mathbf{D}^{-1}\underline{\mathbf{b}}) \tag{3.39}$$

So OK ignores the uncertainty arising due to the fact we do not know the parameters $\underline{\beta}$ and σ^2, which may lead to the reported uncertainty appearing being smaller than it actually is.

We now compare this to the Bayesian method, which treats the parameters of interest as random variables which we wish to learn about. We will call this method Bayesian Kriging (BK). We specify a joint prior distribution for $\underline{\beta}$ and σ^2 as discussed, with a fixed $\underline{\theta}$ and κ value, and consider the equivalent formulae to equations for OK. We use the standard updating equations here, so we can make a more direct comparison to the results of OK, but the differences intuitively translate to the "full" model we have developed.

$$\hat{\underline{\beta}} = \mathbf{m}^* = (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}(\mathbf{V}^{-1}\mathbf{m} + \mathbf{X}'\mathbf{D}^{-1}\hat{y}_{\Delta}(\underline{\mathbf{x}}))$$

$$\begin{aligned}\mathrm{E}(\hat{y}(\underline{\mathbf{x}}^p)) &= (\mathbf{X}^p - \underline{\mathbf{b}}'\mathbf{D}^{-1}\mathbf{X})(\mathbf{V}^{-1} + \mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{V}^{-1}\mathbf{m} \\ &+ (\underline{\mathbf{b}}'\mathbf{D}^{-1} + (\mathbf{X}^* - \underline{\mathbf{b}}'\mathbf{D}^{-1}\mathbf{X})(\mathbf{V}^{-1} + \mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1})\hat{y}_{\Delta}(\underline{\mathbf{x}})\end{aligned}$$

We introduced the predictive variance for a location (or vector of locations) in Equation (3.35). We see that the two sets of equations feature similarities, but that the Bayesian equations involve more terms. This reflects the fact we are accounting for extra uncertainties. In both cases the prediction is a compromise between the

mean parameter and the data. In the case of OK the mean parameter is estimated from the data and in BK it is included as a subjectively chosen prior value updated by the observation of the data. Comparing the variance equations shows that the BK includes an extra term, again to account for the uncertainty related to the unknown parameters, which are treated as known in the OK system.

3.5.5 Conditional conjugate approach and MCMC

An alternative approach to prior specification and analysis is to exploit the benefits of conditional conjugacy. In this set-up we only need to be able to write down the distribution of each parameter conditional on every other parameter. Conditional conjugacy requires that the conditional posterior distribution of parameter α belongs to the same family as the prior. If the expert does not wish to be constrained to a model which says that $\underline{\beta}$ and σ^2 are dependent in the prior, then this approach allows for independence.

In the single contaminant case, we have two unknown parameters, $\underline{\beta}$ and σ^2 and we require $\underline{\beta} \mid \sigma^2$ and $\sigma^2 \mid \underline{\beta}$ to be of the required form. If we follow a similar approach to that of the closed form in Section 3.5.2, but now allow $\underline{\beta}$ and σ^2 to be independent then we have a conditional conjugate set up. $\underline{\beta}$ is Normal, and σ^2 is Inverse Gamma; learning about one parameter a priori no longer affects beliefs about the other. We could use this approach for the single contaminant case if the expert wishes. However, when we no longer have a closed form update we have to deal with computational issues.

In the case of multiple contaminants, we will no longer be able to perform a closed form update, and so the conditional conjugate method is a simple solution, as it facilitates the use of Markov Chain Monte Carlo methods.

Markov Chain Monte Carlo methods

When the form of our posterior distribution involves intractable integrals, MCMC (Markov Chain Monte Carlo) methods allow samples to be drawn directly from the normalised density. The class of algorithms makes use of the fact that we can generate Markov chains with ease if we can simulate from the initial distribution and

each conditional. The chain created will be a dependent sample from the target distribution. However, the memoryless property ensures that each sample drawn is only dependent on the previous sample. We may take a subset of the chain (known as thinning) in order to reduce this dependency and use it to make inferential statements. We will use Gibbs Sampling, and also the Metropolis-Hastings algorithm, of which the Gibbs Sampler is a special case.

Gibbs Sampler

Gibbs sampling is an effective and relatively simple MCMC method. If we are able to sample from each of the conditional posterior distributions then Gibbs Sampling draws from each successive distribution iteratively. The methodology is outlined below:

- Define the parameter vector $\phi = (\phi_1, \ldots, \phi_l)$ as containing the l subvectors containing the unknown parameters
- The Gibbs sampler begins with initial parameter values.
- The k^{th} iteration will draw a value of each ϕ_j^k conditional on all the other $(\phi_1^k, \ldots, \phi_{j-1}^k, \phi_{j+1}^{k-1}, \ldots, \phi_l^{k-1})$.
- Each ϕ_j is updated using the latest (k^{th}) values for the previous parameters, and the $(k-1)^{th}$ values for the parameters not yet drawn in this iteration.

Gelman et al. [48] gives a trivial Gibbs sampler to demonstrate the procedure. For details into the theory behind the use of the Gibbs Sampler see [8]. The distribution of ϕ^k depends only on $\phi^{(k-1)}$. This memoryless property is called the Markov property, the reason the chain of samples is called a Markov Chain. After an (often large) number of iterations, the Markov Chain produced should converge to its stationary distribution, which can be shown to be the posterior distribution of the parameters given the sample, and then the draws are made from the required posterior distribution. A large number of runs are required, as a number will be thrown away as "burn-in" at the beginning and the chain will be thinned out so only every 10^{th} (for example) value is stored.

Metropolis-Hastings

The Metropolis-Hastings algorithm allows for sampling from non-standard distributions. If we no longer obtain the required form of the distribution of a parameter conditional on all the others (as we will see for the τ^2 distribution), then we can no longer use Gibbs sampling alone, we require a Metropolis-Hastings step within the Gibbs sampler. An introduction to the Metropolis-Hastings method is given in [9]. Instead of generating from the conditional distribution at each stage, we perform a Metropolis random walk step each time. The algorithm involves the steps outlined below, but also a tuning parameter. The tuning parameter, Ξ, is the variance of the distribution we use to select the random walk step. This figure requires "tuning" until an appropriate acceptance rate is obtained in step 4. The acceptance rate is the proportion of time that the proposed value is accepted. If the tuning parameter is too small then the acceptance rate will be too high and the chain will converge slowly to the equilibrium distibrution. However, is the tuning parameter is too large then the acceptance rate will be very low and the sampler will also take a long time to reach equilibrium.

1 Generate a random walk step ξ, using a prespecified tuning parameter Ξ, e.g. $\xi \sim \mathrm{N}(0, \Xi)$

2 Propose a new ϕ_j value: $\phi_j^* = \phi_j + \xi$

3 Compute $p^* = \frac{p(\phi_j^* | \phi_{\setminus j})}{p(\phi_j | \phi_{\setminus j})}$

4 Generate $u \sim \mathrm{U}(0, 1)$

   - If $u \leq p^*$ then the new value of $\phi_j = \phi_j^*$
   - else the new value of $\phi_j = \phi_j$

3.5.6 Conditional distributions

There are two cases to take account of when determining the MCMC algorithm we apply. If we are not learning about τ^2, then we may apply a straight Gibbs

Sampler. However, if we would like to learn about the relative nugget variance, we add a Metropolis Hastings step within the Gibbs Sampler as introduced above.

We have the same likelihood as in the closed form update, as shown in Equation (3.22), but we now have an independent prior specification for $\underline{\beta}$ and σ^2. If we would like to learn about τ^2, then we must include a prior distribution on this. Recall that we introduced the parameter ν in Equation (3.11), the relative nugget variance which is the parameter we will specify a prior distribution for. If we have a known value for τ^2_{ME}, as discussed in Section 3.4.1, then we account for this by learning about the overall τ^2 and removing this afterwards to look at the microscale variation contribution.

We have more freedom in the choice of prior distribution for ν. Options used in the literature include Uniform and Inverse Gamma. We choose to use an Inverse Gamma, as this enables informative prior selection. So, combining this new prior with the likelihood yields the following conditional distributions (the calculations are detailed in Appendix A.3.1)

$$p(\underline{\beta} \mid \sigma^2, \nu, \hat{y}_\Delta(\underline{\mathbf{x}})) \sim \mathrm{N}(\mathbf{m}^*, \mathbf{V}^*) \tag{3.40}$$

$$p(\sigma^2 \mid \underline{\beta}, \nu, \hat{y}_\Delta(\underline{\mathbf{x}})) \sim \mathrm{IG}\left(\frac{a + (\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta})'(\mathbf{D_G} + \nu \mathrm{I}_m)^{-1}(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta})}{2}, \frac{d+m}{2}\right) \tag{3.41}$$

$$\begin{aligned} p(\nu \mid \underline{\beta}, \sigma^2, \hat{y}_\Delta(\underline{\mathbf{x}})) &\propto \nu^{-(d_\nu+1)} \exp\left(\frac{-a_\nu}{\nu}\right) |\mathbf{D_G} + \nu \mathrm{I}_m|^{-\frac{1}{2}} \\ &\times \exp\left(-\frac{1}{2}(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta})'(\sigma^2)^{-1}(\mathbf{D_G} + \nu \mathrm{I}_m)^{-1}(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta})\right) \end{aligned} \tag{3.42}$$

To perform the update we select a suitable set of parameters for the MCMC (including the number of runs, amount of burn in, thinning and tuning values for the Metropolis step if required) and let the computer do the work. Convergence diagnostics for the MCMC algorithm will be considered in Section 3.6.2. We can select which output we require depending on whether our interest is in the updated parameters, or the predictive values for particular locations. Bearing in mind that we aim to use the Bayesian modelling approach to select sampling designs, we may come into difficulty if a large number of MCMC runs are required.

3.6 Examples

In this section, examples are used to demonstrate how this model works in practice, by looking at updated parameters as well as point prediction. The examples show how we can make improved uncertainty statements about contamination over a site if we adopt a spatial approach and introduce parameter uncertainty. We also consider sensitivity analysis and model robustness to examine how "wrong" the expert specifications can be; and compare the closed update to the conditional conjugate approach.

3.6.1 Demonstrating the update, hypothetical site

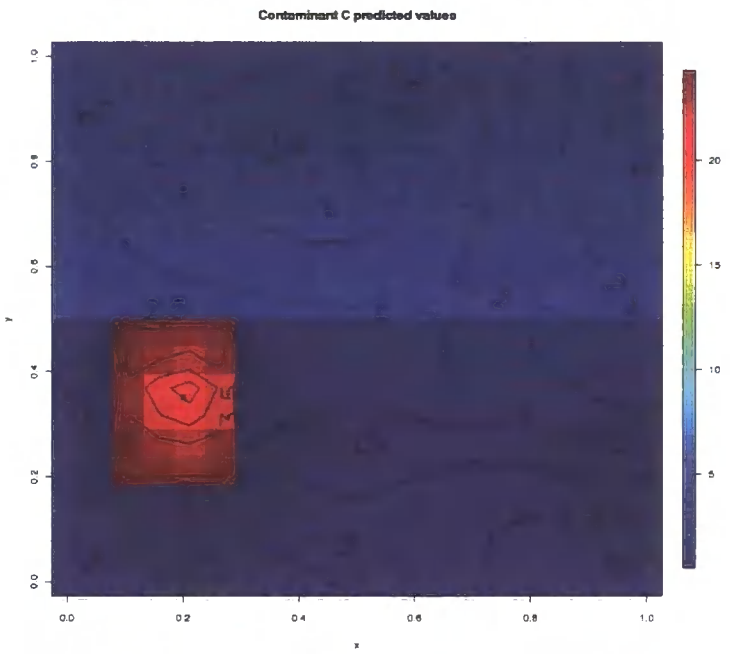


Figure 3.8: Example of prediction over grid for hypothetical contaminant C data, with standard deviation contours overlaid and observation locations displayed

We briefly consider prediction for points on our hypothetical site introduced in Chapter 2. For this example we look at the closed update only, and compare the results when we have a zonal specification, or no zoning over the site. Figure 3.8 demonstrates prediction over a 20×20 grid where we have used a detailed prior

specification shown in Table 3.3 and Equation (3.43). For the hypothetical example the prior specifications were not elicited from an expert, but chosen as reasonable values in order to demonstrate the methodology. $m_c(\underline{x})$ gives the expert defined prior expected contamination function over the site. Table 3.4 gives the locations and values for contaminant C, and they are depicted in Figure 3.9 (left hand image).

a	d	**m**	**V**	E($y(\underline{\mathbf{x}})$)	**G**	ν	Cor pars	Cor wgts
20	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{pmatrix}$	$m_c(\underline{x})$	Sec 3.5.1	0.05	Sec 3.5.1	Tab 3.1

Table 3.3: Prior specification for Figure 3.8 example

Observation	1	2	3	4	5	6	7	8	9	10
x_e	0.50	0.35	0.45	0.95	0.20	0.45	0.80	0.60	0.20	0.00
x_n	0.25	0.80	0.00	0.25	0.35	0.70	0.45	0.95	0.75	0.35
Value	1.0	2.0	1.0	0.5	20.0	5.0	2.0	1.5	4.0	1.0

Table 3.4: Hypothetical contaminant C values and location coordinates

$$m_c(\underline{x}) = \begin{cases} 8 - 5 \times x_n & \text{if } (x_n > 0.5) \\ 25 & \text{if } x_e < 0.3,\ x_e > 0.1,\ x_n \leq 0.5,\ x_n > 0.2 \\ 1 + x_e & \text{otherwise} \end{cases} \tag{3.43}$$

The shading refers to the mean level expected at each point on the grid, and the overlaid contours give the standard deviation.

The clear feature of this image is the hard boundary between predictions in the north and south zones of the site. This demonstrates the need for smoother zonal boundaries, which will be considered when analysing the real data set. Intuitively the variance is reduced around the observation locations, and is higher where there is no data, or where the expert selected a higher relative uncertainty.

This example has been chosen to show how a good specification helps reduce uncertainty and aid decisions. We will consider how sensitive the model is to "bad"

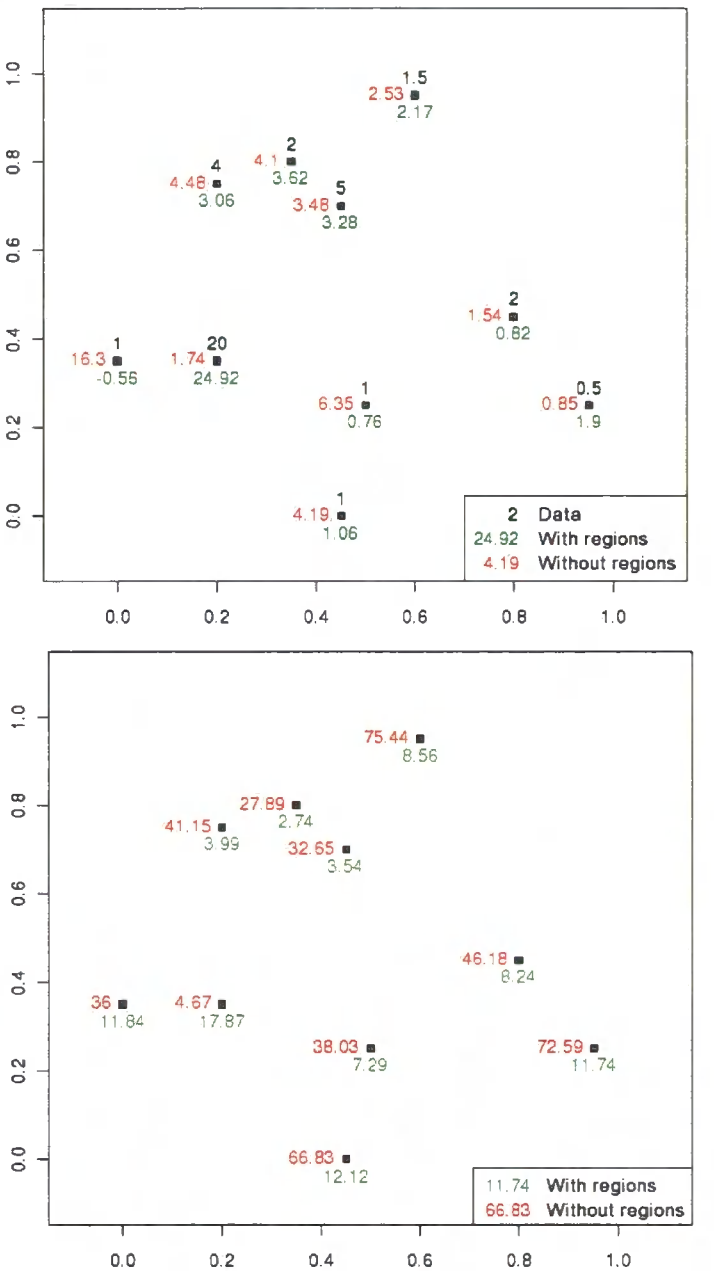


Figure 3.9: Comparison of "leave one out" prediction means (left image) and variances (right image) for the original method and the "full" model

expert specifications in detail in due course. In order to assess how the model is performing, we can carry out a cross validation exercise on the observations. Figure 3.9 shows the large differences, both in the predictive mean and variance, when we compare the "full" model with all the additional features to the basic

model. The green values are calculated using the detailed prior specification, and the red values come from the basic model set up.

Cross validation using the basic model does not predict a value near 20 for the location within the expert specified "high" area, as the remaining values are low, showing how a hotspot may be missed if not suspected and quantified in the prior elicitation. A value of $\theta_e = \theta_n = 0.5$ was specified in the basic setup and so a particularly small variance of 4.67 is attached to this location, due to all the "influential" points around it having a value of 1. In reality, our expert may not have selected this location as a high area in their prior specification either, if the contamination was arising from some unknown/undetected source, and this highlights the problem of missed hotspots. We can never resolve all the uncertainty in the site, but we can consider that any small hotspot we pick up may be an anomaly, and it will be the decision of the investigator and LPA to make case by case decisions on the importance of extreme values.

3.6.2 Comparison between the closed update and the conditional conjugate approach

For a single contaminant we have the choice of which prior specification to adopt, the closed form or the conditional conjugate. We now compare prediction and parameter estimation for the two methods as this allows us to see how varying prior values will affect the Gibbs Sampler. We will also consider MCMC diagnostics by varying the amount of burn in and thinning. In order to compare the two approaches, we must bear in mind that the two options are essentially different modelling approaches, and therefore cannot be directly compared. Also, we need to "scale" the prior information when we move from the closed approach to the conditional conjugate, as the two models treat the parameters differently in terms of independence. While we can use the same zonal specifications and expert mean/uncertainty functions for both the closed and MCMC update, we require a little extra thought regarding the selection of a and $\mathbf{V}$. This is because of the dependence between $\underline{\beta}$ and σ^2 in the closed model which is removed in the initial set-up for the conditional conjugate approach. Essentially the key difference is that $\mathrm{p}(\underline{\beta} \mid \sigma^2) \sim \mathrm{N}(\mathbf{m}, \sigma^2\mathbf{V})$ in the closed

case, where $E(\sigma^2) = \frac{a}{d-2}$ and $p(\underline{\beta}) \sim N(\mathbf{m}, \mathbf{V})$ in the alternative approach.

Figure 3.10 shows the differences between the predictive means and standard deviations for the two approaches with the same priors, over a herringbone grid. We see similar predictive means, but slightly higher predictive standard deviations to the north of the site in the conditional conjugate case. In this case we used dataset A2 as shown in Table 3.5, and the expert zone specification as in the bottom right of Figure 3.13. The regional specification here is different to that of contaminant C, and only includes two zones. Given that the closed form approach is computationally much quicker, and does not appear to give a marked change to the MCMC approach we will use this when possible. The added benefit of using the MCMC approach is the ability to learn about the parameter ν.

Dataset	Loc 1	2	3	4	5	6	7	8	9	10
A	4	22	2	2	5	12	5	10	14	6
A2	4	16	2	2	5	12	5	10	14	6
Afits	5	11.6	2	5	6.2	10.4	7.4	12.1	11	6.2

Table 3.5: Values used to demonstrate poor prior specification. Dataset A are the initial hypothetical values, with a hotspot at location 2. A2 has a lower value for location 2, and Afits give the data which fit the prior mean specification exactly

a	d	$\mathbf{m}$	$\mathbf{V}$	$E(y(\underline{x}))$	$\mathbf{G}$	ν	Cor pars	Cor wgts
20	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$	$m_a(\underline{x})$	$sd_a(\underline{x})$	0.05	Tab 3.7	Tab 3.7

Table 3.6: Prior specification for hypothetical contaminant A.

Zone	$\underline{\theta}$	κ	$\underline{x} \in$ Zone	1	2
1	1	2	$a_1(\underline{x})$	0.8	0.6
2	0.5	2	$a_2(\underline{x})$	0.6	0.8

Table 3.7: Prior information for hypothetical contaminant A

$$m_a(\underline{x}) = \begin{cases} 2 + 12 \times x_n & \text{if } (x_n < 0.85) \\ 14 - 2x_n & \text{otherwise} \end{cases} \tag{3.44}$$

$$sd_a(\underline{x}) = sd_b(\underline{x}) = \begin{cases} 1 & \text{if } (0.1 < x_e < 0.6, 0.6 < x_n < 1) \\ 1.5 & \text{otherwise} \end{cases} \tag{3.45}$$

$$\tag{3.46}$$

Learning about ν

We can choose whether to fix the relative nugget variance, or learn about it by including the Metropolis step within the Gibbs Sampler as introduced in Section 3.5.5. We implement this within R, as with the Gibbs Sampling. To demonstrate the way in which this may affect the posterior parameters, we look at two examples where we change only this factor. The first two lines of Table 3.8 relate to updating with

	Posterior ν		Posterior σ^2	
ν prior	Mean	SD	Mean	SD
Fixed 0	NA	NA	7.727	3.776
IG(2,0.05)	0.0437	0.043	7.964	3.906
Fixed 0	NA	NA	30.941	14.611
IG(2,0.05)	0.149	0.306	22.53	15.680

Table 3.8: Effect of enabling learning about the ν parameter in the MCMC. Top two rows are the posterior values when we use dataset A2 and bottom two use dataset A as in Table 3.5

the A2 dataset, and the bottom two are when a large value is included in the data, that is dataset A. By large, we mean that it falls outside the range of expected values as determined by the prior σ^2 specification, in terms of the predicted mean plus or minus two standard deviations. If an extreme value occurs in the data, the posterior variance parameter can increase to reflect this, see Section 3.6.3 for further information about specification of "sensible" priors. Also the posterior ν increases

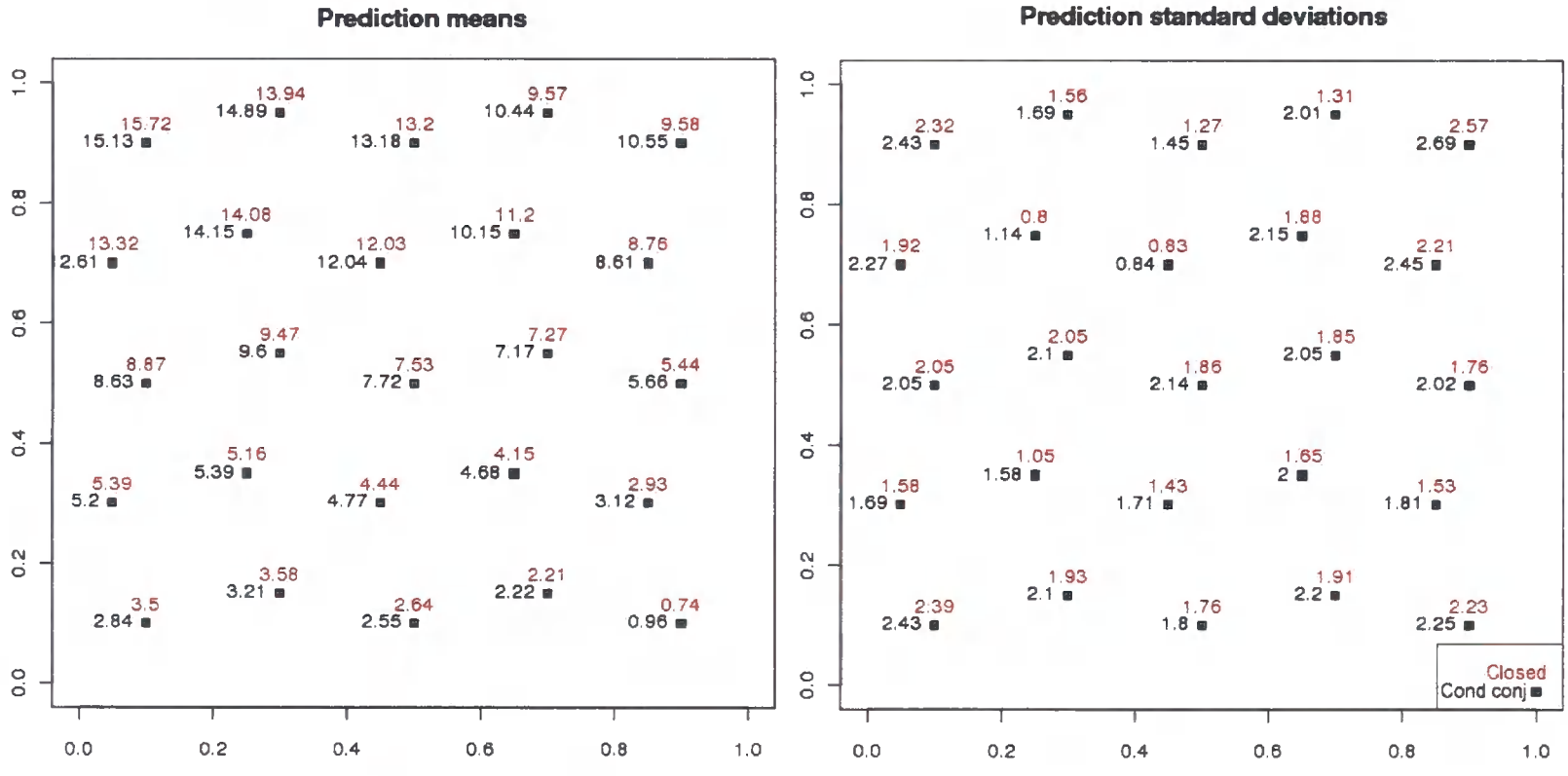


Figure 3.10: Comparison of the closed and MCMC update for prediction over a herringbone grid, given hypothetical dataset A2. Left figure shows the predictive means, closed update values in red and MCMC in black. Right figure shows standard deviations for the predicted values

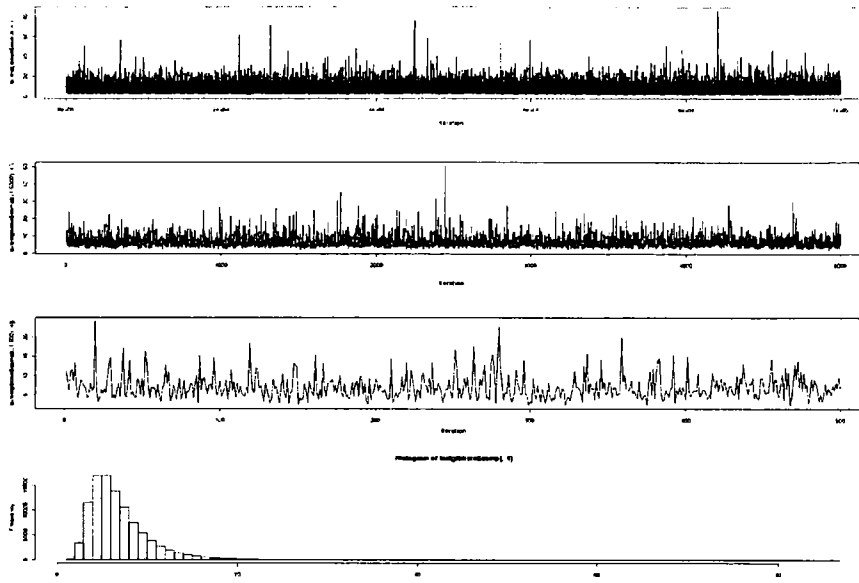


Figure 3.11: Visualisation of the posterior σ^2 simulations, from top to bottom: Full chain, first 5000 iterations, first 500 iterations, histogram of full chain

as the values become more spread out, to account for the relative nugget error and the heterogeneity of the land.

MCMC diagnostics

There are several ways in which we can assess the performance of the Gibbs Sampler, in order to check for convergence and stability. By varying the starting point of the chain we can see whether it approaches the true value. By removing a number of the initial iterations (defined in the literature as the burn-in period before the chain converges), we can remove the effect the starting parameter has on the variance of the chain.

Thinning the chain means that we can remove some of the dependence which is present in the chain. We select an amount to thin by, and then select every k^{th} value along the chain. The dependence caused by the use of a Markov Chain will always be present, and we see here in Table 3.9 that there is not a marked change in the posterior σ^2 mean and standard deviation.

By running the chain for a long time we can look at the output for evidence of convergence. We use the prior specification of Section 3.6.1. By looking at both

Thin, Take every	Mean	SD
1	6.868	3.438
5	6.918	3.591
10	6.947	3.638
15	6.914	3.455
20	6.957	3.652

Table 3.9: Mean and standard deviation of the posterior σ^2 distribution, for different amounts of thinning from the 100000 length chain

Chain Length	Mean	SD
1000	7.005	3.195
5000	6.909	3.428
25000	6.863	3.372
50000	6.866	3.420
100000	6.868	3.438

Table 3.10: Mean and standard deviation of the posterior σ^2 distribution, for different chain lengths

Table 3.10 and Figure 3.11 we see that the chain appears to settle fairly quickly. While we do see a couple of spikes, the histogram shows that these do not affect the distribution as the tail is long but there are very few points in this high range. Table 3.11 demonstrates that the amount of burn-in required is negligible. In fact there is barely a difference in the posterior σ^2 whether we remove 0 or 10000 iterations. In practice there will always be a burn-in period of at least 1000 in case the starting parameter is further from the posterior value than is reasonable in terms of the prior. However, Figure 3.12 shows that even with a large starting value, the chain settles down on quickly, as we would hope.

Burn-in	Mean	SD
0	6.868	3.438
100	6.868	3.437
1000	6.867	3.440
10000	6.868	3.441

Table 3.11: Mean and standard deviation of the posterior σ^2 distribution, for different burn-in values from the 100000 length chain

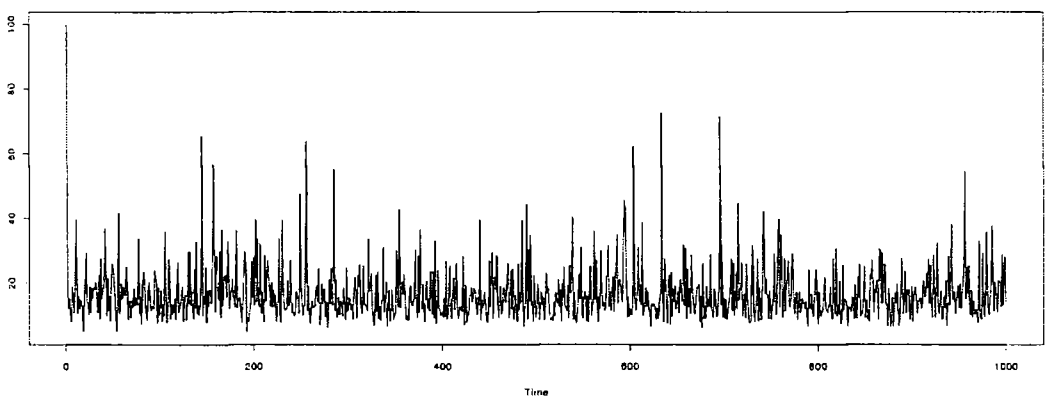


Figure 3.12: Demonstration that starting the chain from a large value can lead to large jumps in the sampler

3.6.3 Effect of prior specification and fixed correlation parameters

We now look in more detail at how the update works and how we can judge the suitability of the prior specification. As we have decided to fix the correlation parameters (although we may vary them by zone), we look at how changing them affects the update. If the observed values are far away from the prior expectation, we are likely to see an increase in the σ^2 parameter, which should alert us to the possible unsuitability of the prior specification. For example, consider the three datasets in Table 3.5. If we give a prior specification with a variance of 20 for example, we would begin to question the specification if several observations fall outside a range of $\pm$ 8.944 around the prior expected value. In this case, the third row of the table shows the expected prior values as determined by the expert's prior specification of $\mathrm{E}[y(\underline{\mathbf{x}})]$, and so the value of 22 included in the A dataset is 10.4 above the expected value. This leads to the high predictions and high predictive variance as depicted in the upper left hand image of Figure 3.13. The upper right image shows the prediction for points over a 20×20 grid when we use the closed update with a linear trend and the data A2. We see that the predictive variance is reduced as a reflection of the fact that the observed values fall within the expected prior range. As an extension of this, the bottom left image shows the predicted mean and variance over the grid when the data observed matches the prior specification exactly. The bottom right image depicts the zones specified by the expert as having different correlation structures. The uncertainty at the zone border is depicted by the steep contours in all of the images.

To demonstrate the effect of changing the correlation parameter we will use the basic model. As we hope to use expert's opinion to build up a zonal correlation structure this is not necessarily of huge importance, but as we still have a within zone fixed correlation we can highlight the problems associated with a bad $\underline{\theta}$ and κ selection. We see in Table 3.12 that the posterior variance and posterior predictive variance do change markedly with a varying $\underline{\theta}$. The mean changes to a lesser extent, as each predictive mean is a weighted combination of the other data points and the

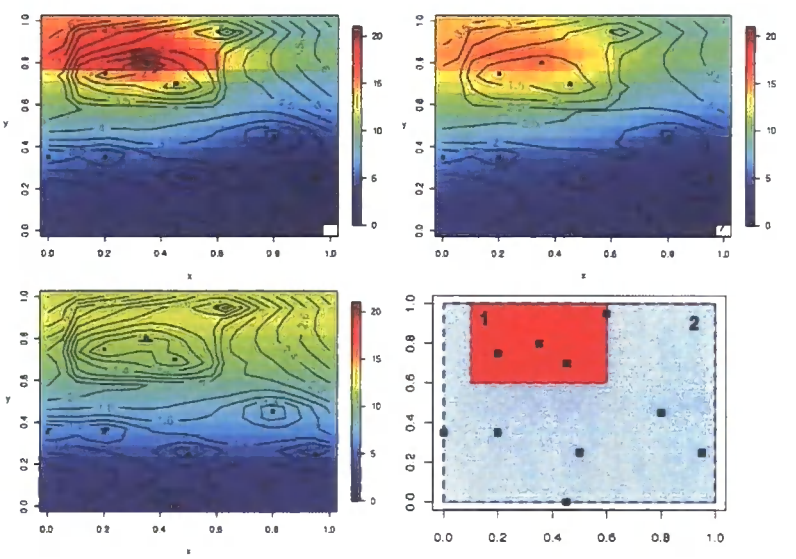


Figure 3.13: Demonstrating how one "bad" data point affects predictions. The plots depict the posterior predicted contamination levels, with overlaid standard deviation contours. Top left uses dataset A from Table 3.5, top right uses A2, bottom left Afits. The bottom right image gives the observation locations and the experts zone specification for contaminant A

prior mean. Here the only thing changing is the influence that other points have on the predicted value. As the correlation length decreases, so does the weighting to certain data points. The predictive variance is larger when we increase the correlation length. This is due to an increase in the strength of the relationship between different areas of the site, which means that more locations are being used to calculate the predictive mean. As such, when more values are used, there will be a larger predictive variance when the observations are further away from the expected values. A correlation length of 10 gives an increase in the variance as this is considerably longer than the size of the site (1×1). While it is an unreasonable value, it reminds us to select correlation parameters with care, and to carry out diagnostic assessment of our model, given the observed data. We will check the suitability of the priors selected by the expert for the real case study in the next chapter.

$\underline{\theta}$	κ	Post σ^2	Pred mean Loc 1	Pred var Loc 1
1	2	8.456301	2.954448	4.943619
0.1	2	4.105711	3.29603	8.137232
0.01	2	4.239489	3.423487	8.75354
10	2	22.07736	3.172359	3.586264
1	1	7.113746	3.215735	4.70642
1	0.01	5.123874	4.860765	10.4471
c(1, 0.5)	2	4.12314	3.963545	6.69606

Table 3.12: Effect of varying the correlation parameters

3.6.4 Comparison with the CLR7 Mean Value Test

One way we can demonstrate the potential of implementing a Bayesian procedure is to compare results with the CLR7 Mean Value Test. We introduced the MVT in Section 2.2 and talked about the issues of treating samples as independent. If we place prior beliefs on the site and update with the available data to obtain the posterior distribution for β and σ^2, we can consider a detailed description of the predicted contamination levels and related uncertainty. The MVT reports an upper

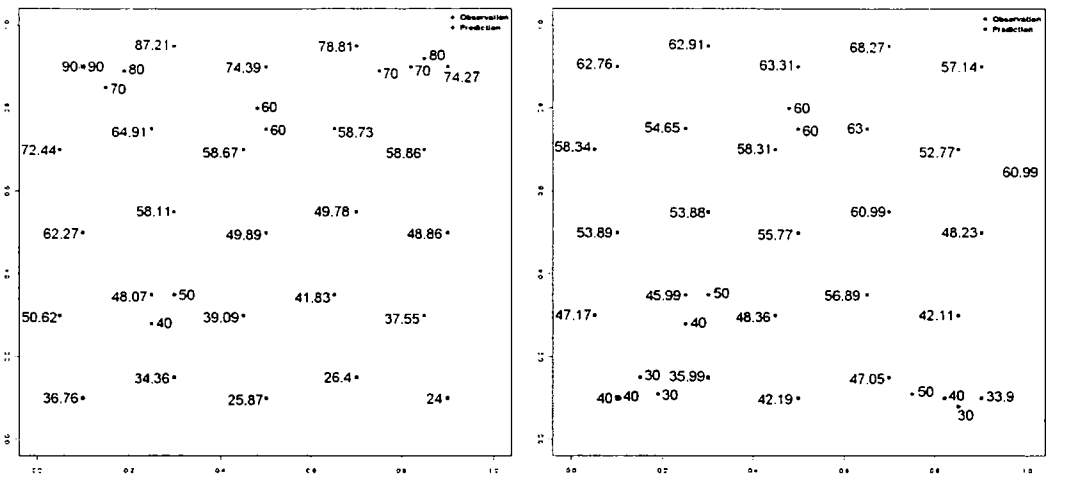


Figure 3.14: Prediction with linear trend on extreme example of Section 2.2 to show potential of including a spatial element. Left figure shows predictions for the high values, and the right image used the low values.

bound level for the mean contamination over the site. While this is seen as a conservative value, the MVT does not report the uncertainty associated with different areas of the site. When the spatial structure of the contamination is disregarded, the test may not report the true level of uncertainty, and this is why a detailed decision analysis is required.

Case	n	Value
Low	10	49.72
Low	3	55.27
High	10	75.66
High	3	82.81

Table 3.13: Effect of n on the results of the MVT

As a simple demonstration we use the extreme example of Section 2.2 to demonstrate the point clearly. While the sampling scheme used in this example would hopefully not get past the local authorities as a valid description of the site, it helps show the use of a spatial approach. We use the basic closed update here with a linear trend, and predict over a herringbone grid. The observation and prediction locations and values are given in Figure 3.14. The figure shows the values predicted

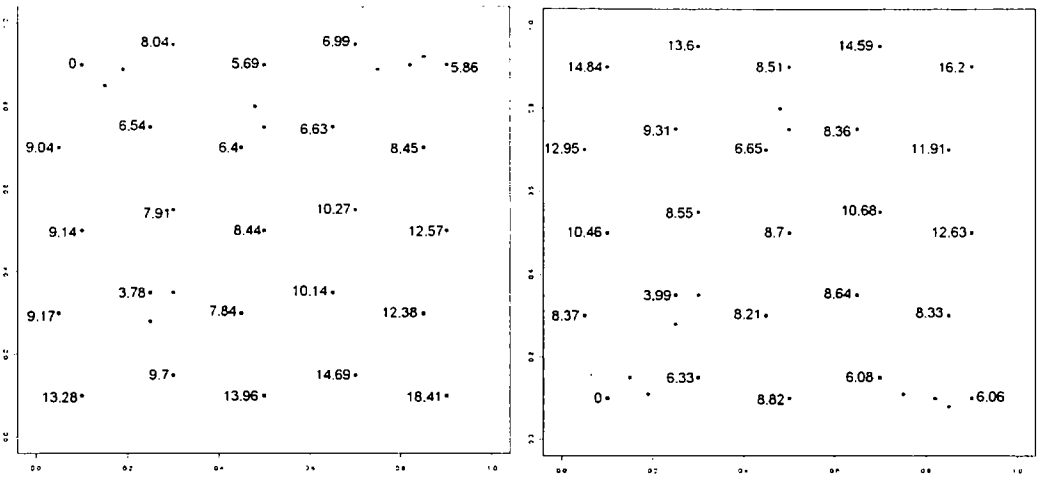


Figure 3.15: Prediction standard deviations for extreme example of Section 2.2 to show potential of including uncertainty in the analysis. The left figure shows the standard deviations when the high values were used to predict, and the right figure uses the low values. We see higher uncertainty in the zones where there were fewer observations.

over the grid, for the two extreme cases, while Figure 3.15 shows the related standard deviations. This output shows that allowing the incorporation of a linear trend immediately shows how different the predictions become. While this appears to back up the MVT result, if we take the average of the two sets of predictions, we get 52.55 for the "low" observations, and 54.07 for the "high". If we look at the results of the MVT which use the observations to calculate an upper bound for the mean of the site, we got 49.72 and 75.66 for low and high observations respectively. As a rough comparison, using the mean and standard deviation of the predicted values the equivalent upper bound values would be 67.63 and 84.73, showing the impact of allowing for the trend. We can also look directly at the uncertainty associated with each prediction location to consider where we are most unsure. This will help a DM consider where it is most important to sample in order to resolve uncertainty regarding a significant possibility of significant harm. We will use this description of uncertainty when developing a sample search algorithm in later chapters.

If we alter the correlation parameter in the Bayesian model with a constant mean, as $\underline{\theta}$ decreases the values would become closer to the equivalent MVT result

for n=8. This is because a low $\underline{\theta}$ relates to a very short correlation distance, and so samples are treated as nearly independent as in the MVT (where they are treated as independent). Equivalently, applying the mean value test with a lower n gives higher values, as we are including fewer locations, as shown in Table 3.13. This example is useful to demonstrate the shortcomings of the MVT, showing the need for a spatial modelling approach to replace it.

3.7 Multiple contaminants

We have already discussed the importance of being able to include multiple contaminants in our model. We need to consider the best way to do this. One way to model a relationship between contaminants is to introduce a "between contaminant" correlation structure, acting on the $\underline{\beta}$ parameters. We will explain this in further detail when we look at the form of the prior distributions. We also need to look at ways to elicit beliefs regarding between contaminant correlation. This adds a level of complexity to the views the expert will have to express. We require a simple and practical elicitation approach to ensure we collect as much prior information as possible.

3.7.1 Modelling the relationship between contaminants

We will look at a three contaminant case here, as this may be extended to any number of contaminants as required. We bear in mind that if two contaminants are deemed to have no relationship (such as a metal and an organic), we may analyse them independently, and use the closed update if we do not wish to learn about the nugget variance.

We now have a set of observations for each contaminant, and an individual expert specification for each contaminant, so we have $\hat{y}_{\Delta 1}(\underline{\mathbf{x}}), \hat{y}_{\Delta 2}(\underline{\mathbf{x}})$ and $\hat{y}_{\Delta 3}(\underline{\mathbf{x}})$. We assume that we have no missing data. Each contaminant has a multivariate likelihood of the same form as Equation (3.22), i.e.

$$p(\hat{y}_{\Delta i}(\underline{\mathbf{x}}) \mid \underline{\beta}_i, \sigma_i^2, \tau_i^2, \underline{\theta}_i, \kappa_i) \sim \mathrm{N}(\mathbf{X}\underline{\beta}_i, \sigma_i^2 \mathbf{D}_{\mathbf{G}_i} + \tau_i^2 \mathrm{I}_m) \tag{3.47}$$

To account for the between contaminant correlation, we are going to assume that the $\underline{\beta}$'s are correlated but the σ^2's and ν's are not. We make the following specification

$$\begin{aligned} p(\underline{\beta}_1, \underline{\beta}_2, \underline{\beta}_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, \nu_1, \nu_2, \nu_3) &\sim p(\underline{\beta}_1, \underline{\beta}_2, \underline{\beta}_3) p(\sigma_1^2) p(\sigma_2^2) p(\sigma_3^2) p(\nu_1) p(\nu_2) p(\nu_3) \\ &\sim \mathrm{N}\left(\begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \mathbf{m}_3 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31}, & \Sigma_{32} & \Sigma_{33} \end{pmatrix} \right) \\ &\times \mathrm{IG}\left(\frac{a_1}{2}, \frac{d_1}{2}\right) \mathrm{IG}\left(\frac{a_2}{2}, \frac{d_2}{2}\right) \mathrm{IG}\left(\frac{a_3}{2}, \frac{d_3}{2}\right) \\ &\times \mathrm{IG}\left(a_{\nu 1}, d_{\nu 1}\right) \mathrm{IG}\left(a_{\nu 2}, d_{\nu 2}\right) \mathrm{IG}\left(a_{\nu 3}, d_{\nu 3}\right) \quad (3.48) \end{aligned}$$

where Σ_{ij} $(i \neq j)$ will express the between contaminant correlation. We still require that this specification remains positive definite, and so we build up the Σ matrix as a whole rather than contaminant by contaminant. There are several ways in which elicitation may be carried out, depending on the technical background of the expert being consulted. To ensure overall positive definiteness, a relatively simple approach is to build up the correlation structure as follows. First, we scale all contaminants onto the same "contamination scale" in order to ensure we can compare them in a coherent way. Each contaminant will be described in terms of an overall contribution and a per contaminant contribution.

For example, if we have three contaminants, then

$$\begin{aligned} \underline{\beta}_1 &= \alpha_1 \underline{\beta} + \alpha'_1 \underline{\beta}_1^* \\ \underline{\beta}_2 &= \alpha_2 \underline{\beta} + \alpha'_2 \underline{\beta}_2^* \\ \underline{\beta}_3 &= \alpha_3 \underline{\beta} + \alpha'_3 \underline{\beta}_3^* \end{aligned}$$

The α_i's determine the contribution of the overall $\underline{\beta}$, and the α'_i's determine the contribution related to the individual contaminant specifications. We assume that individual contaminant contributions $\underline{\beta}_i^*$ are independent, and so we determine the following

$$\begin{aligned} \mathrm{Cov}(\underline{\beta}_i, \underline{\beta}_j) &= \alpha_i \alpha_j \mathrm{Var}(\underline{\beta}) \\ \mathrm{Var}(\underline{\beta}_i) &= \alpha_i^2 \mathrm{Var}(\underline{\beta}) + \alpha_i'^2 \mathrm{Var}(\underline{\beta}_i^*) \end{aligned}$$

So, Σ, the prior variance matrix for the jointly distributed $\underline{\beta}$ parameters, will be made up of:

$$\begin{aligned}\Sigma_{ii} &= \text{Var}(\underline{\beta}_i) \\ \Sigma_{ij} &= \text{Cov}(\underline{\beta}_i, \underline{\beta}_j) \\ \Sigma_{ji} &= \Sigma'_{ij}\end{aligned}$$

This method involves the specification of two α's per contaminant, as well as $n+1$ variance matrices (scalars if constant mean case). This will become increasingly time consuming and complex as the number of contaminants increases.

This method is flexible and can incorporate extra factors which may be necessary with more contaminants. If we are investigating a larger number of contaminants, we might introduce a larger collection of common factors, such as one for metals and one for organics. Each contaminant can still retain different correlation structures, with differing values of $\underline{\theta}$ and κ.

Performing this elicitation in practice may become quite complicated when trying to determine *alpha* values whilst ensuring the questions asked to the expert do not become too complex. In some cases it may be as useful to work backwards to obtain the necessary values, through determining which contaminants will be completely unrelated first, and then effectively asking for an ordering of similarity. Ensuring the values elicited remain coherent may become increasingly difficult when more contaminants are added to a multiple analysis, and we may have to consider alternative approaches to elicitation in the future.

The aim of this construction is to allow as much, or as little, prior information to be introduced as the expert feels is appropriate and useful for the investigation. We will need to consider how much information can be obtained from one contaminant in terms of learning about another, and whether it is worth going through such a procedure. We will look at the pros and cons of implementing an elicitation procedure of this format in the next chapter, when we look at the real site with expert elicitation.

3.7.2 Performing the update

We will have one multivariate normal likelihood function per contaminant as in Equation (3.47). The updating procedure per contaminant is the same as that of the single contaminant case, except for one extra calculation. Before we multiply the prior with the likelihood, we use the properties of the multivariate normal distribution (See Equation (A.2)) in order to determine the distribution of $\underline{\beta}_i$ conditional on the other $\underline{\beta}_{\setminus i}$'s. The conditioning step required is included in Appendix A.3.2, along with an extension to any number of contaminants. We then condition on each parameter in turn as in Section 3.5.6 to obtain the conditional distributions. The resulting conditional posteriors are as follows, where i, j, k refer to the three contaminants.

$$
\begin{aligned}
p(\underline{\beta}_i \mid \underline{\beta}_j, \underline{\beta}_k, \sigma_i^2, \sigma_j^2, \sigma_k^2, \nu_i, \nu_j, \nu_k) &\sim \mathrm{N}(\mathbf{m}_i^*, \mathbf{V}_i^*) \\
p(\sigma_i^2 \mid \underline{\beta}_i, \underline{\beta}_j, \underline{\beta}_k, \sigma_j^2, \sigma_k^2, \nu_i, \nu_j, \nu_k) &\sim \mathrm{IG}\left(\frac{a_i^*}{2}, \frac{d_i + n}{2}\right) \\
p(\nu_i \mid \underline{\beta}_i, \underline{\beta}_j, \underline{\beta}_k, \sigma_i^2, \sigma_j^2, \sigma_k^2, \nu_j, \nu_k) &\propto (\nu_i)^{-(d_{\nu i}+1)} \exp\left(\frac{-a_{\nu i}}{\nu_i}\right) |\mathbf{D}_{\mathbf{G}_i} + \nu_i \mathrm{I}|^{-1/2} \\
&\times \exp\left(-\frac{1}{2\sigma_i^2}\left(\hat{y}_{\Delta i}(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}_i\right)' (\mathbf{D}_{\mathbf{G}_i} + \nu_i \mathrm{I})^{-1}\right. \\
&\times \left.\left(\hat{y}_{\Delta i}(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}_i\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
a_i^* &= a_i + (\hat{y}_{\Delta i}(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}_i)'(\mathbf{D}_{\mathbf{G}_i} + \nu_i \mathrm{I})^{-1}(\hat{y}_{\Delta i}(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}_i) \\
\mathbf{V}_i^* &= \left(\frac{\mathbf{X}'(\mathbf{D}_{\mathbf{G}_i} + \nu_i \mathrm{I})^{-1}\mathbf{X}}{\sigma_i^2} + \mathbf{W}_i^{-1}\right)^{-1} \\
\mathbf{m}_i^* &= \mathbf{V}_i^* \left(\frac{\mathbf{X}'(\mathbf{D}_{\mathbf{G}_i} + \nu_i \mathrm{I})^{-1}\hat{y}_{\Delta i}(\underline{\mathbf{x}})}{\sigma_i^2} + \mathbf{W}_i^{-1}\mathbf{m}_i \right. \\
&\quad + \mathbf{W}_i^{-1}\left[\begin{pmatrix} \Sigma_{ij} & \Sigma_{ik} \end{pmatrix} \begin{pmatrix} \Sigma_{jj} & \Sigma_{jk} \\ \Sigma_{kj} & \Sigma_{kk} \end{pmatrix}^{-1} \begin{pmatrix} \underline{\beta}_j \\ \underline{\beta}_k \end{pmatrix}\right] \\
&\quad \left. - \mathbf{W}_i^{-1}\left[\begin{pmatrix} \Sigma_{ij} & \Sigma_{ik} \end{pmatrix} \begin{pmatrix} \Sigma_{jj} & \Sigma_{jk} \\ \Sigma_{kj} & \Sigma_{kk} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{m}_j \\ \mathbf{m}_k \end{pmatrix}\right]\right) \\
\mathbf{W}_i &= \Sigma_{ii} - \begin{pmatrix} \Sigma_{ij} & \Sigma_{ik} \end{pmatrix} \begin{pmatrix} \Sigma_{jj} & \Sigma_{jk} \\ \Sigma_{kj} & \Sigma_{kk} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{ji} \\ \Sigma_{ki} \end{pmatrix}
\end{aligned}
$$

Computational aspects

It is important to consider how the computational effort increases with contaminants. Using the multiple approach for two contaminants takes nearly seven times as long as the single contaminant case. So, if we treat two contaminants as independent we can learn about them both three and a half times faster than we can learn about them together. The increase in time arises from the inversion of larger matrices and the computation of extra elements relating to the $\underline{\beta}$ conditioning step included. These extra elements require calculation at each iteration of the MCMC sampler, hence the marked increase. As a result we should think carefully about which contaminants to group together, to reduce computing time. If the correlation between contaminants is very low, we may prefer to analyse the contaminants separately as this is computationally more effective, especially if we use the closed update. We may also run a multiple update in order to learn about the parameters together, and then perform further updates as single runs with the new parameters.

3.7.3 Example

We now give a brief example of updating for three contaminants on our hypothetical site, comparing this to the equivalent single MCMC updates. We ran the Gibbs Sampler for two million iterations in the case of contaminant A and B with no correlation prior, to check that the sampler was behaving as expected. We retain some non-zero elements due to rounding, and because we haven't taken an infinitely large sample.

Prior specifications

For each contaminant we will use the same prior information as outlined below, and one of a number of between contaminant correlation structures. We will give the per contaminant prior specifications here, and any differing values will be mentioned on a case by case basis. The prior information is given in Tables 3.15-3.17 and Equations (3.49-3.53). Figure 3.16 shows the zonal specifications, and observation locations and values, which are also given in Table 3.14. Contaminant A and B are

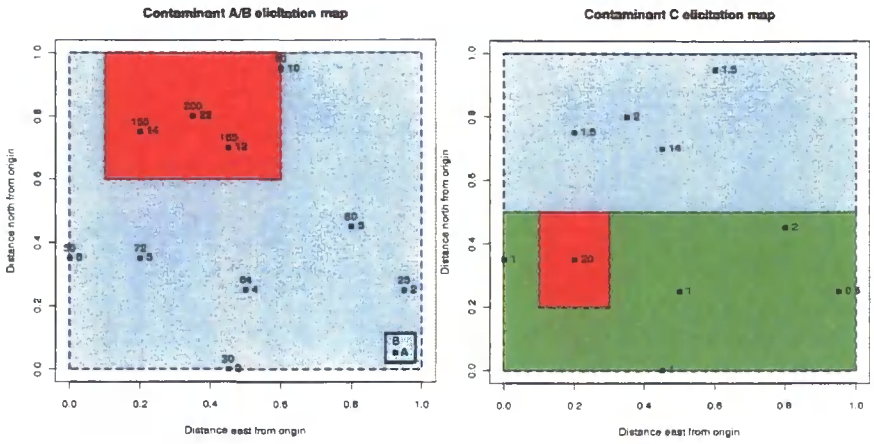


Figure 3.16: Expert zonal specification for the hypothetical example

given the same zones, and C is separate as it has no relationship with the others. For this example, and throughout the examples in this chapter, we have used a linear set up for the mean parameter of each contaminant, as this demonstrates more complexity than a constant mean. (For one of the ten cases detailed below we consider a constant mean per contaminant).

Dataset	Loc 1	2	3	4	5	6	7	8	9	10
A2	4	16	2	2	5	12	5	10	14	6
B	64	200	30	25	72	165	80	80	155	50
C	1	2	1	0.5	20	5	2	1.5	4	1

Table 3.14: Hypothetical example data

$$m_a(\underline{x}) = \begin{cases} 2 + 12 \times x_n & \text{if } (x_n < 0.85) \\ 14 - 2x_n & \text{otherwise} \end{cases} \tag{3.49}$$

$$m_b(\underline{x}) = \begin{cases} 170x_n + 20 & \text{if } (x_n < 0.85) \\ 680 - 600x_n & \text{otherwise} \end{cases} \tag{3.50}$$

$$m_c(\underline{x}) = \begin{cases} 8 - 5 \times x_n & \text{if } (x_n > 0.5) \\ 25 & \text{if } x_e < 0.3,\ x_e > 0.1,\ x_n \leq 0.5,\ x_n > 0.2 \\ 1 + x_e & \text{otherwise} \end{cases} \tag{3.51}$$

$$sd_a(\underline{x}) = sd_b(\underline{x}) = \begin{cases} 1 & \text{if } (0.1 < x_e < 0.6, 0.6 < x_n < 1) \\ 1.5 & \text{otherwise} \end{cases} \tag{3.52}$$

$$sd_c(\underline{x}) = \begin{cases} 0.5 & \text{if } (x_n > 0.5) \\ 2 & \text{if } (0.1 < x_e < 0.3, 0.2 < x_n \leq 0.5) \\ 1 & \text{otherwise} \end{cases} \tag{3.53}$$

Comparison of the updates

Tables 3.18-3.21, give the posterior $\underline{\beta}$ variance and correlation, along with the predictive mean and variance for a particular location. The location chosen was the point (0.3,0.55), and was chosen as a relatively central value for demonstration of point prediction. To consider how different factors affect the update we have changed several factors.

For some of the prediction locations we obtain a predictive standard deviation which leads to a large number of negative predictions. We can look at ways of dealing with this problem, either by truncating all negative values to zero, or possibly by adding a weighting function to the predictive variance dependent on the size of the predictive mean. However, this problem is predominantly caused by poor prior specification, and the presence of many negative predictions should alert the analyst to a potential problem. If we are predicting a small value at a location, we would expect not to see large values and as such should have a small predictive variance. This will be considered in further detail when we look at the consistency of the beliefs elicited from the expert in the next chapter.

Case 1-3 Are the single updates using no changes from the given specifications above. We use these to compare with the multiple updates. In all these cases we see a posterior σ^2 reduction and reduction in the predictive standard deviations.

Case 4 and 6 By jointly updating pairwise with no between contaminant correlation in the prior we expect to obtain the same as in the separate formulation. These updates were 200,000 runs long, and so we have some rounding error, but the updates are close to their single equivalents in cases 1-3.

Case 5,7,9 Give updates with between contaminant correlations included as described by the prior $\underline{\beta}$ variance. We again obtain similar values for the predictions and standard deviations.

Case 8 Performing the update for A and B with a constant mean and high between contaminant correlation performs well in this case due to the selection of a good prior, and hypothetical observations.

Case 10 Here we alter the prior within correlation weights from those of Table 3.17, to a $\underline{\theta}$ of 0.01 throughout the site. As a result the posterior σ^2 is reduced, and the predictive standard deviations increase slightly due to the lack of surrounding points to influence the prediction.

A point of note is that the predictive means and standard deviations don't vary significantly with type of update used (i.e. single, multiple, no correlation or with correlation present). We shall investigate this further when we consider the real case study in the next chapter. If it makes little difference to the update then we would prefer to update each contaminant separately as this is much more computationally efficient. We will have to think about how information relating to multiple contaminants will affect the selection procedure for sampling schemes. This will be considered in Chapter 5 and 6. The examples in this chapter were run predominantly to demonstrate the methodology, and raise any potential issues to consider carefully. In the next chapter we will have to deal with real data which is unlikely to be as easy to deal with as the hypothetical example.

We have now developed a general method with which to model contamination over a site, quantify and update beliefs. We consider the practical issues involved in modelling a real site in the next chapter.

Contaminant	a	d	$\mathbf{m}$	$\mathbf{V}$	$\mathrm{E}(y(\underline{\mathbf{x}}))$	$\mathbf{G}$	ν	Cor pars/reg wgts
A	20	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$	$m_a(\underline{x})$	$sd_a(x)$	0.02	See Table 3.17
B	30	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 20 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 20 \end{pmatrix}$	$m_b(\underline{x})$	$sd_b(\underline{x})$	0.02	See Table 3.17
C	10	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$	$m_c(\underline{x})$	$sd_c(\underline{x})$	0.02	See Table 3.16

Table 3.15: Priors for multiple contaminant example

Zone	$\underline{\theta}$	κ	$\underline{x} \in$ Zone	1	2	3
1	1	2	$a_1(\underline{x})$	0.975	0.2	0.15
2	0.5	2	$a_2(\underline{x})$	0.2	0.975	0.15
3	0.1	2	$a_3(\underline{x})$	0.097	0.097	0.977

Table 3.16: Prior information for multiple contaminant example. Correlation and zonal weighting functions for contaminant C

Zone	$\underline{\theta}$	κ	$\underline{x} \in$ Zone	1	2
1	1	2	$a_1(\underline{x})$	0.8	0.6
2	0.5	2	$a_2(\underline{x})$	0.6	0.8

Table 3.17: Prior information for multiple contaminant example. Correlation and zonal weighting functions for contaminant A and B

Case	Pred loc Mean	Pred loc SD
1. A	9.575	2.057
2. B	118.146	3.719
3. C	5.259	1.171
4. A,B	9.584, 118.128	2.072, 3.736
5. A,B	9.539, 118.108	2.059,3.708
6. A,C	9.587,5.255	2.074,1.176
7. A,C	9.611, 5.251	2.086,1.182
8. A,B	9.528,118.215	2.064,3.638
9. A,B	9.565,118.021	2.091,3.741
10. A,B	9.109,118.188	3.016,3.955

Table 3.18: Comparison of single and multiple MCMC approaches

Case	Post σ^2	Prior $\underline{\beta}$ variance	Post $\underline{\beta}$ var
1. A	8.215	$\begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$	$\begin{pmatrix} 2.503 & -0.662 & -0.584 \\ -0.662 & 3.138 & -0.309 \\ -0.584 & -0.309 & 2.889 \end{pmatrix}$
2. B	26.857	$\begin{pmatrix} 20 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 20 \end{pmatrix}$	$\begin{pmatrix} 10.526 & -2.999 & -3.471 \\ -2.999 & 13.237 & -0.675 \\ -3.471 & -0.675 & 12.587 \end{pmatrix}$
3. C	4.456	$\begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$	$\begin{pmatrix} 1.737 & -0.607 & -0.734 \\ -0.607 & 2.453 & -0.306 \\ -0.734 & -0.306 & 2.454 \end{pmatrix}$
4. A,B	8.185 26.578	$\begin{pmatrix} 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 20 & 0 & 0 \\ 0 & 0 & 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 0 & 0 & 20 \end{pmatrix}$	$\begin{pmatrix} 2.455 & -0.593 & -0.593 & 0.022 & -0.014 & -0.004 \\ -0.593 & 3.033 & -0.224 & 0.029 & -0.071 & -0.057 \\ -0.593 & -0.224 & 2.902 & 0.061 & 0.001 & -0.032 \\ 0.022 & 0.029 & 0.061 & 10.513 & -2.761 & -3.364 \\ -0.014 & -0.071 & 0.001 & -2.761 & 13.226 & -0.705 \\ -0.004 & -0.057 & -0.032 & -3.364 & -0.705 & 12.959 \end{pmatrix}$
5. A,B	8.322 26.713	$\begin{pmatrix} 4 & 0.8 & 0.8 & 5.367 & 1.789 & 1.789 \\ 0.8 & 4 & 0.8 & 1.789 & 5.367 & 1.789 \\ 0.8 & 0.8 & 4 & 1.789 & 1.789 & 5.367 \\ 5.367 & 1.789 & 1.789 & 20 & 4 & 4 \\ 1.789 & 5.367 & 1.789 & 4 & 20 & 4 \\ 1.789 & 1.789 & 5.367 & 4 & 4 & 20 \end{pmatrix}$	$\begin{pmatrix} 1.991 & -0.345 & -0.412 & 1.553 & -0.489 & -0.544 \\ -0.345 & 2.648 & -0.047 & -0.513 & 2.515 & 0.053 \\ -0.412 & -0.047 & 2.401 & -0.633 & 0.051 & 2.195 \\ 1.553 & -0.513 & -0.633 & 8.996 & -2.04 & -2.483 \\ -0.489 & 2.515 & 0.051 & -2.04 & 11.745 & -0.228 \\ -0.544 & 0.053 & 2.195 & -2.483 & -0.228 & 11.043 \end{pmatrix}$

Table 3.19: Comparison of single and multiple MCMC approaches

Case	Post σ^2	Prior $\underline{\beta}$ var	Post $\underline{\beta}$ var
6. A,C	8.215 4.433	$\begin{pmatrix} 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{pmatrix}$	$\begin{pmatrix} 2.438 & -0.569 & -0.596 & 0.001 & 0.003 & -0.007 \\ -0.569 & 3.049 & -0.194 & 0.007 & -0.031 & 0.046 \\ -0.596 & -0.194 & 2.904 & -0.034 & 0.004 & 0.043 \\ 0.001 & 0.007 & -0.034 & 1.708 & -0.609 & -0.776 \\ 0.003 & -0.031 & 0.004 & -0.609 & 2.355 & -0.254 \\ -0.007 & 0.046 & 0.043 & -0.776 & -0.254 & 2.444 \end{pmatrix}$
7. A,C	8.347 4.455	$\begin{pmatrix} 4 & 0.8 & 0.4 & 0.8 & 0 & 0 \\ 0.8 & 4 & 0.8 & 0 & 0.8 & 0 \\ 0.4 & 0.8 & 4 & 0 & 0 & 0.8 \\ 0.8 & 0 & 0 & 4 & 0.8 & 0.4 \\ 0 & 0.8 & 0 & 0.8 & 4 & 0.8 \\ 0 & 0 & 0.8 & 0.4 & 0.8 & 4 \end{pmatrix}$	$\begin{pmatrix} 2.228 & -0.25 & -0.524 & 0.285 & -0.171 & -0.184 \\ -0.25 & 2.771 & 0.115 & -0.178 & 0.428 & -0.062 \\ -0.524 & 0.115 & 2.778 & -0.179 & -0.093 & 0.412 \\ 0.285 & -0.178 & -0.179 & 1.588 & -0.45 & -0.694 \\ -0.171 & 0.428 & -0.093 & -0.45 & 2.171 & -0.067 \\ -0.184 & -0.062 & 0.412 & -0.694 & -0.067 & 2.289 \end{pmatrix}$
8. A,B cte	8.358 25.67	$\begin{pmatrix} 4 & 8.05 \\ 8.05 & 20 \end{pmatrix}$	$\begin{pmatrix} 1.506 & 2.422 \\ 2.422 & 6.986 \end{pmatrix}$

Table 3.20: Comparison of the MCMC approaches

Case	Post σ^2	Prior $\underline{\beta}$ var	Post $\underline{\beta}$ var
9. A,B	8.259 26.40	$\begin{pmatrix} 4 & 0 & 0 & 5.367 & 0 & 0 \\ 0 & 4 & 0 & 0 & 5.367 & 0 \\ 0 & 0 & 4 & 0 & 0 & 5.367 \\ 5.367 & 0 & 0 & 20 & 10 & 10 \\ 0 & 5.367 & 0 & 10 & 20 & 10 \\ 0 & 0 & 5.367 & 10 & 10 & 20 \end{pmatrix}$	$\begin{pmatrix} 2.257 & -0.537 & -0.5 & 2.276 & -1.412 & -1.351 \\ -0.537 & 2.785 & -0.145 & -1.216 & 3.827 & -0.759 \\ -0.5 & -0.145 & 2.534 & -1.371 & -0.805 & 3.354 \\ 2.276 & -1.216 & -1.371 & 6.739 & 0.056 & -0.174 \\ -1.412 & 3.827 & -0.805 & 0.056 & 8.876 & 1.44 \\ -1.351 & -0.759 & 3.354 & -0.174 & 1.44 & 8.03 \end{pmatrix}$
10. A,B	5.537 9.382	$\begin{pmatrix} 4 & 0 & 0 & 5.367 & 0 & 0 \\ 0 & 4 & 0 & 0 & 5.367 & 0 \\ 0 & 0 & 4 & 0 & 0 & 5.367 \\ 5.367 & 0 & 0 & 20 & 10 & 10 \\ 0 & 5.367 & 0 & 10 & 20 & 10 \\ 0 & 0 & 5.367 & 10 & 10 & 20 \end{pmatrix}$	$\begin{pmatrix} 1.113 & -0.67 & -0.737 & 1.214 & -1.118 & -1.224 \\ -0.67 & 2.077 & -0.11 & -1.073 & 2.91 & -0.476 \\ -0.737 & -0.11 & 1.854 & -1.179 & -0.491 & 2.542 \\ 1.214 & -1.073 & -1.179 & 2.595 & -1.8 & -1.973 \\ -1.118 & 2.91 & -0.491 & -1.8 & 5.314 & -0.155 \\ -1.224 & -0.476 & 2.542 & -1.973 & -0.155 & 4.698 \end{pmatrix}$

Table 3.21: Comparison of single and multiple MCMC approaches

Chapter 4

Analysis for a real site

In the previous chapter we introduced and demonstrated the methods with which we model the contamination levels over a site. We used simple, hypothetical examples with varying priors in order to show how the updates work. We will now look in more detail at the elicitation and modelling process for the real case study in order to highlight practical issues arising in the appication of the methodology. We can also see how the model deals with the fluctuations present in a real dataset. We will consider elicitation issues, how to deal with the computational issues of prediction over a large grid, dealing with large uncertainties arising in undersampled areas, and what information is important for the consideration of the sampling selection problem.

4.1 The Site

As introduced in Section 2.5, we are using an anonymised site, shown in Figure 4.1, which we refer to as Site R. We now describe the main features of the site so that we can explain any patterns which may appear in the data. The site boundary is the official boundary given in the real site investigation report. We will ask questions about this area and a wider area around the boundary in order to obtain a detailed prior specification, so that we can understand where contamination may be entering/leaving the site. Also, we may consider the wider boundary when we look to select a sampling scheme. This may not be reasonable given that there

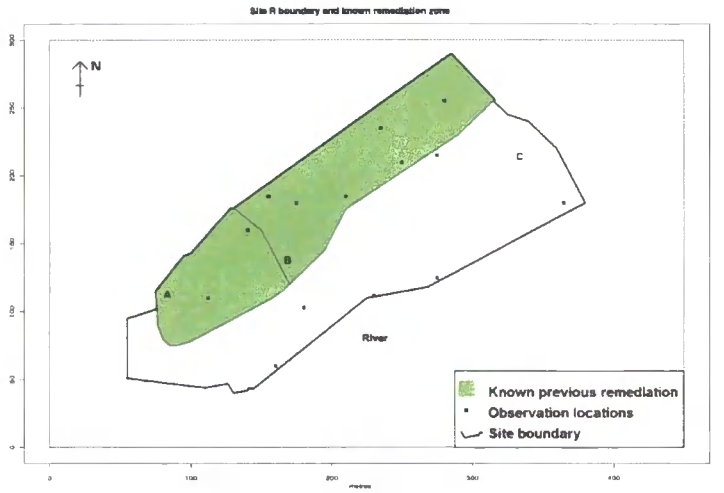


Figure 4.1: Site R site boundary and observation locations

will be large uncertainties involved when we are far away from the initially sampled locations, and this may give a sampling design entirely outside the site boundary. However, the selection of points will also depend on the loss function we use when designing sampling search algorithms. While giving information outside the site boundary is of interest and will affect the beliefs within the site boundary, the remit of the investigator is to consider points within this boundary.

4.1.1 Site description and history

The desk study for the site considered previous site investigation and remediation reports, historical site plans, details of previous remediation undertaken, and information gained from consultation with a former employee of the tar works which was on the site in the past. The key aspects of the desk study are summarised below.

Ground levels on the site slope downwards from north to south towards the river. There are two slopes, which are both very steep. Slope 1 separates two plateaus (A and B) present in the green area of Figure 4.1, and slope 2 is between the green area and the riverside path (C). Slope 2 is densely covered with vegetation and, while not impossible to reach by foot, it would be very difficult to access with any machinery (i.e. to dig a trial pit or sink a borehole). We should bear in mind that, given the slope towards the river, contamination may be spreading from northwest

to southeast but it is unlikely to be spreading in the opposite direction. The site is mainly recreational grassland, along with vegetation and woodland. The riverside area is a tarmaced path, mainly used by dog walkers and cyclists. For the purposes of the site investigation the land use has been assessed as residential with plants for the assignment of SGVs.

The site has a history of various industries as on much of the land in the UK. The most recent was a tar works, which covered much of the area in and around the site boundary, with a lead works to the left of the site boundary. Previously to this, a chemical works was situated on the site, in the southwest of the site.

The green shaded area in Figure 4.1 is known to have been infilled with uncontaminated soil approximately 9 years ago. Also, all above ground and known below ground structures have been previously removed and pump and treat wells installed. Previous investigations over the past three decades have indicated high levels of polyaromatic hydrocarbons (PAH) within the site, and Lead and Arsenic outside the site boundary where the lead works was situated.

Whilst there are no plans to develop the site, for the purposes of this assessment it is assumed that the site will be redeveloped to residential housing with gardens. Children may be playing on the land and so should be treated as the critical receptors.

4.1.2 Contaminants of interest, SGVs and sampling locations

As with all ex-industrial sites in the UK and further afield, it is not unusual to expect a large number of contaminants to be present at some level. Here we choose four contaminants of interest in order to demonstrate our methodology. Extending this to a large number of contaminants would not cause any theoretical problems, only increase the computational burden. It would potentially increase the complexity of the elicitation if we wish to learn about the relationships between contaminants.

We should consider the levels of PAH over the site as this has been highlighted as a previous contaminant of concern. Rather than picking specific contaminants, we have initially grouped together several to give a level of total PAH. However,

Figure 4.2: Collecting soil samples at the site

Contaminant	Critical Value
Arsenic	20mg/kg
Lead	450mg/kg
Zinc	648mg/kg [1]
Total PAH	36mg/kg [2]
BaP	1.04 mg/kg

Table 4.1: Critical values for use in the case study. SGVs taken from the site report where available, for residential use

after considering the site report and noting that Benzo(a)pyrene (BaP) is the main contaminant of concern we have decided to model this individually as well. After discussion with the expert, we have also chosen three metals of interest, Arsenic, Lead and Zinc in order to look at correlated contaminants in our model.

Use of critical values

Table 4.1 gives the values which we will use throughout this case study (here and in Chapter 7). As mentioned above, the values used are for the residential land use type as we would consider the critical receptor to be a 0-6 year old girl playing on the site. While these are seen as screening values by the regulatory bodies, we use them for

[1] As taken from the official site report

[2] This value is taken from the US EPA total PAH screening value. Based on 16 PAHs [82]

now as indicators of potential SPOSH. In practice, these values are generally used as indicators of a requirement for further investigation, rather than an indication that remediation is definitely required. However, to demonstrate the methodology we will use these values as a threshold for the requirement for remediation, as the model can be adapted for any choice of critical value the expert/legislators wish to specify. (Whilst the values for Arsenic and Zinc have since been revised, we will use the values which were present at the time of this case study).

Prediction grid

In order to assess the contamination, we would like to predict a value for every location on the site. Whilst we could place finer and finer grids on the site to ensure maximum coverage, this would not be computationally feasible. A more effective approach is to predict over a coarser grid, and use methods in R to create the images we will see later in this chapter. This is done using interpolation functions which smoothen between prediction locations, and give a simple visual description of the results. Another option is to display the results as in Figure 4.12. In this image we see the prediction locations only. As discussed in Section 2.1.2 the preferred choice of grid from the guidance is the herringbone design. We will not use a herringbone grid for the final sampling design, but we require a starting grid to search over, and it is sensible to use the best current grid layout as this starting grid.

We have written code to produce a herringbone grid over any site (not necessarily a regular shape). However, whilst the benefit of producing a grid within the irregular site boundary is clear, it sometimes produces unwanted results. Figure 4.3 shows in pink two areas where we have no points. The lower area is particularly important, as we discover from prediction and looking at the initial observation values, as this area potentially has high levels of contamination present, and we would like to investigate further.

To create a herringbone grid, we initially place a herringbone design over a rectangle, and then the points outside the required site boundary are discarded, as shown in Figure 4.3. The filled in locations indicate the final herringbone grid within the site boundary. The remaining circles complete the initial grid over a rectangle

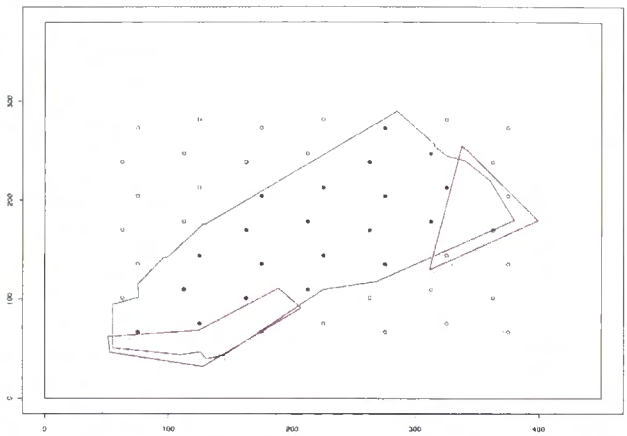


Figure 4.3: Potential problems with the herringbone grid which leaves some areas within the site R boundary unsampled, shown in pink. This is a potential problem when placing a grid over an irregular site, and we can rectify this by altering the grid size/staring point as necessary.

covering the area, these are removed as they fall outside the site boundary. In order to avoid the problem of large areas without sampling locations we can create several candidate grids until we find a suitable choice. The grid we have chosen for prediction has 100 points and covers the site well. This grid may also be used as the choice for the sampling search, or we may use a finer/coarser grid depending on computational capabilities.

4.2 Elicitation

The elicitation of a "sensible" prior specification is crucial to any Bayesian analysis. It is particularly important when we will not receive a vast amount of data with which to update the expert's beliefs [68]. During a site investigation, the compilation of the desk study collates into one place all relevant information which will aid the selection of a sampling scheme. Combining this information with a detailed discussion about the site, with one or more experts, will lead us to a formal description of the parameters of the prior distribution required for our model. Whilst it is inevitable that, even between experts, different parameter values are likely to be elicited, we would expect small deviations in the prior not to have a major influence

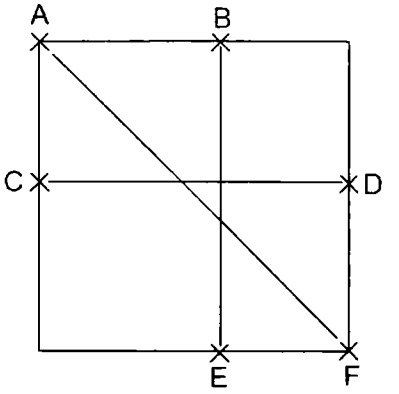


Figure 4.4: Demonstration of points used to learn about the prior **V** parameter

on the posterior distributions. We have explored this using the hypothetical dataset in the previous chapter, and will also perform diagnostic and sensitivity checks on this site, to identify which aspects of the prior specification are not important for our conclusions. We also check that the family of prior distributions we have elected to use are performing effectively, given the expert's beliefs and the initial data.

Elicitation tends to be a very case specific procedure, depending on the statistical knowledge of the expert and the site in question. There are many articles discussing the problems associated with the elicitation process, and ways to deal with these problems, including [18, 54, 68].

It would not be easy or effective to ask an expert questions directly relating to the parameters of a Normal Inverse Gamma distribution, and so we take a different approach. We require answers to a number of qualitative and quantatitive questions in order to determine a sensible prior specification. Using a series of site maps, the expert was asked a number of increasingly detailed questions regarding her beliefs about the presence of contamination. The following scheme proposes the required steps for an effective elicitation. In practice this would be guidance rather than a strict step by step scheme, and may be a several stage elicitation. An expert may become tired or bored if asked many questions at once, and undertaking two or more sessions allows the refinement and adjustment of views and values.

4.2.1 Elicitation scheme

1. Initial questions to place the expert in the correct frame of mind would be qualitative. These might include her opinion on the groundwater flow direction in order to help consider the direction a trend may occur in through the transport of contamination in water; and which contaminants she would expect to see on the site in elevated levels. This, combined with any preliminary data would help decide the contaminants we investigate.

2. The expert would be asked to use coloured pens and delineate the site into "high", "medium" and "low" zones. These definitions were discussed in terms of the SGV if available, or some other critical value which was decided upon. One map was used per contaminant, these are shown in Figures 4.5, B.5-B.7. This gives us the zonal breakdown as introduced in Section 3.4.2.

3. Once the zones have been decided upon we require a set of values for each in order to specify $\mathrm{E}[y(\underline{\mathbf{x}})]$. For this the expert could be asked to describe the average value she expected to see in each zone, and about the extremes of the zone to determine her uncertainty levels, for the choice of $\mathbf{V}$, a and $\mathbf{G}(\underline{\mathbf{x}})$. Questions asked could include "What values would you be surprised to see?", or "What are the biggest and smallest values you would expect to observe in this zone?" The expert's answers should be checked for consistency and coherence. Normal distribution theory tells us that we expect most observations to be within plus or minus two standard deviations, and so an answer of "I would expect contamination levels of no more than 2 (generic contamination units), but I am very unsure and they could be as big as 30" give us a problem. We cannot specify a mean of 2 and a standard deviation of 14 as this also says the contamination could be as low as -26. The expert would be asked to reconsider her choice of mean or extreme values. We may wish to consider alternative assumptions than normality, but initially we stick with this choice, and deal with any negative predictions.

4. For the choice of $\mathbf{V}$, it is necessary to isolate the behaviour of the contaminant in each direction if a linear or quadratic trend is assumed. To do so similar

questions to those in Stage 3 could be asked, but about points at the same east or north location. For example, Figure 4.4 shows pairs of points which could be used to determine the required values. Asking about the range of expected values at the points C and D would give an idea of the expected west-east trend, and points A and F would give information relating to the NW-SE direction.

5. In order to determine the strength of relationship between zones, we asked whether the zonal boundaries were "soft" or "hard". That is, would she expect a smooth graduation of contamination from zone to zone, or would she expect a jump from zone to zone (for example there may be a buried structure preventing the spread of contamination). Also it would be necessary to determine how much each zone may influence the neighbouring or further afield zones. This will enable the specification of the weights a_i as introduced on page 46.

6. Each zone would require a within contaminant correlation level, in order to build up a correlation structure as defined by Equation (3.13-3.14). Here we would ask questions including "How much would you expect an observation at location A in this zone to affect an observation at location B in the same zone", "At what distance from this point would you expect no relationship to exist for values of contaminant x?"

7. As well as determining the prior map for each individual contaminant, the expert would be encouraged to express a belief regarding the between contaminant correlation. This could be undertaken using the method proposed in Section 3.7.1, or any way the expert is comfortable expressing her views.

8. It is important to bear in mind that we can use either the closed update, or the conditional conjugate approach. As such the hyperparameters we select during the elicitation exercise are not suitable for both approaches, due to the dependence in one case. However, we can determine hyperparameters for each model if required. In this case we will look at both options in order to compare and assess the performance of the models.

The responses were used to produce maps such as Figure 4.5. At present these maps are created by hand, transforming the hand drawn images to a set of coordinates to enter in the software. The maps for the other contaminants are included in Appendix B. For this case study we have decided to fit one linear trend per contaminant throughout the site. While we could place a linear trend on each zone and learn specifically about the departures from the mean within each zone, we are not certain about the delineation of the zones, and so we specify zonal expectations and variances, and then learn about the overall linear trend on $\hat{y}(\underline{x}) = y(\underline{x}) - \mathrm{E}[y(\underline{\mathbf{x}})]$ for the site.

If this methodology were to be offered for routine use, then it would be useful to develop a user-friendly interface. We would envisage some sort of screen where a map could be clicked and delineated into zones on screen and beliefs input as a series of answers to questions from the software (such as that of the US EPAs Visual Sampling Plan software [86]). At present I have used the responses of the expert, combined with her annotated map to produce images in R such as Figure 4.6. Computationally, we can select a zone of any shape rather than simple regular areas as we saw in the hypothetical example, by determining the expected value of a point dependent on its presence within one of the specified zones. As we discussed in Section 3.6.1 there can be a steep change at zonal boundaries which is unlikely to accurately describe the distribution of contamination over the site. To deal with this we can place a smoothing zone onto the site, or have functions within the original zones to reflect the changing expected mean function. In the case of Lead for example, we decided to place an extra smoothing zone onto the prior description, as depicted in Figure B.1. We decided on this rather than a decreasing function within the blue zone as the expert was fairly uncertain as to the extent of the spread. The correlation lengths chosen within and across these zones will allow for a further description of the uncertainty expressed by the expert regarding the relationship between locations.

We will look first at the priors for each contaminant individually, and update and cross validate before going on to considering multiple updates. The update for Lead and some additional figures are included in Appendix B. Again it is important

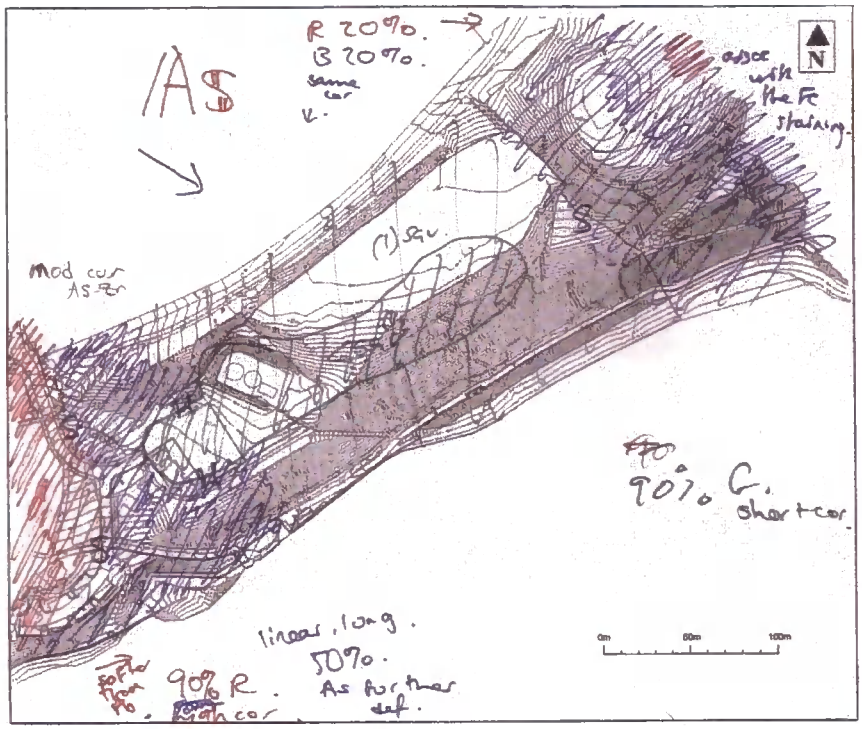


Figure 4.5: Demonstration of the method used to elicit beliefs from the expert, in this case for the levels of Arsenic contamination

Figure 4.6: Expert elicitation map for Arsenic, including observation locations

throughout these examples to remember that we make a scaling on the priors when switching from the closed form update to the MCMC approach in order to ensure a level of comparibility between the methods.

4.2.2 Arsenic contamination

Prior specification

The SGV for Arsenic is 20mg/kg. The expert split the site into 7 zones to represent her beliefs regarding the levels of contamination as shown in Figure 4.6. The beliefs for each contaminant were based upon the previous site uses and historical maps of the site. While the regional specification covers the whole map, our prediction will be confined to locations within the site boundary (as shown by the solid line in Figure 4.6). If data was obtained then the site boundary could be extended as necessary. After discussion of the definitions of very low, low, medium and high we gave the prior specification as listed in Table 4.2 and 4.3. The mean function given is in the form of a constant regional mean, and the linear trend we learn about will tell us about the departures from this specification. Along with the extra information we wish to elicit, the overall hyperparameters for the Normal Inverse Gamma distribution had to be selected. The expert gave an indication of how uncertain she was about each of these values, which we turned into zonal standard deviation specifications. It was also important to bear in mind that the choice of prior should not allow for negative predicted values. As the model uses a Normal distribution it is possible that if we predict a small mean with a large associated variance we will obtain negative predictions when making draws from the distribution. We clearly cannot expect to see negative contamination levels. We will truncate these to zero when they occur in the simulations, or adjust our prior specification accordingly. The presence of the large variance/small mean combination should act as a diagnostic warning, and lead to further questions being raised. If this occurs frequently, a lognormal distribution may be preferred as a way to remove the problem of mnegative predictions. However, this raises further questions about the shape of the distribution required.

When specifying the zonal weightings as in Section 3.4.2 (and Table 3.1), we decided that zones would only be influenced by those adjacent to them. Most of the weighting will be assigned to the zone itself, i.e. the values on the diagonal of the 7×7 array will be the largest. The correlation lengths given reflect the expert's opinion,

	a	d	$\mathbf{m}$	$\mathbf{V}$
Closed prior	400	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{pmatrix}$
Closed posterior	3789.26	17	$\begin{pmatrix} 47.81 \\ -0.03 \\ -0.19 \end{pmatrix}$	$\begin{pmatrix} 0.600 & -0.001 & -0.002 \\ -0.001 & 0.000 & -0.000 \\ 0.002 & -0.000 & 0.000 \end{pmatrix}$
Cond conj prior	400	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 2000 & 0 & 0 \\ 0 & 2000 & 0 \\ 0 & 0 & 2000 \end{pmatrix}$
Cond conj posterior	4112.59	17	$\begin{pmatrix} 49.63 \\ -0.03 \\ -0.19 \end{pmatrix}$	$\begin{pmatrix} 164.105 & -0.334 & -0.458 \\ -0.334 & 0.003 & -0.002 \\ -0.458 & -0.002 & 0.005 \end{pmatrix}$

Table 4.2: Prior and posterior hyperparameters in the closed and conditional conjugate case for Arsenic

in that the green and red areas were deemed to have very short correlation lengths, and the blue zones would be longer to reflect the spread of contamination from the high to low zones. As in Section 3.6.2, we need to be careful when implementing the prior information for the two approaches, by remembering $\mathrm{p}(\underline{\beta} \mid \sigma^2) \sim \mathrm{N}(\mathbf{m}, \sigma^2\mathbf{V})$ in the closed case, where $\mathrm{E}(\sigma^2) = \frac{a}{d-2}$ and $\mathrm{p}(\underline{\beta}) \sim \mathrm{N}(\mathbf{m}, \mathbf{V})$ in the alternative approach.

The prior specification for Arsenic was the simplest. It was decided that no smoothing zones were needed. Bearing in mind that the expert gave a prior $\mathrm{E}(y(\underline{\mathbf{x}}))$, we give a prior $\mathbf{m}$ of $\underline{0}$ and then the a and $\mathbf{V}$ prior values describe how far away from the prior expectation we could reasonably expect to observe contamination levels.

Updating and diagnostics

As we are performing a single update, we may use the closed update as this is computationally quicker. However, we will also perform the single MCMC update, to compare the two models. As mentioned above, the two are not directly comparable as they are different models, and require the specification of a modified set of prior

Zone	$E(y(\underline{x}))$	$\mathbf{G}$	$\underline{\theta}$	κ	$\underline{x} \in$	1	2	3	4	5	6	7
1	200	4	2	5	$a_1(\underline{x})$	0.950	0.230	0	0.100	0.100	0	0
2	60	1	2	100	$a_2(\underline{x})$	0.229	0.95	0.094	0.150	0.150	0	0
3	15	0.5	2	5	$a_3(\underline{x})$	0	0.104	0.95	0.150	0.150	0.104	0
4	20	0.5	2	15	$a_4(\underline{x})$	0.150	0.130	0.200	0.950	0.100	0.130	0.150
5	20	0.5	2	15	$a_5(\underline{x})$	0.150	0.130	0.200	0.100	0.950	0.130	0.150
6	60	1	2	100	$a_6(\underline{x})$	0	0	0.094	0.150	0.150	0.950	0.229
7	200	4	2	5	$a_7(\underline{x})$	0	0	0	0.100	0.100	0.230	0.950

Table 4.3: Prior specification for Arsenic contamination at Site R

parameters.

Figure 4.7 shows the predicted values in both cases, along with predictive standard deviations, again bearing in mind that this is prediction within the site boundary. The predictions were made over a herringbone grid of 100 points and then interpolated to produce a smooth map over the site. The upper left image in Figure 4.7 shows the prediction surface with overlaid standard deviation contours for the closed case. To further depict the posterior predictive variance the upper right image shows the smoothed map of standard deviations. The bottom images show the equivalent results obtained from 100,000 runs of the single contaminant Gibbs Sampler with linear trend.

As would be expected, the standard deviation maps show higher uncertainty where there are few "local" observations and also where the expert was less sure in the prior specification. The maps show that the two methods are giving similar results. This suggests that it would be beneficial where possible to use the closed update, as it is considerably quicker.

Table 4.2 gives the posterior hyperparameters, although the more descriptive number is the posterior variance parameter σ^2. In the closed case this was reduced from 400 to 252 (i.e. a reduction in standard deviation from 20 to 15.87), and in the MCMC update it reduced from 400 to 274 (i.e a reduction in standard deviation from 20 to 16.55). Again, the posterior $\mathbf{V}$ matrices are not directly comparable, as in the closed model this would be multiplied by σ^2. We can scale the $\mathbf{V}$ matrix in the second row of the table by multiplying it by $\frac{a^*}{d^*-2}$ which gives a value of 151.57 for the [1,1] entry, showing that the closed case gives a slightly reduced variance. The directional terms in the posterior $\mathbf{m}$ suggest the zonal specification is describing the contamination levels well, and there is no linear trend in the residuals.

Cross validation for Arsenic

To check the model further we can use a leave one out method of cross validation. By leaving out each observation point one at a time and using the other values to predict the left-out point, we can see how the model and update are performing. This is a useful method of checking the model without having to go out into the site

and obtain extra information, and will help us refine the model to avoid undertaking a sampling search routine using a poor model.

Table 4.4 gives the results of this exercise for Arsenic. Most values are close to the actual observed value, and within two posterior predictive standard deviations. We use this measure of closeness based on Normal distribution theory, as a rough indicator that we would expect 95% of observations to lie within two standard deviations of the mean. However, a couple of the values are more different, looking at their placement suggests the blue zone in the west of the site could be extending a little further than the expert suspects. Again, the results are similar for both methods, with slightly lower predictive standard deviations in the MCMC update.

4.2.3 Zinc contamination

Prior Specification

The expert gave a seven zone specification for Zinc, which was fairly similar to that of the Arsenic zonal specification. With Zinc it was again deemed that the blue zone in the west was possibly more extensive than originally described. We dealt with this as by adding a transitional zone as discussed in Section 3.4.2, this time on top of part of the original blue zone, and onto the light green zone, to account for the possibility of the spread of Zinc contamination after the remediation of the site. The prior variance parameter was chosen to be $\sigma^2 = 250^2$. Figure 4.9 and Table 4.6 give the prior specifications, where *smoothz* is a linearly decreasing function from 3000 to 648.

$$smoothz = 3000 - 1352\left(\frac{x_e - 65}{95}\right) \tag{4.1}$$

Updating and cross validation

Figure 4.10 depicts the predictive Zinc levels smoothed over the site, and the standard deviation levels. Interestingly in this case, the standard deviations to the very west of the site become somewhat lower than might be expected. This may be due

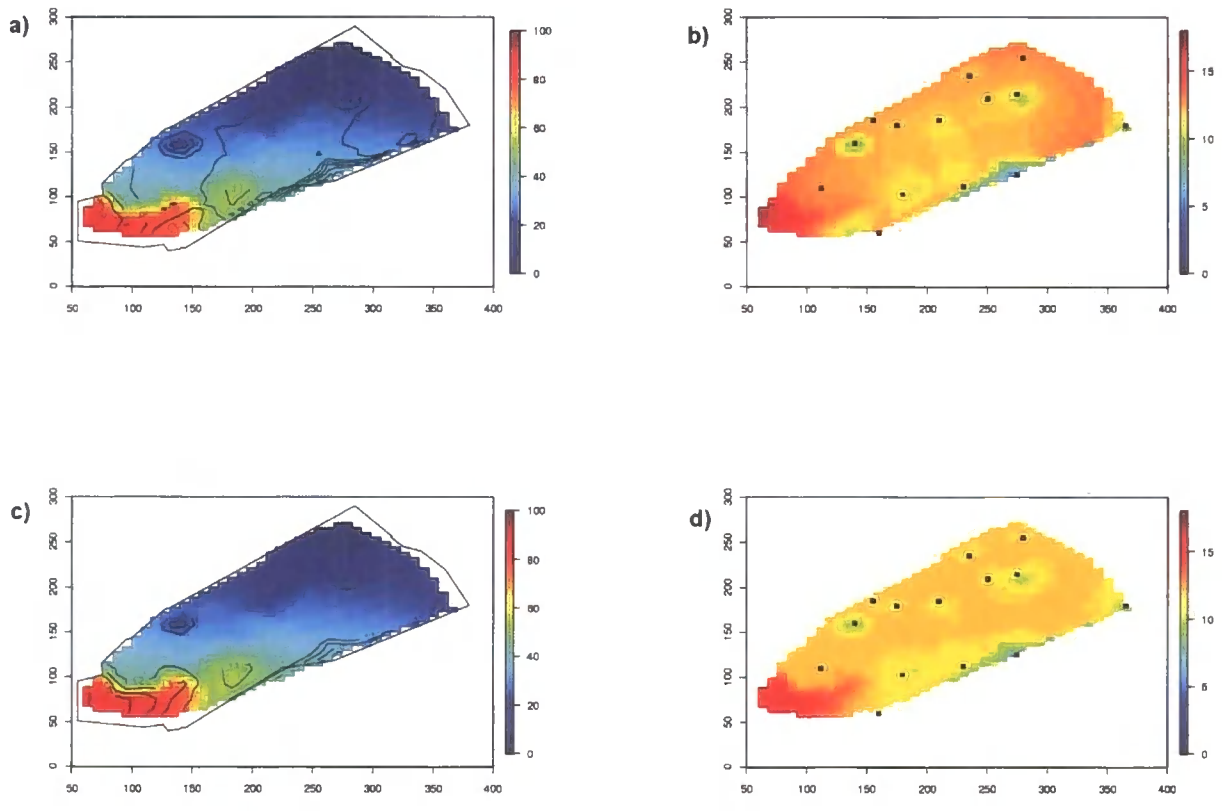


Figure 4.7: Arsenic posterior predictive distributions, comparison of methods. a) shows the closed update, mean with standard deviation overlaid, b) shows standard deviation. Single MCMC approach, mean with standard deviation overlaid in c), and d) shows standard deviation.

		Closed			Cond conj		
Location	Value	Pred mean	Pred sd	Standardised	Pred mean	Pred sd	Standardised
(160,60)	72	98.84	16.64	-1.61	100.17	15.49	-1.82
(230,112)	42	39.96	13.10	0.16	40.96	12.43	0.08
(275,125)	38	36.20	13.73	0.13	36.10	12.51	0.15
(365,180)	15.5	31.30	15.84	-1.00	30.62	12.23	-1.24
(112,110)	44	36.91	13.69	0.52	37.68	12.48	0.51
(140,160)	12	32.49	11.90	-1.72	33.38	10.92	-1.96
(180,103)	68	37.61	10.03	3.03	38.53	9.67	3.06
(275,215)	19	20.03	12.71	-0.08	19.73	12.49	-0.06
(155,185)	24	23.67	13.13	0.03	24.13	12.49	-0.01
(175,180)	37	22.22	12.04	1.23	22.40	11.89	1.23
(210,185)	15	23.00	12.23	-0.65	23.15	12.25	-0.67
(235,23)	8.1	13.33	13.20	-0.40	12.94	12.48	-0.39
(250,210)	23	15.56	12.45	0.60	15.29	12.30	0.63
(280,255)	7.4	7.11	13.90	0.02	6.348	12.52	0.08

Table 4.4: Cross validation for Arsenic observations

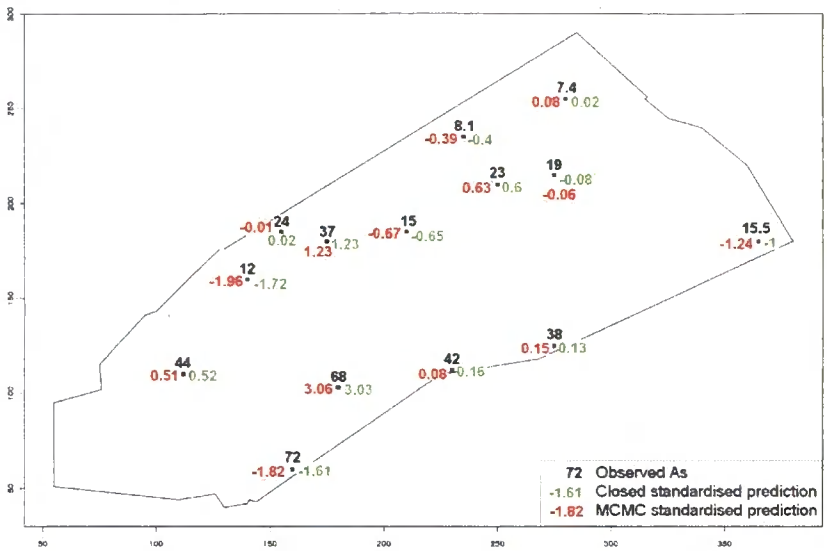


Figure 4.8: Cross validation results: Standardised prediction values for Arsenic at observation locations, for closed update in green and MCMC in red

	a	d	$\mathbf{m}$	$\mathbf{V}$
Closed prior	62500	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix}$
Closed posterior	469921.4	17	$\begin{pmatrix} -64.37 \\ -0.19 \\ -0.40 \end{pmatrix}$	$\begin{pmatrix} 1.020 & -0.003 & -0.002 \\ -0.003 & 0.000 & -0.000 \\ 0.002 & -0.000 & 0.000 \end{pmatrix}$
Cond conj prior	62500	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 625000 & 0 & 0 \\ 0 & 625000 & 0 \\ 0 & 0 & 625000 \end{pmatrix}$
Cond conj posterior	590233.2	17	$\begin{pmatrix} -68.03 \\ -0.19 \\ -0.38 \end{pmatrix}$	$\begin{pmatrix} 40872.12 & -121.60 & -70.50 \\ -121.60 & 1.05 & -0.54 \\ -70.50 & -0.54 & 1.02 \end{pmatrix}$

Table 4.5: Prior and posterior hyperparameters in the closed and conditional conjugate case for Zinc

Zone	$E(y(\underline{\mathbf{x}}))$	**G**	$\underline{\theta}$	κ	$\underline{x} \in$	1	2	3	4	5	6	7	8
1	4000	10	2	5	$a_1(\underline{x})$	0.975	0	0.222	0	0	0	0	0
2	*smoothz*	5	2	100	$a_2(\underline{x})$	0	0.950	0.220	0.222	0.200	0.132	0	0
3	3000	3	2	100	$a_3(\underline{x})$	0.222	0.950	0.222	0	0	0	0	0
4	350	0.5	2	5	$a_4(\underline{x})$	0	0.127	0	0.950	0.132	0.200	0.126	0
5	648	2	2	15	$a_5(\underline{x})$	0	0.127	0	0.127	0.950	0	0.127	0
6	648	2	2	15	$a_6(\underline{x})$	0	0.126	0	0.127	0	0.950	0.127	0
7	2000	3	2	100	$a_7(\underline{x})$	0	0	0	0.126	0.200	0.200	0.950	0.222
8	4000	10	2	5	$a_8(\underline{x})$	0	0	0	0	0	0	0.222	0.975

Table 4.6: Prior specification for Zinc contamination at Site R

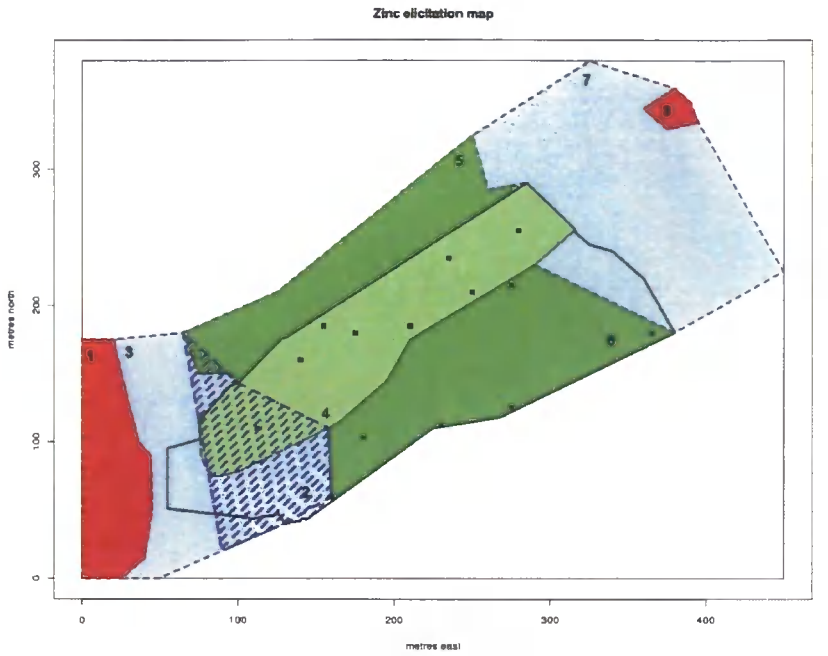


Figure 4.9: Expert elicitation map for Zinc, including observation locations

to the values observed at the two nearest locations fitting the prior specification better than in the case of Arsenic and Lead. The standard deviation plot shows that we are more uncertain in the lower portion of the site, which is intuitive as there is less data. The posterior σ^2 is 177^2 in the closed case, and 198^2 in the MCMC, which is a marked reduction from the prior of 250^2. The cross validation again shows promising results, with all predicted values within one or two standard deviations of the actual observation. The exception being the third value, where 400 was predicted when 930 was observed. While this is within three standard deviations, it still highlights a potential isolated hotspot or diagnostic warning. We will bear this in mind when looking at the sampling schemes and related uncertainty reporting. Both approaches perform similarly as expected.

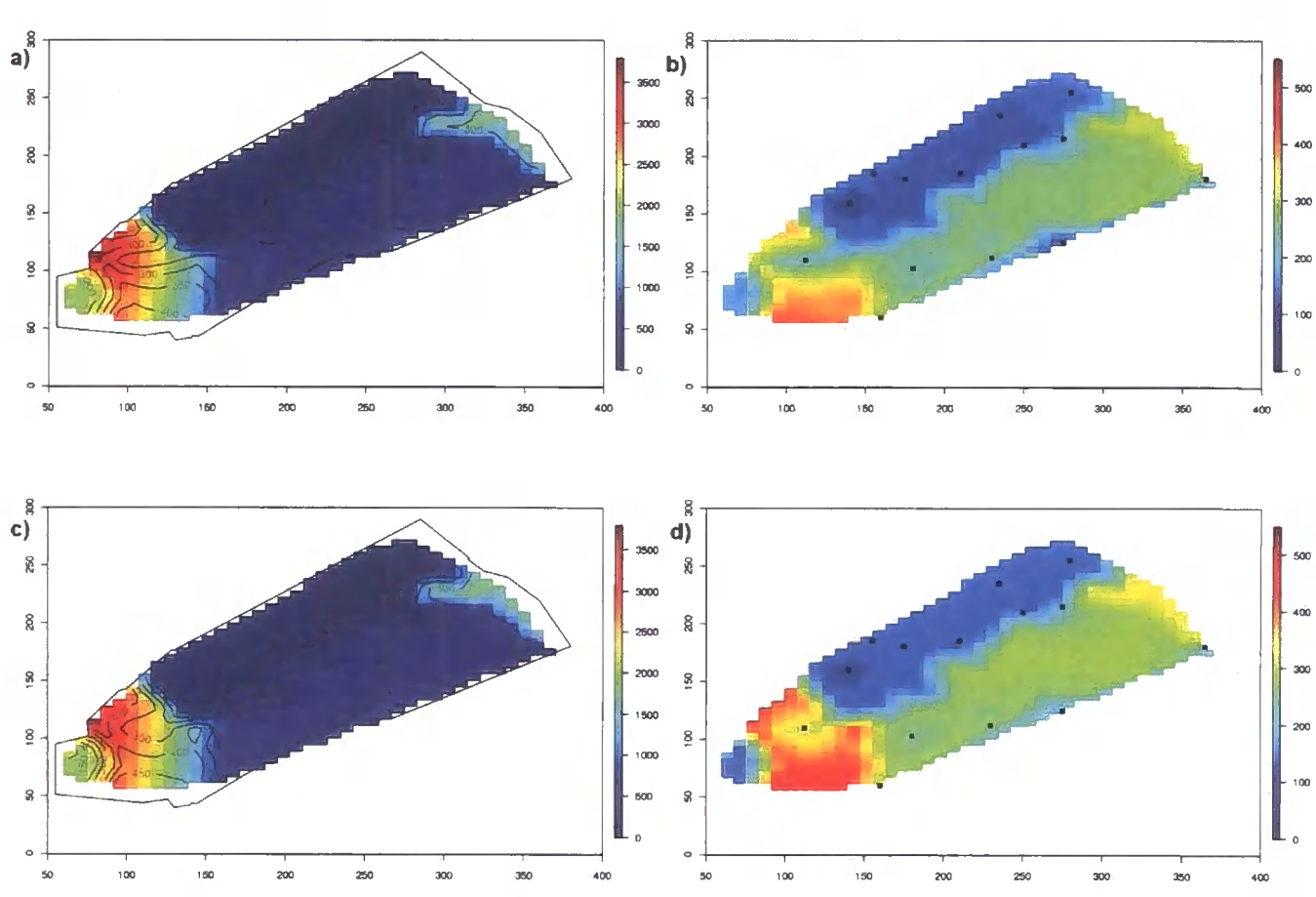


Figure 4.10: Zinc posterior predictive distributions, comparison of methods. Closed update, mean with standard deviation overlaid in a),b) shows standard deviation. Single MCMC approach, mean with standard deviation overlaid in c), d) shows standard deviation.

		Closed			Cond conj		
Location	Value	Pred mean	Pred sd	Standardised	Pred mean	Pred sd	Standardised
(160,60)	1800	2110.7	184.2	-1.69	2117.59	135.32	-2.35
(230,112)	380	541.6	274.1	-0.59	558.72	293.19	-0.61
(275,125)	930	400.0	250.0	2.12	397.28	259.50	2.05
(365,180)	230	552.6	311.9	-1.03	551.25	288.83	-1.11
(112,110)	2700	2014.3	508.9	1.35	2014.26	561.24	1.22
(140,160)	100	254.5	148.3	-1.04	253.78	143.26	-1.07
(180,103)	870	497.6	254.3	1.46	486.58	276.80	1.39
(275,215)	160	475.8	256.4	-1.23	478.75	283.66	-1.12
(155,185)	170	204.0	151.1	-0.22	202.59	149.05	-0.22
(175,180)	310	168.0	136.4	1.04	167.56	143.53	0.99
(210,185)	200	180.0	137.6	0.15	179.62	149.04	0.14
(235,23)	79	172.5	145.1	-0.64	174.58	146.79	-0.65
(250,210)	310	124.9	133.6	1.39	122.66	139.79	1.34
(280,255)	100	146.7	162.0	-0.29	147.15	148.43	-0.32

Table 4.7: Cross validation for Zinc observations

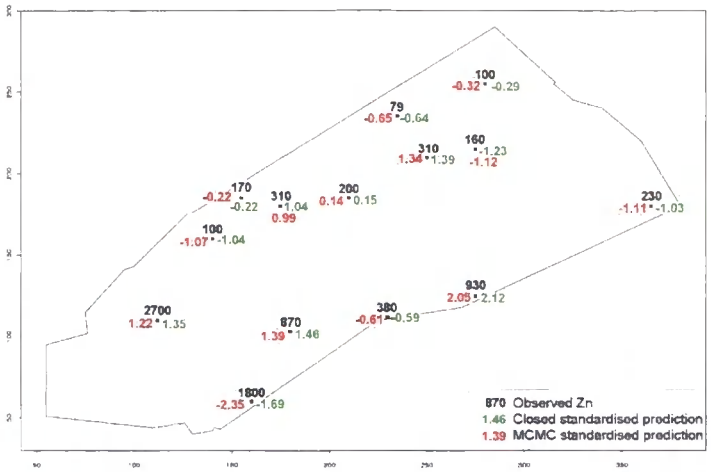


Figure 4.11: Cross validation results: Standardised prediction values for Zinc at observation locations, for closed update in green and MCMC in red

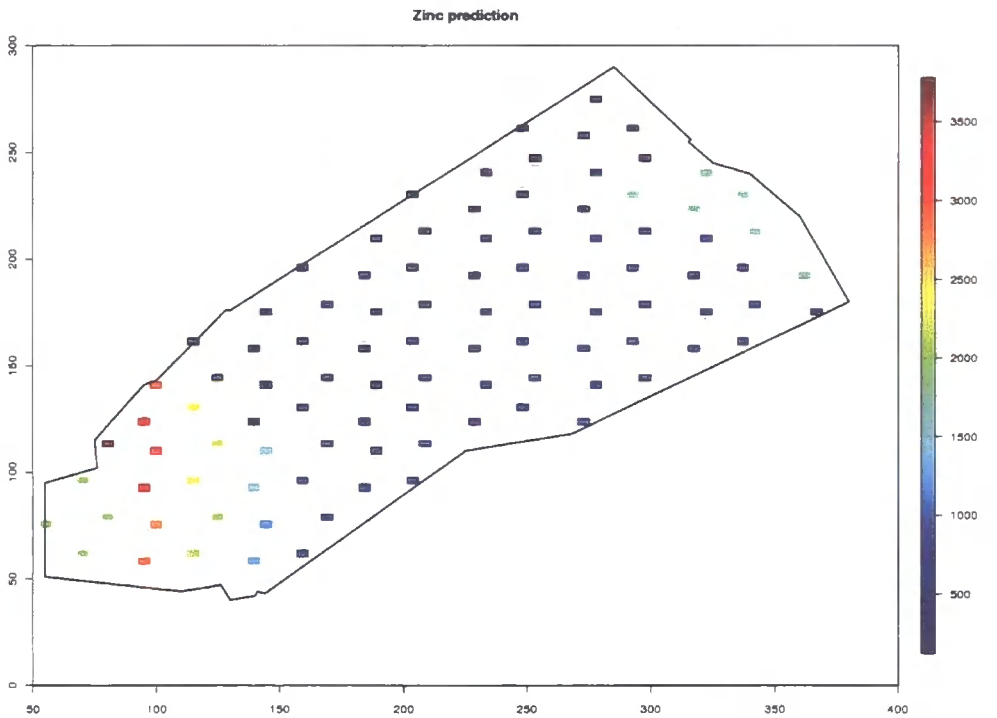


Figure 4.12: Plot displaying the prediction locations only, for Zinc closed prediction, rather than the interpolated smooth surface.

4.2.4 Total PAH contamination

Prior specification

This case is completely unrelated to the three metals we have analysed. The analysis of total PAH is more problematic than the first three contaminants. From discussion with the expert, the distribution of PAH can be extremely variable and can comprise very specific hotspots. The contaminants we combine in this case to comprise total PAH were chosen by the expert, and are Dibenzo(a,h)anthracene, Benzo(a)pyrene, Benzo(b)fluoranthene, Benzo(a)anthracene and Napthalene. The belief is that it will not be possible to resolve as much uncertainty in this case as we can in the metals. This will have implications for the decision process, as we shall encounter in Chapter 5. On inspection of the data we had to revise the original prior specification, as even more variation was encountered than expected. The main problem is the very large value we observe at the riverside, of 14826 mg/kg. Application of the maximum value test as suggested in CLR 7 does not suggest that this value is an outlier. However, as it skews the data to such an extreme we have several options to consider in order to attempt an improvement on our modelling capabilities.

1. Remove this point and reanalyse. We can either go back to the site and investigate whether this measurement was affected by a human error, or an isolated hotspot which can be dealt with separately. While we have discussed the problems associated with the tests described in CLR 7, the fact that the maximum value test is passed suggests retaining the large value in our update.

2. Include the point and deal with the huge uncertainties associated with this. This contaminant demonstrates the difficulties that can arise when attempting to model the spread and presence of contamination. We should not spend too long trying to manipulate the data to ensure our model can cope with it, rather accept the limitations and think about ways to deal with them.

3. Remove all Napthalene from the calculations as this is the dominant contaminant. Again, we don't want to remove a contaminant just because we have difficulty describing its spread, particularly when the levels observed are so

high. A high level of uncertainty will simply inform the investigator that further investigation must be undertaken to achieve better levels of confidence, or that remediation should be undertaken immediately as the cost of sampling to end up coming to the likely same conclusion will be high.

4. Add a very specific zone around this point in the revision of the prior specification. Going back to the prior specification and making adjustments in light of "surprising" observations is a suitable way to deal with a prior-data conflict. The aim of a Bayesian analysis is to update and refine our beliefs. If the data we see informs us that our prior specification was unreasonable, this is a different problem to that of poor model assumptions and we can adjust for this. This is discussed in detail in [43].

On consideration of these options, we will reconsider the expert's prior specification and also compare the updates with and without the extreme point included. Figure 4.13 shows the revised prior, with an added smoothing zone at the riverside, and Table 4.10 gives the zonal specifications. The original zonal specification the expert selected can be seen in Appendix B. The smoothing zone we use for this case is an exponentially decreasing function. The values were chosen to create a function which decreases smoothly from 4000 at the left of the zone, down to 10 at the far right of the zone.

$$smopah = 4000 \exp\left(-6.096825\left(\frac{x-270}{110}\right)\right) \tag{4.2}$$

Updating and cross validation

Recall that the posterior $\mathbf{V}$ matrices are not comparable for the two updates. In this case for the closed update $\mathbf{V}$ becomes very small. This is because the posterior mean for σ^2 becomes very large, particularly so when we include the extreme value. This is an example of our posterior variance increasing as a result of observing a very surprising value in comparison to what we expected to see. This is also reflected in Table 4.11, where we see standardised predictions for the extreme observation of 14.94 and 13.36 for the closed and MCMC approaches respectively. These results are also shown with the actual observation values in Figure 4.16.

	a	d	$\mathbf{m}$	$\mathbf{V}$
Closed prior	500000	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{pmatrix}$
Closed posterior	60715207	17	$\begin{pmatrix} 12.42 \\ -0.14 \\ -0.23 \end{pmatrix}$	$\begin{pmatrix} 0.003 & 0.000 & -0.000 \\ 0.000 & 0.000 & -0.000 \\ -0.000 & -0.000 & 0.000 \end{pmatrix}$
Cond conj prior	500000	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 2500000 & 0 & 0 \\ 0 & 2500000 & 0 \\ 0 & 0 & 2500000 \end{pmatrix}$
Cond conj posterior	76827270	17	$\begin{pmatrix} 12.49 \\ -0.13 \\ -0.22 \end{pmatrix}$	$\begin{pmatrix} 17189.130 & 31.043 & -117.250 \\ 31.043 & 1.067 & -1.305 \\ -117.250 & -1.305 & 2.000 \end{pmatrix}$

Table 4.8: Prior and posterior hyperparameters in the closed and conditional conjugate case for PAH. Case 1, all points included

	a	d	$\mathbf{m}$	$\mathbf{V}$
Closed prior	500000	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{pmatrix}$
Closed posterior	3584524	16	$\begin{pmatrix} 5.74 \\ 0.05 \\ -0.10 \end{pmatrix}$	$\begin{pmatrix} 0.003 & -0.000 & -0.000 \\ 0.000 & 0.000 & -0.000 \\ -0.000 & -0.000 & 0.000 \end{pmatrix}$
Cond conj prior	500000	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 2500000 & 0 & 0 \\ 0 & 2500000 & 0 \\ 0 & 0 & 2500000 \end{pmatrix}$
Cond conj posterior	4543263	16	$\begin{pmatrix} 5.96 \\ 0.05 \\ -0.11 \end{pmatrix}$	$\begin{pmatrix} 1081.611 & 2.308 & -7.764 \\ 2.308 & 0.068 & -0.085 \\ -7.764 & -0.085 & 0.131 \end{pmatrix}$

Table 4.9: Prior and posterior hyperparameters in the closed and conditional conjugate case for PAH. Case 2, extreme value removed

Zone	$E(y(\underline{\mathbf{x}}))$	$\mathbf{G}$	$\underline{\theta}$	κ	$\underline{x} \in$	1	2	3	4	5	6	7
1	100	0.1	2	2	$a_1(\underline{x})$	0.975	0.157	0	0	0	0	0
2	50	0.1	2	30	$a_2(\underline{x})$	0.157	0.975	0	0	0	0	0
3	10	0.001	2	5	$a_3(\underline{x})$	0.157	0.157	1	0	0	0	0
4	4000	2	2	25	$a_4(\underline{x})$	0	0	0	0.975	0	0.128	0
5	4000	2	2	25	$a_5(\underline{x})$	0	0	0	0	0.975	0.128	0
6	250	0.2	2	100	$a_6(\underline{x})$	0	0	0	0.222	0.157	0.975	0.141
7	*smopah*	1	2	50	$a_7(\underline{x})$	0	0	0	0	0.157	0.128	0.990

Table 4.10: Prior specification for total PAH contamination at Site R

Figure 4.13: Revised expert elicitation map for total PAH, including observation locations

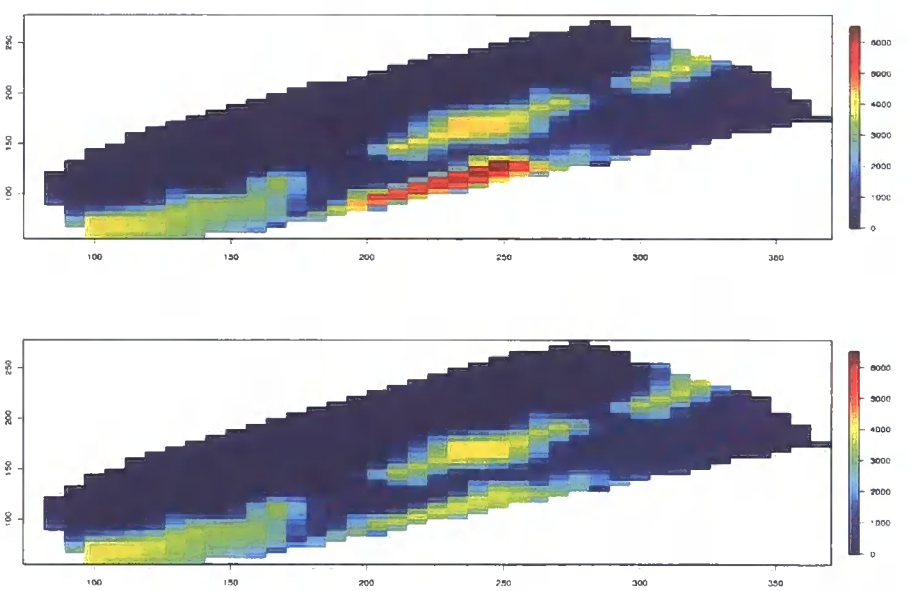


Figure 4.14: Effect of removing the large point on the posterior predictive means of total PAH. The upper figure has the large point included and shows the red region around this point, while the effect is removed in the lower figure where the value has been removed.

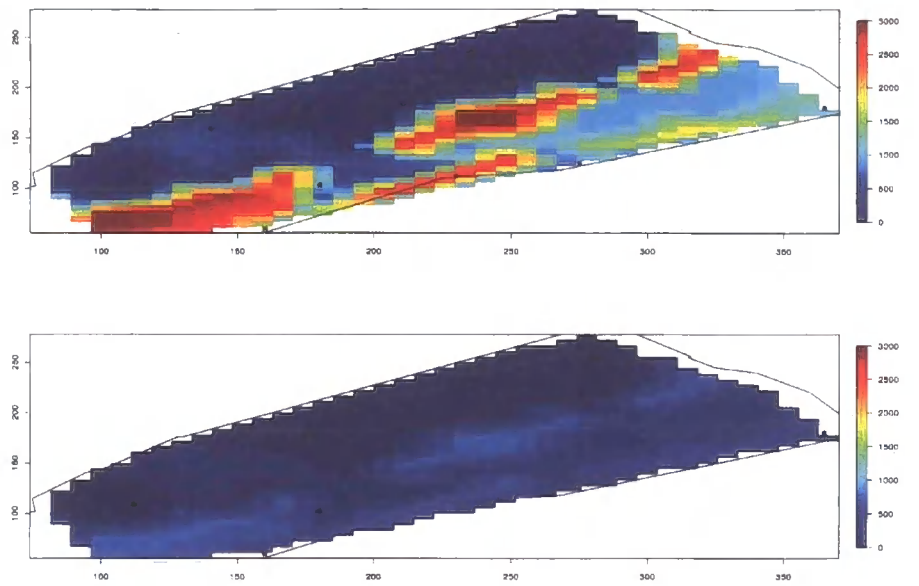


Figure 4.15: Effect of removing the large point on the posterior predictive standard deviations of total PAH. There is a marked effect on the standard deviations throughout the site when we remove the large point. The upper figure, where the point is included, shows areas of high uncertainty where the contamination is believed to be high, and few observations have been made. The lower figure, where the point has been removed before the update, shows the same regions to have a much reduced level of uncertainty.

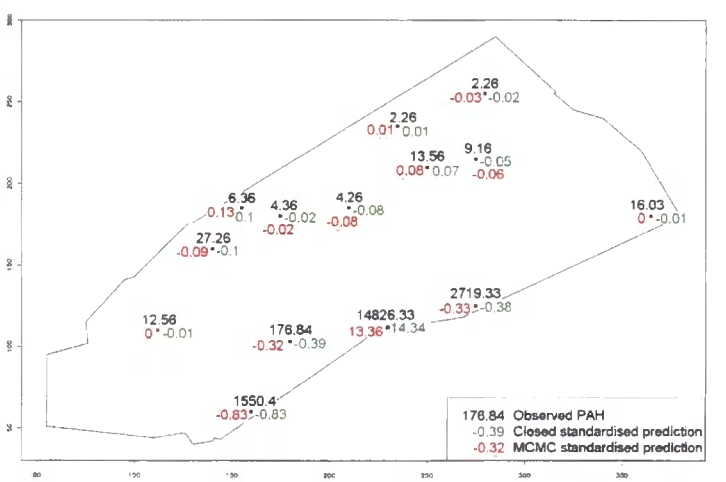


Figure 4.16: Cross validation results: Standardised prediction values for PAH at observation locations, for closed update in green and MCMC in red

		Closed			Cond conj		
Location	Value	Pred mean	Pred sd	Standardised	Pred mean	Pred sd	Standardised
(160,60)	1550.4	3969.36	2931.03	-0.83	4299.23	3293.25	-0.84
(230,112)	14826.3	3957.62	727.6	14.94	3881.65	819.30	13.36
(275,125)	2719.3	3523.75	2120.89	-0.38	3514.905	2378.67	-0.33
(365,180)	16.03	41.35	2139.94	-0.01	14.84	2400.90	0.00
(112,110)	12.56	13.72	121.16	-0.01	12.43	76.69	0.00
(140,160)	27.26	95.63	675.79	-0.1	95.32	767.56	-0.089
(180,103)	176.84	545.51	940.98	-0.39	519.22	1061.34	-0.32
(275,215)	9.16	14.13	94	-0.05	13.55	76.53	-0.06
(155,185)	6.36	-2.55	93.81	0.1	-3.37	76.80	0.13
(175,180)	4.36	6.03	75.23	-0.02	6.07	76.62	-0.02
(210,185)	4.26	10.53	73.96	-0.08	10.41	76.61	-0.08
(235,23)	2.26	1.13	82.96	0.01	1.36	76.35	0.01
(250,210)	13.56	8.05	77.21	0.07	7.78	76.53	0.08
(280,255)	2.26	3.84	85.49	-0.02	4.74	76.80	-0.03

Table 4.11: Cross validation for PAH observations

The closed update with the extreme observation removed gives a variance reduction (standard deviation is reduced from 707.1 to 506.00) whereas retaining the point for the Bayesian analysis gives an increase from 707.1 to 2011.9.

Figures 4.14 and 4.15 show the marked effect on the predictive surface depending on the inclusion or exclusion of the extreme values. While the mean surface is similar except for the area immediately in the vicinity of the large value, we see a large difference in the standard deviation surface. It would appear to suggest that leaving out the point would be beneficial in reducing our uncertainties. We could go ahead and do this, and decide separately and immediately that that area requires remediation. However, it may be counterproductive to remove this point and continue with the assumption that there are no more large hotspots of contamination to discover. This example reminds us that we can only hope to provide a tool with which to aid decision making, and that judgements regarding specific factors will often be required.

4.2.5 BaP contamination

Prior specification

The analysis of Benzo(a)pyrene alone as suggested by the expert is relatively easy once the total PAH prior specification has been made. The expert believed that the regional and correlation specifications could remain the same for BaP, the only change that would be required was a scaling on the expected values per region and the uncertainty levels. As the level of detection and SGV for BaP are both low, there is a good chance that the predictive surface will give values above the SGV for much of the site. However, this is reflective of the situation that often arises with BaP, and as such is a much debated guideline value.

Table 4.12 gives the prior specification for BaP, with the regions as displayed in Figure 4.13.

$$smobap = 650 \exp\left(-4.5\left(\frac{x - 270}{110}\right)\right) \tag{4.3}$$

Updating and cross validation

Table 4.13 and Figure 4.17 show the results of the update. We performed the closed update and the MCMC conditional conjugate approach. We see, as with the other contaminants, similar results for both approaches. The posterior σ^2 is reduced from 158^2 to 145.5^2 in the closed case, and from 158.1^2 to 154^2 in the MCMC update. Again we cannot resolve a lot of the uncertainty, as we see a large variation in the observations, and we have some regions with few or no observations.

Table 4.14 gives the results of the cross validation. Again, we see similar results to the PAH cross validation, where the values at the riverside are predicted badly, this is due to the extremely large values we see here. We will consider the large uncertainty here when we look at the decision and sampling analysis in Chapter 7.

Zone	$E(y(\underline{x}))$	$\mathbf{G}$	$\underline{\theta}$	κ	$\underline{x} \in$	1	2	3	4	5	6	7
1	10	0.5	2	2	$a_1(\underline{x})$	0.975	0.157	0	0	0	0	0
2	5	0.5	2	30	$a_2(\underline{x})$	0.157	0.975	0	0	0	0	0
3	2.5	0.01	2	5	$a_3(\underline{x})$	0.157	0.157	1	0	0	0	0
4	500	2	2	25	$a_4(\underline{x})$	0	0	0	0.975	0	0.128	0
5	400	2	2	25	$a_5(\underline{x})$	0	0	0	0	0.975	0.128	0
6	40	0.5	2	100	$a_6(\underline{x})$	0	0	0	0.222	0.157	0.975	0.141
7	*smobap*	1	2	50	$a_7(\underline{x})$	0	0	0	0	0.157	0.128	0.99

Table 4.12: Prior specification for BaP contamination at Site R

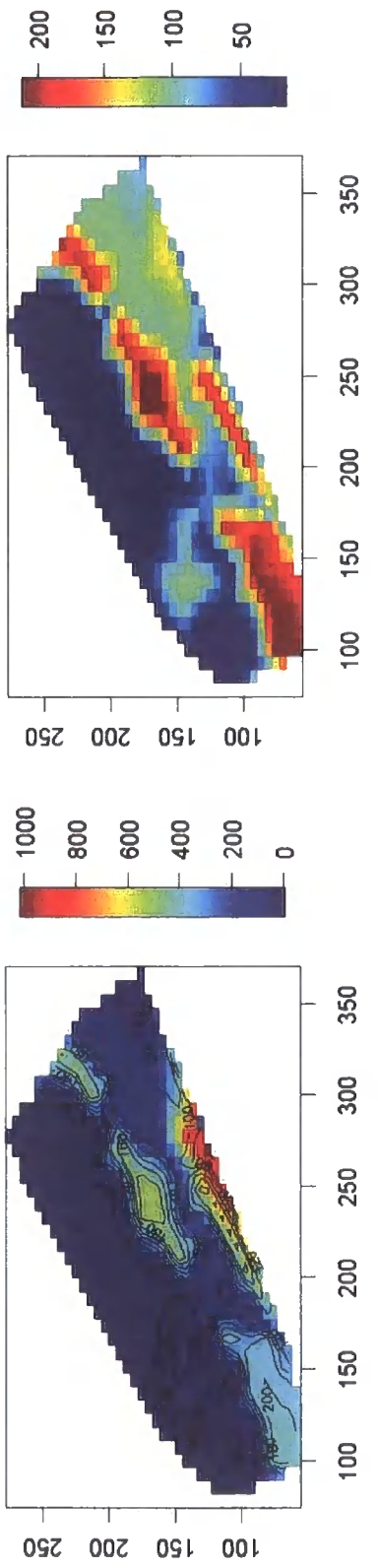


Figure 4.17: Posterior BaP predictive distribution, top image shows predictive mean surface with standard deviation overlaid, and bottom image shows standard deviation surface.

	a	d	$\mathbf{m}$	$\mathbf{V}$
Closed prior	25000	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{pmatrix}$
Closed posterior	317624	17	$\begin{pmatrix} 12.31 \\ 0.16 \\ -0.24 \end{pmatrix}$	$\begin{pmatrix} 0.003 & 0.00 & -0.000 \\ 0.000 & 0.000 & -0.000 \\ -0.000 & -0.000 & 0.000 \end{pmatrix}$
Cond conj prior	100	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 250 & 0 & 0 \\ 0 & 250 & 0 \\ 0 & 0 & 250 \end{pmatrix}$
Cond conj posterior	335470	17	$\begin{pmatrix} 13.19 \\ 0.13 \\ -0.17 \end{pmatrix}$	$\begin{pmatrix} 180.659 & 0.334 & -1.249 \\ 0.334 & 0.041 & -0.047 \\ -1.249 & -0.047 & 0.057 \end{pmatrix}$

Table 4.13: Prior and posterior hyperparameters in the closed and conditional conjugate case for BaP.

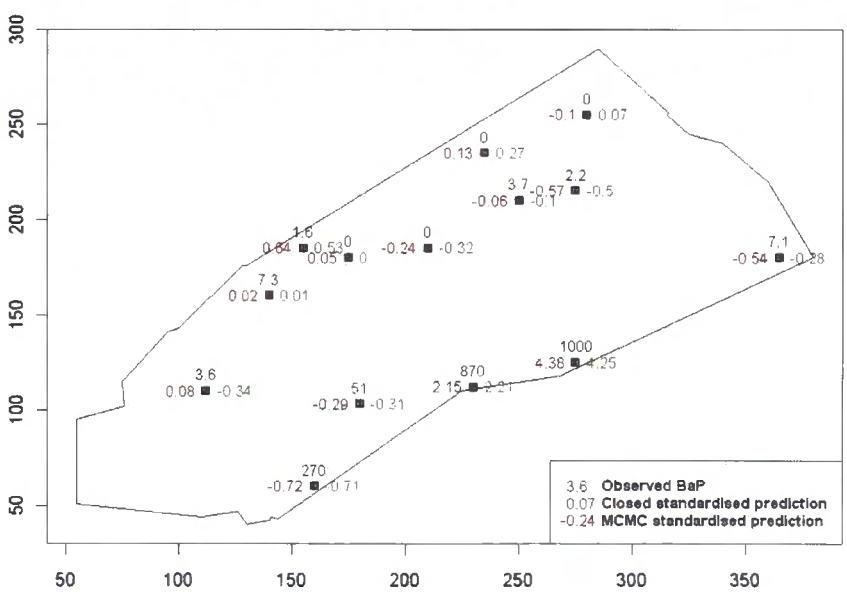


Figure 4.18: Cross validation results: Standardised prediction values for BaP at observation locations, for closed update in green and MCMC in red

		Closed			Cond conj		
Location	Value	Pred mean	Pred sd	Standardised	Pred mean	Pred sd	Standardised
(160,60)	270	422.68	214.58	-0.71	433.39	226.43	-0.72
(230,112)	870	456.68	187.31	2.21	449.9	195.83	2.15
(275,125)	1000	559.50	103.54	4.25	556.27	101.22	4.38
(365,180)	7.1	51.91	158.05	-0.28	94.14	161.45	-0.54
(112,110)	3.6	12.60	26.54	-0.34	2.21	16.35	0.08
(140,160)	7.3	6.65	109.53	0.01	5.4	116.64	0.02
(180,103)	51	85.20	108.78	-0.31	83.63	113.89	-0.29
(275,215)	2.2	12.55	20.53	-0.50	11.42	16.31	-0.57
(155,185)	1.6	-9.42	20.68	0.53	-8.89	16.28	0.64
(175,180)	0	0.06	17.18	0.00	-0.81	16.38	0.05
(210,185)	0	5.42	16.77	-0.32	3.91	16.44	-0.24
(235,23)	0	-5.02	18.71	0.27	-2.07	16.37	0.13
(250,210)	3.7	5.46	17.50	-0.10	4.75	16.4	-0.06
(280,255)	0	-1.38	19.49	0.07	1.65	16.44	-0.1

Table 4.14: Cross validation for BaP observations

4.3 Multiple contaminant update, MCMC approach

In order to demonstrate the multiple contaminant update we use the pairwise update of Arsenic and Zinc, and an update on the three metal contaminants Arsenic, Lead and Zinc to determine the effectiveness of this option against three single updates. We put into practice the discussion of between contaminant correlation from Section 3.7.1.

4.3.1 Selection of between contaminant correlation values

We will use the same prior information per contaminant, as in the previous section, but we now require a description of the relationship between contaminants. The expert did not give specific numbers, but answered questions about whether high, medium or low between contaminant correlation was expected. We use these in combination with the choice of variance parameters to decide upon reasonable values.

4.3.2 Multiple update and comparison with single approaches

We initially ran the multiple update with zero prior correlation as in Section 3.7.2 to ensure the results were equivalent to the single updates, and the results we got were as in the relevant single cases, with small rounding error.

Two contaminants

Looking first at the relationship between Arsenic and Zinc, we compare the zonal specification given for each, and asked the expert whether she expected the distribution of Arsenic and Zinc to be similar. We decided to include a non-zero between contaminant correlation only for the diagonal elements. That is, we compare the relationship between contaminants in the x_e and x_n directions, as well as the constant value, but not other terms such as $\text{Corr}(x_{As-e}, x_{Zn-n})$. It was deemed that comparing contaminants in the same direction only gave enough information, and any other elements were quite difficult to elicit using simple qualitative questions.

Using the methodology of Section 3.7.1 we proceed as follows, bearing in mind that we are giving a detailed zonal specification and so we give a prior $\underline{\beta}$ of $\underline{0}$. How-

ever, we can still think about using the α specifications to determine the correlation structure by looking at the variances.

1. Scale the contaminants to ensure they are comparable. SGV for Arsenic is 20mg/kg, so Zinc is about 32 times that at 648. We have considered the overall $\underline{\beta}$ variance contributions for each contaminant individually as 2000 and 625,000, and so leave Var(As)=2000, and use Var(Zn)=625000/31.25=20000 (slight rounding for computational convenience).

2. We decide to specify

$$\text{Var}\left(\underline{\beta}\right) = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix}$$

for the similar variance contribution, and for the individual contributions we give

$$\text{Var}\left(\underline{\beta}_{As^*}\right) = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix}$$

for Arsenic and

$$\text{Var}\left(\underline{\beta}_{Zn^*}\right) = \begin{pmatrix} 312.5 & 0 & 0 \\ 0 & 312.5 & 0 \\ 0 & 0 & 312.5 \end{pmatrix}$$

for Zinc.

3. This gives us the setup:

$$\begin{aligned} \text{Var(As)} &= \alpha_{As}^2 \text{Var}(\underline{\beta}) + \alpha_{As}'^2 \text{Var(As}^*) \\ \text{Var(Zn)} &= \alpha_{Zn}^2 \text{Var}(\underline{\beta}) + \alpha_{Zn}'^2 \text{Var(Zn}^*) \end{aligned}$$

Putting in the numbers we have gives

$$\begin{pmatrix} 2000 & 0 & 0 \\ 0 & 2000 & 0 \\ 0 & 0 & 2000 \end{pmatrix} = \alpha_{As}^2 \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix} + \alpha_{As}^{'2} \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix}$$

$$\begin{pmatrix} 20000 & 0 & 0 \\ 0 & 20000 & 0 \\ 0 & 0 & 20000 \end{pmatrix} = \alpha_{Zn}^2 \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix} + \alpha_{Zn}^{'2} \begin{pmatrix} 312.5 & 0 & 0 \\ 0 & 312.5 & 0 \\ 0 & 0 & 312.5 \end{pmatrix}$$

4. Assigning suitable numbers for the α's to match the experts opinion we get $\alpha_{As} = \sqrt{150} \approx 12.25$, $\alpha'_{As} = \sqrt{50} \approx 7.07$, $\alpha_{Zn} = \sqrt{1500} \approx 38.73$ and $\alpha'_{Zn} = \sqrt{16} = 4$.

5. This leads to a correlation of approximately 0.75 for each of the diagonal elements. The overall prior correlation structure is shown in Equation 4.4, we used

$$\text{Corr(A, B)} = \frac{\text{Cov(A, B)}}{\sqrt{\text{Var(A)}}\sqrt{\text{Var(B)}}}$$

to get the value of 0.75.

$$\begin{pmatrix} 1 & 0 & 0 & 0.75 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0.75 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0.75 \\ 0.75 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0.75 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0.75 & 0 & 0 & 1 \end{pmatrix} \tag{4.4}$$

The elements of interest in the update are the posterior σ^2 and the posterior joint $\mathbf{V}$. These, are summarised in Table 4.15, and we see that the values we obtain are very similar to the relevant single MCMC update. The posterior correlation

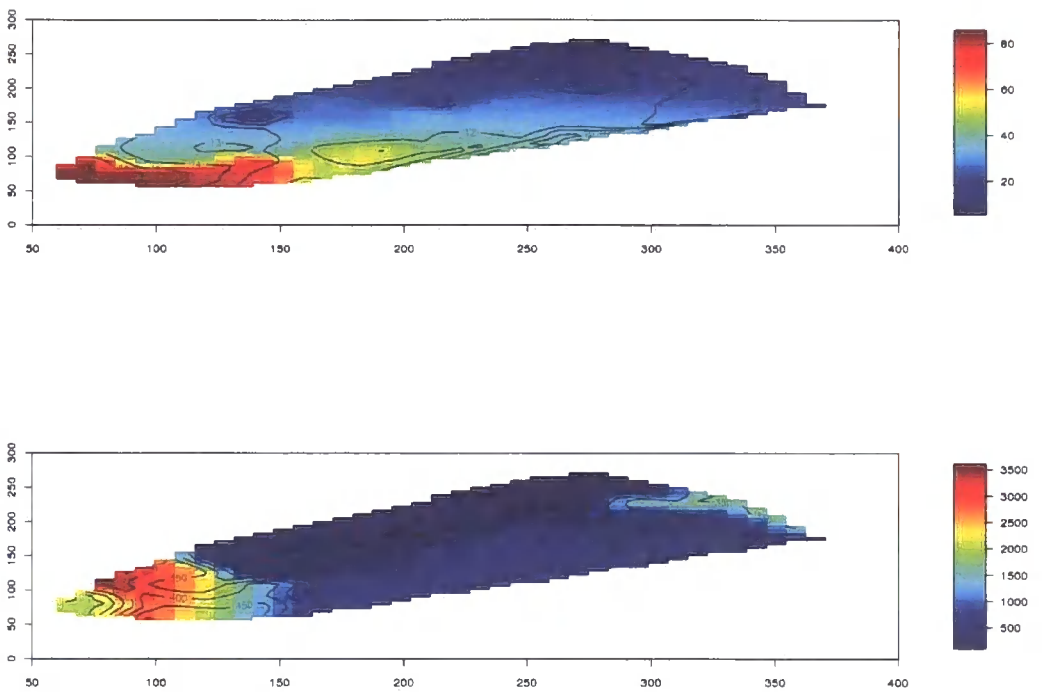


Figure 4.19: Prediction surface for Arsenic (top) and Zinc (bottom) when updated using a multiple approach, standard deviation contours overlaid. Compare with Figures 4.7, 4.10

matrix we compare with Equation 4.4 is

$$\begin{pmatrix} 1 & -0.295 & -0.558 & 0.12 & -0.059 & -0.048 \\ -0.295 & 1 & -0.516 & -0.069 & 0.035 & 0.035 \\ -0.558 & -0.516 & 1 & -0.056 & 0.04 & 0.014 \\ 0.12 & -0.069 & -0.056 & 1 & -0.579 & -0.313 \\ -0.059 & 0.035 & 0.04 & -0.579 & 1 & -0.562 \\ -0.048 & 0.035 & 0.014 & -0.313 & -0.562 & 1 \end{pmatrix}$$

and we see little between contaminant correlation remaining after the update. Also, Figure 4.19 shows the posterior predicted surface over the site, and we can compare this with the single update cases to see that there is little difference in the inference.

Case	Posterior σ^2 (and σ)	Posterior $\mathbf{V}$
Arsenic single	274.2 (16.56)	$\begin{pmatrix} 164.105 & -0.334 & -0.458 \\ -0.334 & 0.003 & -0.002 \\ -0.458 & -0.002 & 0.005 \end{pmatrix}$
Zinc single	39348.9 (198.37)	$\begin{pmatrix} 40872.12 & -121.60 & -70.50 \\ -121.60 & 1.05 & -0.54 \\ -70.50 & -0.54 & 1.02 \end{pmatrix}$
Multiple	281.6 (16.78) 40185.8 (200.5)	$\begin{pmatrix} 169.492 & -0.321 & -0.513 & 312.89 & -0.811 & -0.653 \\ -0.321 & 0.003 & -0.002 & -0.758 & 0.002 & 0.002 \\ -0.513 & -0.002 & 0.005 & -0.787 & 0.003 & 0.001 \\ 312.89 & -0.758 & -0.787 & 39985.231 & -122.016 & -65.346 \\ -0.811 & 0.002 & 0.003 & -122.016 & 1.112 & -0.618 \\ -0.653 & 0.002 & 0.001 & -65.346 & -0.618 & 1.088 \end{pmatrix}$

Table 4.15: Summary of results in single MCMC and multiple MCMC update for Arsenic and Zinc.

Three contaminants

We also ran the three contaminant update for Arsenic, Lead and Zinc. We follow a similar methodology for the two contaminant case to determine a correlation set up of:

$$\begin{pmatrix} 1 & 0 & 0 & 0.25 & 0 & 0 & 0.75 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0.25 & 0 & 0 & 0.75 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0.25 & 0 & 0 & 0.75 \\ 0.25 & 0 & 0 & 1 & 0 & 0 & 0.25 & 0 & 0 \\ 0 & 0.25 & 0 & 0 & 1 & 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0 & 0 & 1 & 0 & 0 & 0.25 \\ 0.75 & 0 & 0 & 0.25 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0.75 & 0 & 0 & 0.25 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0.75 & 0 & 0 & 0.25 & 0 & 0 & 1 \end{pmatrix} \tag{4.5}$$

where the ordering is Arsenic, Lead, Zinc.

Again we perform the multiple update, as before we use the same prior specification as for the individual updates along with the correlation as just specified. The updated parameters show that we obtain very similar results for all three contaminants. Table 4.16 and Figure 4.20 give the parameters of interest and the visual description of the predicted surface. We can again compare these to earlier results to see that there is no marked reduction in uncertainty for this case study.

Conclusions and recommendations

We did not perform any multiple updates for PAH or BaP and one of the metals as the expert has already informed us that the PAH contamination should be analysed separately as the two types of contamination will occur separately, and so we don't wish to learn about metals and organics together.

After performing the multiple update for both two and three of the metals, in this case it appears to be beneficial to use the single closed updates. This is because we yield similar results compared to the MCMC approach, and a multiple approach is deemed unneccessary as the secondary information does not substantially improve our ability to learn about the contaminant of interest. We also reduce the compu-

Case	Posterior σ^2 (and σ)	Posterior $\mathbf{V}$
Arsenic single	274.2 (16.56)	$\begin{pmatrix} 164.105 & -0.334 & -0.458 \\ -0.334 & 0.003 & -0.002 \\ -0.458 & -0.002 & 0.005 \end{pmatrix}$
Zinc single	39348.9 (198.37)	$\begin{pmatrix} 40872.12 & -121.60 & -70.50 \\ -121.60 & 1.05 & -0.54 \\ -70.50 & -0.54 & 1.02 \end{pmatrix}$
Lead single	33648.2 (183.43)	$\begin{pmatrix} 31399.50 & -68.67 & -80.23 \\ -68.67 & 0.62 & -0.36 \\ -80.24 & -0.36 & 0.87 \end{pmatrix}$
Multiple	278.5 (16.69) 33782.2 (183.80) 40158.6 (200.4)	$\begin{pmatrix} 167.81 & -0.33 & -0.49 & 61.48 & -0.17 & -0.14 & 319.67 & -0.93 & -0.61 \\ -0.33 & 0.00 & -0.00 & -0.04 & 0.00 & 0.00 & -0.75 & 0.00 & 0.00 \\ -0.49 & -0.00 & 0.01 & -0.28 & 0.00 & 0.00 & -0.76 & 0.00 & 0.00 \\ 61.48 & -0.04 & -0.28 & 31112.6 & -66.61 & -81.19 & 855.48 & -3.07 & -0.88 \\ -0.17 & 0.00 & 0.00 & -66.61 & 0.61 & -0.36 & -2.37 & 0.00 & 0.01 \\ -0.14 & 0.00 & 0.00 & -81.19 & -0.36 & 0.88 & -1.97 & 0.01 & -0.01 \\ 319.67 & -0.75 & -0.76 & 855.48 & -2.37 & -1.97 & 39085.7 & -117.12 & -66.53 \\ -0.93 & 0.00 & 0.00 & -3.07 & 0.00 & 0.01 & -117.12 & 1.07 & -0.59 \\ -0.61 & 0.00 & 0.00 & -0.88 & 0.01 & -0.01 & -66.53 & -0.59 & 1.06 \end{pmatrix}$

Table 4.16: Summary of results in single MCMC and multiple MCMC update for Arsenic, Lead and Zinc.

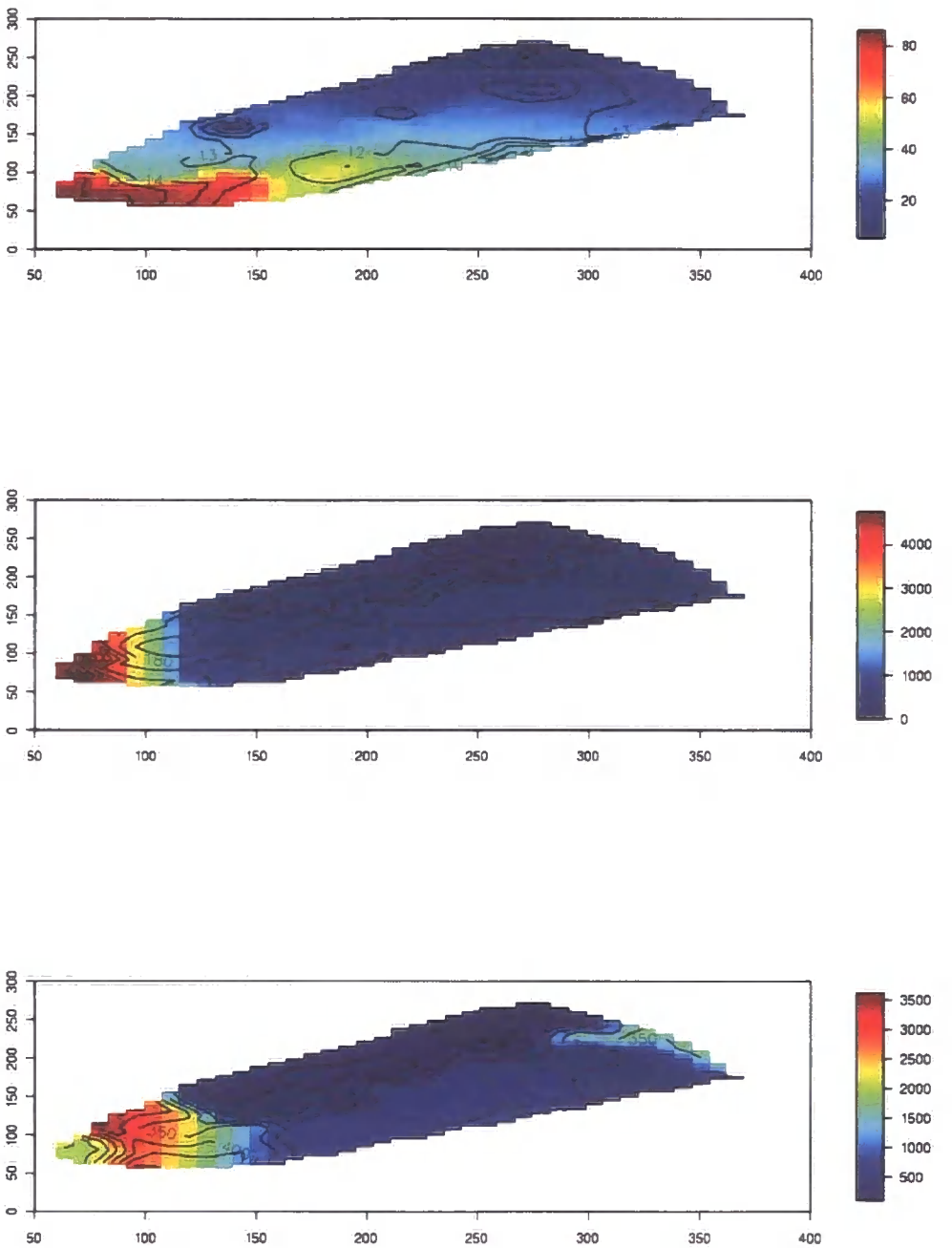


Figure 4.20: Prediction surfaces for Arsenic (top), Lead (middle) and Zinc (bottom), when jointly updated, with standard deviation contours overlaid.

tational time significantly by updating separately. It became apparent, that with a detailed zonal specification, and contaminants which are fitting the expected model well, it is not beneficial to use the multiple approach as we do not learn enough to justify the added computational burden.

Comparison to results from investigation

No prediction was performed in the site investigation report, rather an assessment of risk present from the observed values and recommendations for further sampling. We shall look in detail at these recommendations when we have performed the sampling and decision analysis in Chapter 7. Also, site specific target levels were used in some cases to determine the existence of SPOSH, and so we may end up drawing different conclusions from the site report.

As we have now seen how to model the site in practice, using initial observations combined with expert information, we move on to the decision and sampling process.

Chapter 5

The expected value of a sampling design

Now that we have introduced a suitable spatial model with which we can describe contamination over a site, and can update this model given sample data, we would like to make informed decisions regarding sampling strategies and terminal decisions; that is whether to continue sampling, and if so, where, or whether to stop and remediate (and if so, where and how), or to take no action.

Our key objective is the development of a method with which we can measure the information gained from a candidate sampling design, in order to provide a procedure for the comparison and selection of optimal (or near optimal) designs.

We need to consider how to measure "information gained", and also how we decide whether a design is good enough when we label it as near optimal. In theory, it should always be possible to find the optimal solution. However, in reality we are limited by computational constraints, through time or complexity of calculations.

The next sections discuss the concept of expected loss, and introduce utility theory as a way to compare designs. We will then consider the way in which we can implement a Bayesian decision analysis strategy for the problem of site investigation, and expand the problem to allow for the views of all potential stakeholders. Initially we shall look at the decision procedure when we have no further sampling to undertake, and wish to plan a remediation strategy. We will then introduce candidate designs and consider the expected benefit associated with undertaking sampling at

the candidate locations. Finally we look at decision making for multiple contaminants. Throughout the chapter we will use the hypothetical dataset to demonstrate the methodology. In Chapter 7 we shall implement the theory given in this and the next chapter for the real case study.

5.1 Selection of design criteria

Several methods of selecting optimal sampling designs have been discussed in the literature, and the two key factors to combine are the criterion used for evaluating the value of the design and the method of searching for the optimal design. In this chapter, we shall deal with the first factor, that of determining a way to compare designs in a sensible way, and then we shall look at the selection of designs in the next chapter. Criteria discussed in the literature include variance minimization [87] and maximum entropy (one of the criteria used in [2]). Many of the methods used are aimed at learning about the covariance structure, although there are those that specifically consider the decisions to be made regarding contaminated land, [17,78]. After the introduction of a simple approach, we shall expand our methodology to include an approach similar to that of [78]. A series of papers by Freeze *et al* [45,46,63,79] consider the problem of decision analysis in the investigation of ground water, and several of the ideas discussed in these papers translate to the problem of soil investigation. The main idea of bringing together an uncertainty model for the spread of contamination with a decision framework is the way in which we proceed also.

By using Bayesian decision theory, and considering the probability and the consequences of each outcome, having implemented each candidate design we can compare designs in terms of the expected benefit, and select the best. We will introduce a way to measure the expected benefit. We now summarise the key ideas we will use in the development of our method.

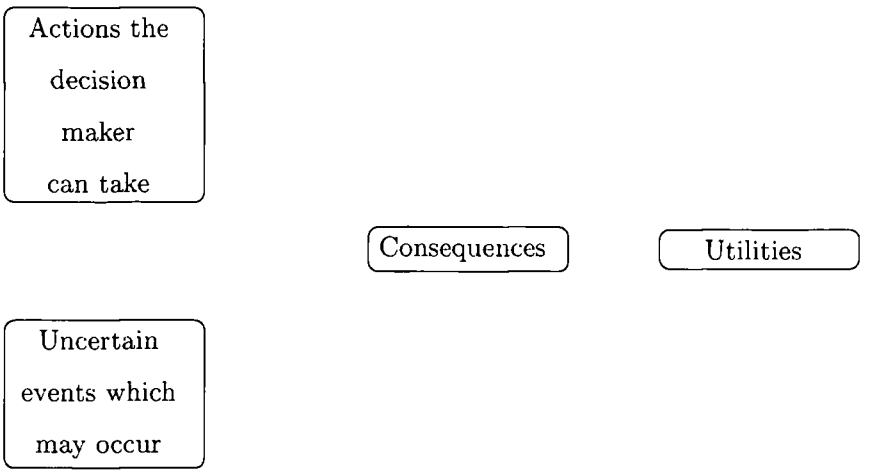


Figure 5.1: The key elements involved in a decision analysis.

5.2 Decision theory

As we will never know the true state of a site at every point, we have to make decisions between remedial options with some uncertainty. As Lindley discusses in [59], we cannot eliminate the need for decision makers, only provide them with tools to help quantify their beliefs and aid rational decision making. If we can list the possible decisions and consequences, and assign probabilities to these consequences, then we can calculate the expected value of taking any particular decision. This is the reason we spent time building a detailed statistical model in Chapter 3, to incorporate as much information from the desk study and experts as possible. In order to make these decisions comparable we assign utility values to the possible outcomes.

For a comprehensive overview of decision theory, [?, 32] introduce and describe the key principles in detail. The structure of decision theory breaks down into the elements shown in Figure 5.1. We then search over all actions to maximise utility and so the decision function will map the probability space to an optimal action.

Making one remediation decision for the whole site, based on one summary statistic, is an unrealistic set-up. In the current practice, if a small zone warrants further investigation, as determined by the results of the MVT or other test, then the investigator will go back to the site and attempt to delineate the extent of the contamination in order to determine the area which requires treatment.

We will need to decide at what scale we would like to make our decisions. There are several options which we can consider. We would like to make the correct decision for every single particle of soil on the site. While this is impossible, we can do our best within the computational constraints to achieve a version of this. If we place a grid over the site, we can consider making a decision regarding remediation at each of the grid points, and sum over these points to obtain an expected cost of remediation for the site. This grid can be as coarse or fine as necessary in the view of the decision maker (DM), within the computational and time constraints. Figure 5.3, gives an example of such a map.

Each inference grid point s is the centre of the "remediation unit" (or spatial unit as defined in [78]), and the decision selected will be for that whole remediation unit. We will sum over the n_I inference points to obtain the total expected loss for the site. The probability we calculate for each point s will be treated as being representative of the related remediation unit.

In general we shall use a regular grid for this, so the remediation units are all the same size, and we can determine the losses for each remediation unit with ease. Therefore the losses per grid square will change if a different density of grid is selected. We could integrate over each remediation unit to obtain an exact probability for the area. However, our approximation will be computationally much faster.

An alternative could be to make a decision for each averaging area as defined in [21], or to make a decision within each of the zones chosen by the expert in the prior specification.

The placement of the inference points may lead to undesirable features. For example, in Figure 5.3, the outermost inference points are on the site boundary, and so the remediation units around these points extend beyond the boundary. That is, having inference points on the site boundary leads to uneven remediation units within the site area. This is not a huge problem as the results obtained will not form a strict plan for remediation which must be followed. Rather, an indication of what decision is taken at each point will guide the DM in creating an effective remediation plan.

We will have to consider all possible consequences of the potential decisions, in

order to determine how we rank them in preference. We will have to ensure that the outcomes we specify are measurable by an available criterion. We could use the value of the predictive mean, expected maximum value or some other value of interest. As long as we always use the same criterion, the expected values will be comparable. In our set up, the outcomes are determined by the posterior probability of exceeding the relevant critical value, where we may have a number of critical values depending on the number of possible decisions. We will discuss such decision problems in more detail in the example of Section 5.5.

5.2.1 Decision alternatives

Initially we are looking at the situation where we have completed sampling, and have to make an immediate decision for each inference point, from the alternatives which were discussed in Section 2.3.1:

- No remediation or MNA. This method may be implemented in several ways, as no remediation does not necessarily mean no action is taken. Monitoring of the site may be one option, or a change in the way the site is used may decrease the risk to human health. The cost of this option may be non-zero.

- In-situ remediation. There are several different types of in-situ remediation, and, as well as the cost and efficacy of each method, other factors may be relevant in judging how desirable the option is, such as time to implement, long term capacity for preventing further contamination or potential for transfer of SPOSH to other receptors associated with a remediation technology.

- Ex-situ remediation. The same factors must be considered as for in-situ remediation.

- A combination of these may be required, as we shall see when we extend the decision analysis to cover multiple contaminants. For example, if we have remedial options for the treatment of organics and metals, the optimal decision may be to implement two types of remediation at a location to deal with both contaminants. If we can find a remedial alternative which is cheaper, and can

effectively treat both contaminants of interest, this should be included in the decision analysis.

Outcomes

In order to consider whether the implementation of a decision will be beneficial, and to compare decisions, we must be able to describe and quantify the consequences of these actions. At the most basic level, any decision we take can be assumed to be a success, or a failure. Once we have decided how we shall measure the success of an option, we also need to attach a cost (not necessarily financial) to describe the effect and impact of this outcome. The next sections will introduce a coherent framework for the quantitative description, and comparison of, these outcomes.

5.2.2 Loss and Utility

The actual loss associated with taking a particular action will depend on the true state of the site, although we do not know this (or we would be making decisions under certainty), and so we consider the expected loss for each decision d.

It is important to present the output of the decision analysis in a form that is interpretable for DMs, and it should be straightforward to compare different decisions or sampling locations. This can be achieved by measuring the performance of a design in terms of its expected utility. Lindley [59] describes utility as: "Utility is a number measuring the attractiveness of a consequence - the higher the utility, the more desirable the consequence ...".

The analysis of the problem at each point now takes the following steps, as given by Lindley [59]

- List possible decisions $(d_1, d_2, \ldots d_n)$
- List the uncertain outcomes $(O_1, O_2, \ldots O_m)$
- Assign probabilities $p(O_1), p(O_2), \ldots p(O_m)$
- Assign utilities $u(d_i, O_j)$

- Choose decision d^* to maximise expected utility, by calculating $\sum_j u(d_i, O_j)p(O_j)$ for each d_i

This is constructed in terms of a discrete situation. However, the idea extends to the continuous case in a natural way as we shall see in later sections.

We would like to choose the decision which maximises expected utility, or equivalently to minimise expected loss, provided that loss is measured on a utility scale. We may for example, convert all losses to money equivalents and then money can be converted to the utility scale [59]. From here on we will use loss, and so we aim to minimise loss.

The expected loss at a single inference point s, incurred by taking decision d, given initial observed data[1] y is

$$\mathrm{E}\left[L(O_s, d)|y\right] = L(O_{1,s}, d)p(O_{1,s} \mid y) + L(O_{2,s}, d)p(O_{2,s} \mid y) + \ldots + L(O_{m,s}, d)p(O_{m,s} \mid y) \tag{5.1}$$

where the expectation is taken over all possible outcomes $O_{i,s}$, $i = 1 \ldots m$ for each inference point, s. The optimal decision $d_s^*(y)$ for each point is the decision which incurs the minimum expected loss,

$$\min_d \mathrm{E}\left[L(O_s, d)|y\right] = \mathrm{E}\left[L(O_s, d_s^*(y))|y\right] = \mathcal{M}_s(y) \tag{5.2}$$

where we minimise over the set of possible decisions $d \in \mathcal{D}$

We calculate this for each inference point, and sum over the inference grid to give the total expected loss for the site $C(y) = \sum_{s=1}^{n_I} \mathcal{M}_s(y)$. When assigning a loss value for remediation or failure, we must bear in mind that for this set up it will refer to a single remediation unit as determined by the inference grid.

The specification of loss functions is again subjective, and will depend on the beliefs and attitude towards risks of the DM. Some aspects of the consequences will be treated as fixed, such as the financial cost of remediation, and an estimate of the cost of failure.

[1] the notation y will be used throughout this chapter to refer to the initial data $y(\underline{\mathbf{x}})$ for notational convenience.

These loss functions can be made as basic or complex as the DM wishes, as long as they are on a valid utility scale for the DM. We may specify loss functions which are a function of the "actual" contamination level at a point. For example, a failure cost may be an increasing function of the contamination level over the critical value, bounded at some point beyond which the impact of failure is judged to be the maximum possible. It may also be judged that different locations can have different consequences, and therefore different losses. This would be a way to account for the existence of complete S-P-R chains on the site, by assigning a higher failure cost in zones where a receptor is more likely or able to come into contact with contaminants. We will look in detail at more complex loss functions as we introduce further examples in this chapter.

5.2.3 Construction of utility functions

Utility functions may be used to demonstrate an individual's attitude to financial gain, that is whether they are risk averse or risk prone [32], or possibly risk neutral. The utility for money is usually decreasing and bounded. It is intuitive that a decision resulting in a higher financial return will have a higher utility than a decision offering a lower return. However, this relationship is not linear. As more money is obtained the attitude to financial gain will change, and an individual will become increasingly indifferent between two returns which are both very large compared to the current wealth and needs of the individual [4].

Utility theory also enables a DM to consider non-financial factors. This is useful in the case of contaminated land investigation as we have many potential positive and negative consequences to weigh up. The construction of a utility function in reality will involve a number of factors, known as the attributes or dimensions of the utility. These are determined subjectively through discussion with all relevant stakeholders, and the different factors can be weighted by importance accordingly. For example, in the previous section, the utilities could be divided into utility for money, and utility for time.

Two attributes are defined as utility independent if the following holds. X is utility independent of Y if preferences between all gambles with varying X and fixed

Y do not depend on the fixed value of Y. If X is utility independent of Y, and Y is utility independent of X, then X and Y are said to be mutually utility independent. This extends to many attributes, and is a feature which allows us to combine utility functions. Multiattribute utility theory is covered in detail in [56].

Mutually utility independent utility functions may be combined using either an additive or multiplicative form. In our work we choose to use the additive form, and standardise the partial utility functions to a [0,1] scale. For example, for the attributes money m, and time t, we construct $U(m,t) = aU(m)+bU(t)$. As discussed later, we will weight partial utilities according to importance, as judged by the decision makers, where the weights sum to 1. As we are working with negative utility in the form of loss, we will now consider partial loss functions specified on a [0,1] scale, where 1 is the worst, and 0 is the best. This [0,1] scaling allows for comparision between attributes.

Some of the possible factors which could be included for the case of soil remediation are discussed in [78], and include:

Possible attributes for multiattribute loss functions

- Human health. Depending on the associated decision, this could be a fixed loss of zero independent of contamination levels, if for example a remediation method is successful for any amount of contamination. It could be a value increasing above some threshold with the amount of "actual" contamination in the ground if the decision does not adequately deal with levels above said threshold. Or, if the associated decision is to take no action it would be a function increasing to a predetermined upper bound of 1, above which all levels of contamination are deemed equally damaging to human health. Some methods will deal with contamination more effectively than others, which will be reflected in this factor.

- Ecological receptors also require consideration during a site investigation, and we may wish to include a separate loss function for the negative effects in cases of particular ecological interest, i.e. a designated site of special scientific interest (SSSI).

- Remediation and failure costs (financial). This factor takes into account actual monetary losses, associated with implementing a remediation method, or alternatively the cost of dealing with contamination when a "failure" is encountered. For the no-action decision, this will be a loss of 0 when the actual contamination levels are below the relevant critical value. However it will increase markedly with contamination over the critical value. Some methods will be cheaper than others, which is reflected in this factor.

- Other factors may include time (until land can satisfactorily be deemed "clean"), productivity of land and future market value of land, health and safety risk to site workers (in investigation and remediation period). The cost of failure may further be broken down into factors such as lost business through damage to reputation, or legal costs arising from receptors who have been exposed. This may be useful in helping the expert build a realistic cost structure, by thinking about all the ways a loss can be incurred.

5.3 Bayesian Decision Analysis

We combine the updated contamination model with a decision theoretic approach, in order to make effective decisions when the state of the land is not fully known. We require the posterior predictive distributions for contamination levels, which is where we use the updated probability distributions described in Chapter 3. This process will now comprise of the following steps, as outlined in [17].

- Assume prior model for the spatial distribution of contaminants.

- Collect data, update beliefs, and perform diagnostic checks on the model specification to ensure we will obtain sensible results.

- Choose decision to minimise expected loss (expectation taken with respect to current posterior).

5.3.1 Losses associated with contaminated land investigation

In order to calculate the total expected loss associated with a sampling design, we must first divide the loss into the three separate factors. First we have the loss incurred by collecting the samples from the site. This value will be influential in deciding how many samples we can take. If the cost of sampling is prohibitively high, then it may never be beneficial to collect any samples. Equivalently, if remediation is free and always successful then we will always remediate. But these are trivial decisions, and we will not usually encounter such simple cases. We can split the sampling cost further into an initial setup cost, L_{init} (i.e. the hiring of equipment and the cost of staff), and then an incremental cost, L_{incr} per individual sampling location. We will label the sampling cost $L_S = L_{init} + n_C \times L_{incr}$, for a sampling design with n_C locations.

We will also have a loss, L_{R_i}, associated with implementing each remedial option (potentially zero if we decide to take no remedial action), and an loss, L_F, from failing to remediate when it is required. We may include a remedial option which can potentially fail. For example, suppose that method A can deal with all levels of contamination, but method B can only deal with "medium" contamination levels. So if we were to decide on method B when the contamination levels were in fact "high", we would incur both the cost of the remediation method L_{R_B}, and a failure cost L_{F_B}. However, when a method is deemed to be suitable for a certain contaminant range, we will assume it is always successful, i.e. we do not perform any faulty remediation. (This is not always the case in practice. As seen in our case study, a previous remediation attempt was deemed to have failed. However, this may be due to an incorrect decision, whereby an inadequate level of remediation was undertaken given the true contaminant levels on site).

These losses may not be strictly financial, and also may depend on the actual value of the contamination level rather than simply whether the contamination exceeds a critical value. Initially, we will consider a simple problem with fixed losses, and only two decisions.

5.3.2 Set up of the contamination decision problem

We begin with a very simple example to demonstrate the approach. We will assume a location on a site can be classified as "contaminated" or "clean" only, and that we allow only two decisions, "remediate" or "no action", so there are four possible decision-outcome pairs. Then we have the set up of Table 5.1. The values given in the table are losses. Therefore we prefer smaller values. If we can assign probabilities to the outcomes as in the process outlined in Section 5.2.2, then we can describe the information from the table in a decision tree as shown in Figure 5.2, and we can solve the problem exactly.

	Contaminated	Clean
Remediate	L_R	L_R
No action	L_F	0

Table 5.1: Simple decision table with four possible consequences

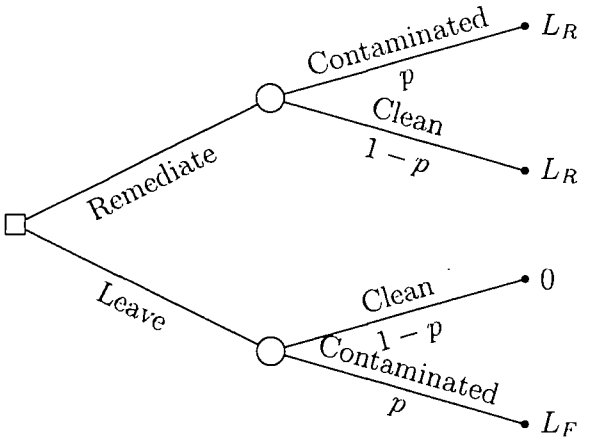


Figure 5.2: Decision tree for two decision, two outcome problem with costs and probabilities attached

For this set up, the most desirable situation would be when we do not undertake any remedial action and the site is clean. This is the most beneficial option as we do not have to spend any money (except possibly the cost of sample collection in order to make this decision), and there is no risk to human health. If the site is contaminated and we take no action there will be a cost incurred, L_F, which is the

most undesirable situation as we would have to remediate the land having discovered we made an erroneous decision. If $L_F < L_R$ then there is no problem to solve as the decision to take no action dominates the decision to remediate, so there is no outcome we could observe that would make remediation a preferable choice.

While the decision to remediate when the land is actually clean is an unneccessary one, it will still cost the same as remediating when the land is contaminated, and so has the same loss value attached to it.

In order to select the decision that minimises expected loss, we calculate the expected loss of each decision individually. Therefore, before looking at different candidate designs, we can identify the optimal decision strategy to implement immediately, when no further sampling is to be undertaken.

5.4 Expected loss for an immediate decision

If we wish to make a decision with no further sampling, we can use the current posterior probability of exceeding the relevant critical value to determine the optimal decision. We have used the initial observations y and updated our beliefs. Following the steps outlined in Section 5.3, we have:

- Listed the possible decisions, d_1, remediate or d_2, take no action
- Listed the uncertain outcomes, that the site is contaminated (above the critical value) O_1, or clean (below the critical value), O_2
- We can assign probabilities to these events using the posterior predictive distribution, $p(O_1 \mid y)$ and $p(O_2 \mid y)$
- Assign loss values by looking at Table 5.1.
- The final step is to choose the Bayes decision $d_s^*(y)$ for each inference point and sum over these points to give the total expected loss.

To select the Bayes decision (the decision which minimises the expected loss) we compare the two possible decisions and decide either to remediate or leave a unit

dependent on $\min\{E[L(O_s, d_1) \mid y], E[L(O_s, d_2) \mid y]\}$

$$\begin{aligned} E[L(O_s, d_1) \mid y] &= L(O_{1,s}, d_1)p(O_{1,s} \mid y) + L(O_{2,s}, d_1)p(O_{2,s} \mid y) \quad (5.3) \\ &= L_R p(O_{1,s} \mid y) + L_R(1 - p(O_{1,s} \mid y)) \quad (5.4) \\ &= L_R \\ E[L(O_s, d_2) \mid y] &= L(O_{1,s}, d_2)p(O_{1,s} \mid y) + L(O_{2,s}, d_2)p(O_{2,s} \mid y) \\ &= L_F p(O_{1,s} \mid y) \quad (5.5) \\ \mathcal{M}_s(y) &= \min\{L_R, L_F p(O_{1,s} \mid y)\} \end{aligned}$$

So the optimal decision is to remediate when the probability of the contamination exceeding the critical value at the point of interest is above the remediation cost divided by the failure cost.

$$d_s^*(y) = \begin{cases} d_1 & \text{if } p(O_{1,s} \mid y) > \frac{L_R}{L_F} \\ d_2 & \text{otherwise} \end{cases} \quad (5.6)$$

We make this assessment of an optimal decision and $\mathcal{M}_s(y)$ at each point on the inference grid and then sum to give the total expected loss $\mathcal{C}(y) = \sum_{s=1}^{n_I} \mathcal{M}_s(y)$. We will also have a decision vector, giving the optimal decision at each location on the inference grid, which is used to plot the "remediation map", as shown in Figure 5.3.

5.4.1 Example

We will now consider the optimal decision procedure for the hypothetical site, for the data set A2, as in Section 3.6.2. We will use the same prior specification as in the example of Section 3.6.2, and use a regular square grid over the unit site as an inference grid. We will assign losses as shown in Table 5.2.

	Below CV	Above CV
Remediate	10	10
Leave	0	50

Table 5.2: Loss table for the simple hypothetical example

So, from Equation (5.6), we see that at each inference point, the decision to remediate or take no action will depend on whether the posterior probability of

exceeding the critical value (CV) is above or below $\frac{10}{50} = 0.2$. We will assign a CV of 10 in this example. We follow the procedure as outlined above, using code written in R. The result of the expected utility calculation gives a total expected loss $\mathcal{C}(y)$ of 156.35 for the whole site. The losses breakdown into 130 assigned to remediation costs, and the remaining 26.35 is the expected loss associated with failure when we decide not to remediate. Figure 5.3 shows the optimal decision at each inference location, and also the observation locations and values. As an example of how the procedure works, we also performed the calculation using a CV of 5, and separately a changed cost structure, to $L_R = 10$, $L_F = 20$. The outcome of each of these calculations is depicted in Figure 5.4. The expected total losses in these cases were 214.25 and 127.87 respectively.

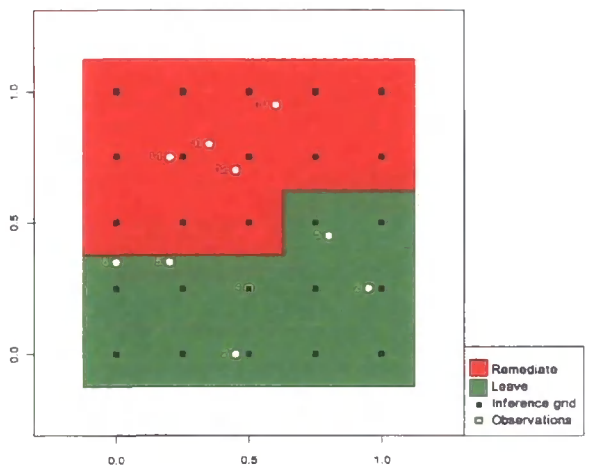


Figure 5.3: Remediation map for dataset A2, using regular 25 point inference grid, and CV of 10, with data and belief specification as in Section 3.6.1

Now that we have considered the expected loss when we make a decision with no further sampling, we require the expected value of implementing a sampling scheme and then making a decision. We can compare the two resulting losses and determine whether sampling is beneficial, i.e. will it reduce expected losses or change the decision.

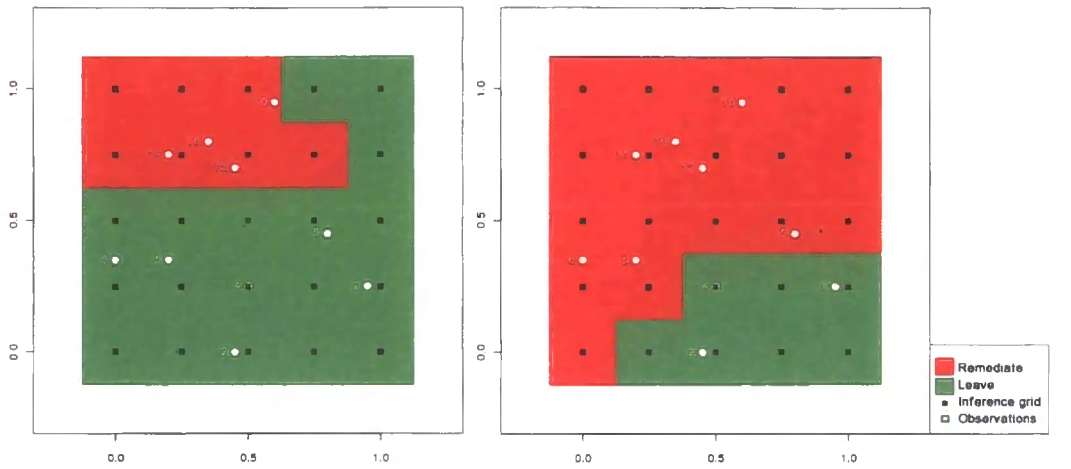


Figure 5.4: Demonstration of how altering values affects the decisions made at each inference point. Image on the left shows decision with lowered CV, and image on the right shows altered cost structure.

5.5 Expected loss for a candidate design

We will now consider the expected loss associated with the implementation of a particular candidate sampling design. There are an infinite number of potential grids we could place over a site, in order to take further samples and update beliefs regarding contaminant levels. Once we can assign a value to the worth of any one of these, we can search over a grid and compare expected losses to determine which points within the grid we should sample at in order to minimise the expected loss. We shall label a single design as δ, which is a specified set of new sampling locations, $\delta = (\underline{x}_{1\delta}, \underline{x}_{2\delta}, \ldots, \underline{x}_{n_\delta \delta})$. Every design does not need to have the same number of points, and so n_δ will not necessarily be the same for every design. We label the set of all candidate designs as Ω, where we start with a full grid, and Ω contains all possible subgrids of this. We will follow a similar procedure to that of calculating the expected loss when no further sampling is required. However, we now have to consider what we may see at the candidate locations, and how influential we believe the new data will be in helping us to make a decision. We will use the current posterior predictive distribution given any previously observed data in order to determine the probabililty distribution of the outcome at each inference point.

We can then look at the expected loss at that point having "observed" data at the candidate grid points and compare to the current decision and associated loss.

Again, the optimal decision strategy will be the one that minimises expected loss. We must now take into account the cost of taking the samples, but also that we do not have the actual observations at the candidate points, only probabilistic statements regarding what we may expect to see. To perform the full expected loss calculation, we would need to solve a large decision tree using backward induction. Figure 5.5 demonstrates the form of the tree. The dashed steps indicate that there are many branches, dependent on the number of decisions, outcomes and sampling designs. Also, for this set up, at the blue decision node we assume that after implementing a candidate design we make a decision. In reality we could carry on with any number of stages of sampling until we have enough information to make a "reliable" decision. The backward induction step required to solve this tree is computationally very expensive. Therefore, we will describe a simulation based approach to the full calculation. We use our updated beliefs, which were obtained using the initial dataset, to draw repeatedly from the posterior predictive distribution for the candidate sampling locations. We update our beliefs with each set of simulated values as if they were actual observations, and consider the resulting optimal decisions over the inference grid.

To calculate the expected loss at each inference point, for a candidate design δ, we now have two sets of data: y are the observed initial observations, y_δ are the observations for the candidate design. We take n_{sim} draws, $y_\delta^1, y_\delta^2, \ldots, y_\delta^{n_{sim}}$, from the current posterior predictive distribution for y_δ to obtain values $y_\delta^i = (y_{\delta 1}^i, y_{\delta 2}^i, \ldots, y_{\delta n_\delta}^i)$ for each simulation. For each draw i, we calculate, at each location s,

$$\mathcal{M}_{s,i}(y, y_\delta^i) = \min_d \mathrm{E}\left[L(O_s, d)|y, y_\delta^i\right], \quad i = 1 \ldots n_{sim} \tag{5.7}$$

To determine the total expected loss for each inference point s we record the optimal decision and associated loss for each set of simulated observations, and average this loss. For a large number of simulations, this average will approximate the expected loss from observing sample y_δ and then choosing the optimal decision, namely

$$\overline{\mathcal{M}_s}(y, y_\delta) = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \left(\mathcal{M}_{s,i}(y, y_\delta^i)\right) \approx \mathrm{E}\left[\mathcal{M}_s(y, y_\delta) \mid y\right] \tag{5.8}$$

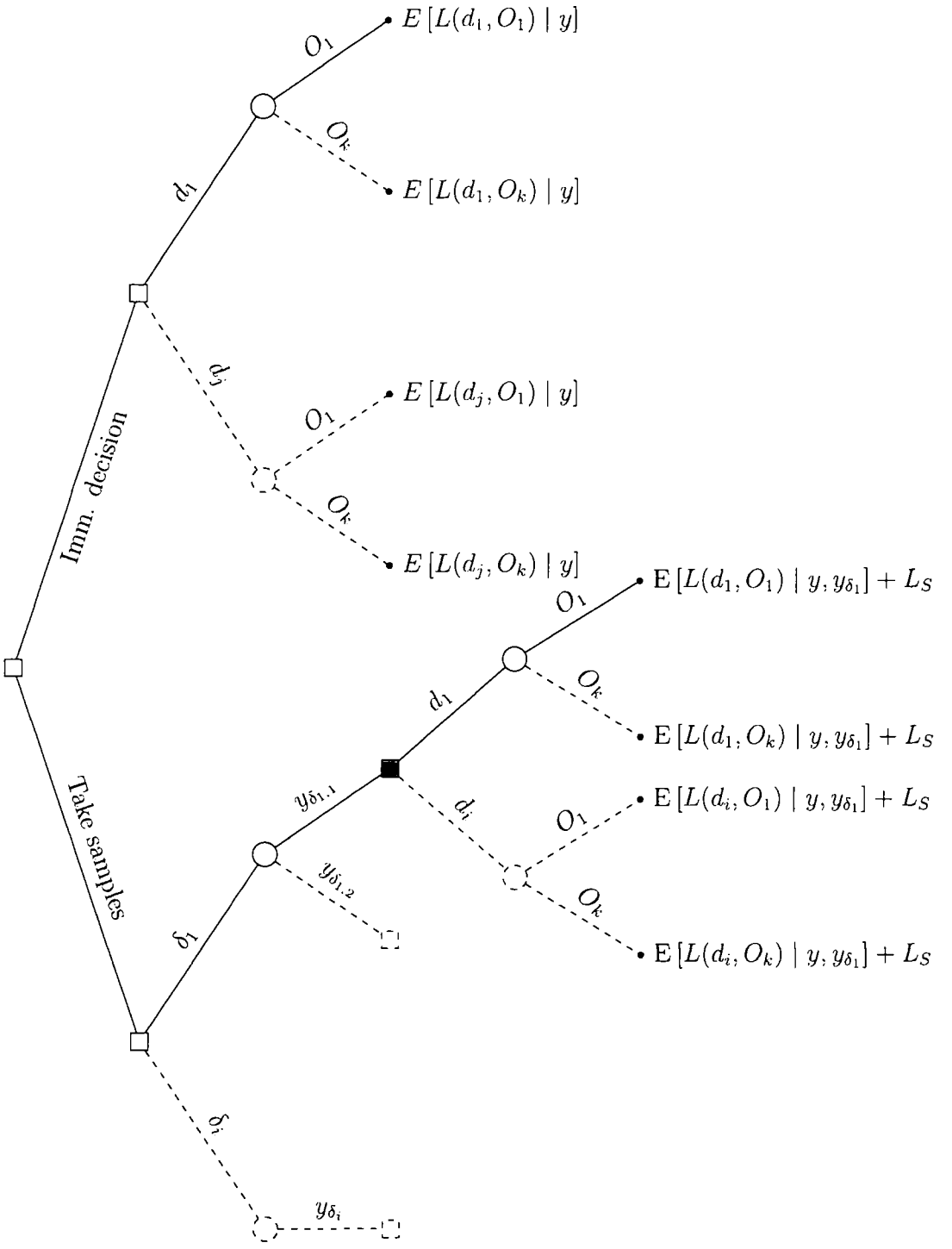


Figure 5.5: Form of the full decision problem, demonstrating the infeasibility of performing backward induction. The solid line shows a single path through the tree. The dotted lines show one of the set of alternative paths, those that do not terminate at end nodes would continue as demonstrated by the relevant full path.

The sample variance of this sum may be used as a guide as to how many simulations are required so that the approximation in Equation (5.8) is good enough for practical purposes. Having done this calculation at each inference point we can sum over the grid to obtain the total expected loss for candidate design δ.

$$C(y, y_\delta) = \sum_{s=1}^{n_I} \overline{\mathcal{M}_s}(y, y_\delta) \approx \sum_{s=1}^{n_I} \mathrm{E}\left[\mathcal{M}_s(y, y_\delta) \mid y\right] \tag{5.9}$$

We can compare $C(y, y_\delta) + L_S$ with the expected cost of making a decision with no further sampling, $C(y)$, to determine whether it will be beneficial to implement the sampling design or not. We can find the sampling variance of $C(y, y_\delta) = \sum \mathrm{Var}(\overline{\mathcal{M}_s}(y, y_\delta)) + 2\sum \mathrm{Cov}(\overline{\mathcal{M}_s}(y, y_\delta), \overline{\mathcal{M}_t}(y, y_\delta))$ directly, and therefore ensure we have taken enough draws from the posterior predictive distribution for y_δ. That is, we judge that our estimate of $C(y, y_\delta)$ is good enough when the value of double the standard error $2\mathrm{SE}(C(y, y_\delta)) = \frac{2\mathrm{SD}(C(y, y_\delta))}{\sqrt{n_{sim}}}$ is much smaller than the estimated difference between $C(y, y_\delta) + L_S$ and $C(y)$ (the expected loss associated with an immediate decision). In particular, we can consider for which zones on the site we are likely to change our decision; or that we may conclude that the difference between $C(y, y_\delta) + L_S$ and $C(y)$ and is sufficiently small that there is no practical difference between the two values.

The expected loss for the candidate design must include the sampling cost L_S (as introduced in Section 5.3.1), dependent on n_δ, the number of locations inspected in the candidate grid:

$$\mathrm{E}\left[L(\delta)\right] = C(y, y_\delta) + \underbrace{L_{init} + n_\delta L_{incr}}_{L_S} \tag{5.10}$$

Let $d^*(y_\delta)$ represent the Bayes decision function over the inference grid, where we choose a decision from the set of all decisions $d \in \mathcal{D}$ having observed the additional sample y_δ. In our hypothetical example, $d^*(y_\delta)$ is a 25 vector for each y_δ, listing the optimal decision at each location for the simulated observations y_δ. The output of the calculation will also give the percentage of times each decision is chosen as a result of the update using simulated values.

Expected loss calculation for design δ - procedure

1. Update beliefs using initial prior specification and original dataset.

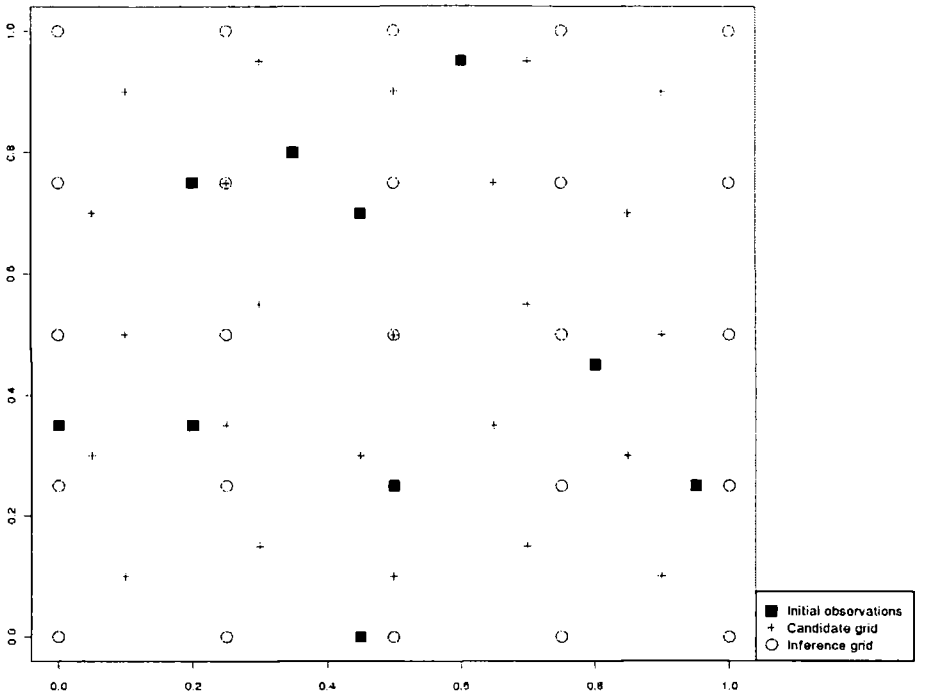


Figure 5.6: Locations used in the calculation of the expected utility of a specific candidate design. See Section 5.5.1 for further explanation

2. Calculate the utility of making a decision now by summing over each point in the "inference grid" as in Section 5.4.

3. Simulate values y_δ^i of the candidate locations, using the current posterior predictive distribution.

4. Use the simulated values, along with the original data to update the original beliefs.

5. Use updated beliefs for simulated values to calculate $\mathcal{M}_{i,s}(y, y_{\delta i})$ and $d^*(y_{\delta i})$ for the i^{th} simulation and for each inference location s

6. Repeat steps 3 to 5 many (n_{sim}) times, and check sample standard deviation to determine when the number of simulations taken is adequate.

7. Evaluate the final expected loss using Equations (5.9) and (5.10).

8. Compare the results of step 2 and 7 to determine whether it would be beneficial to implement this candidate design.

5.5.1 Example

We will now consider the simple set up for the hypothetical example in order to illustrate the methodology. We extend the problem to include three decisions in order to show how the analysis naturally extends to any number of decisions. Table 5.3 shows the loss structure, and gives three outcomes. We have assigned losses in this example on a [0,1] scale, giving a loss of 1 to the worst outcome, and 0 to the best. In this case we have specified a set up whereby:

- Remediation method 1 is more expensive than method 2, but always successful
- Remediation method 2 is successful if the actual contaminant level is below CV2, and cheaper than method 1. However, if the contaminant levels are above CV2 then the method fails and we have to pay the cost of the remediation as well as a failure cost which will be the loss associated with implementing remediation method 1.
- The failure cost is somewhat unrealistic in this example, and tells us that if the actual contaminant level is between CV1 and CV2 then a loss of 0.6 is incurred, and above this level a loss of 1 is incurred (the cost of dealing with the contamination as well as legal/social/health implications). While it is reasonable to suggest that the cost of failure increases with the level of contamination, a discrete cut off as given should be viewed as a simple approximation to the aggregate effect of contamination levels on loss.

	Below CV1	CV1<Value<CV2	Above CV2
Rem method 1	0.5	0.5	**0.5**
Rem method 2	0.25	**0.25**	0.75
Leave	**0**	0.6	1

Table 5.3: Cost tables for the simple decision approach with 3 decisions

The bold entries in Table 5.3 indicate the optimal decision for each potential outcome, showing that each decision is preferable at some level of contamination.

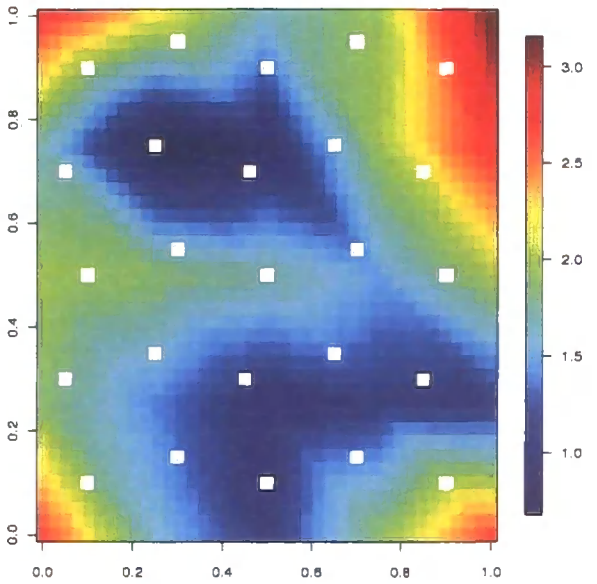


Figure 5.7: Posterior predictive standard deviation following initial update to demonstrate where the areas of high uncertainty are, to guide selection of candidate designs for the next stage of sampling.

Using the same prior information and observation set as in Section 5.4, now with a CV1 of 7 and a CV2 of 10, we calculate the total expected loss when we take no more sampling, and for the 25 point candidate herringbone grid depicted by the crosses in Figure 5.6. As well as the loss values, we require a sampling cost in order to account for the loss incurred by actually implementing the candidate design. We assign $L_{init} = 0$ and $L_{incr} = 0.01$.

As well as the full herringbone grid, we consider four possible candidate grids, each of 5 locations. These were chosen for a number of reasons. The posterior predictive uncertainty was considered, as was the spacing of points, both in relation to each other and to the initial observations. Figure 5.7 shows the posterior predictive uncertainty over the grid following the initial update, with the candidate points overlaid. The four grids compared are shown in Figure 5.8.

In order to consider the amount of information obtained by implementing a "full" expert prior specification with zonal parameters, we will also calculate the expected loss for a "basic" prior structure. That is we will use the prior as specified by the

expert in Section 3.7.3, (Tables 3.15 and 3.17), which gives a full zonal specification and a detailed weighted correlation structure. We will also perform the calculation using a fixed value of $\underline{\theta}$ for the whole site, and only the specification of $a, d, \mathbf{m}$ and $\mathbf{V}$, rather than mean and variance values that vary over the site. However, the basic structure will allow for a linear trend on the mean. Finally, as we obtain considerable information from the ten initial observations, we shall also perform the calculations for a subset of these values, the four locations shown in Figure 5.9.

Table 5.4 details the options used for each case, and gives the expected loss for an immediate decision, along with the expected loss associated with the candidate design. We also see the number of simulations which were "required" for each design. This is the number of simulations that ran before the sampler was deemed "good enough", as determined by comparing the value of $2\text{SE}C(y, y_\delta)$ with the difference between the immediate decision loss and the expected value of the candidate design. Figure 5.11 shows a plot of the standard error against the number of simulations for each of the cases in Table 5.4. The code ran for up to 300 simulations, with a minimum of 50 before the stopping criterion was considered. This choice is for demonstration purposes, in practice more simulations would be run so that $2\text{SE}(C(y, y_\delta))$ is not just below the difference we are comparing with but can further decrease to a minimum, and improve the inference and decision making process. Table 5.4 shows that in most cases approximately 50 simulations were sufficient. However, in cases 4 and 9 the calculation used all 300 simulations, and was still not "good enough". The calculated values for each case show a difference of 1^{-4} and 0.05 respectively between an immediate decision and implementation of the design, meaning the standard error would effectively have to reach 0 to decide between the two alternatives with confidence. Running case 4 for 1000 simulations still gave a standard error of just over 0.005. However, with a difference of 1^{-4} in the expected values, it is likely that that particular sampling design would not be deemed worthwhile and so it would not be neccessary to make a large number of simulations.

The results of Table 5.4 tell us several things:

- For cases 11-15, with the subset of four initial observations as shown in Fig-

Case	Type of prior	Init obs	Cand grid	Imm dec loss	Value of Cand design	Improvement	Num sims "required"
Case 1	Full	All	Full her	7.36	6.65	0.71	51
Case 2	Full	All	Grid 1	7.36	7.01	0.35	51
Case 3	Full	All	Grid 2	7.36	7.19	0.17	51
Case 4	Full	All	Grid 3	7.36	7.36	0.00	300
Case 5	Full	All	Grid 4	7.36	6.80	0.56	51
Case 6	Basic	All	Full her	7.88	7.39	0.49	51
Case 7	Basic	All	Grid 1	7.88	7.56	0.32	51
Case 8	Basic	All	Grid 2	7.88	7.59	0.29	51
Case 9	Basic	All	Grid 3	7.88	7.83	0.05	300
Case 10	Basic	All	Grid 4	7.88	7.45	0.43	51
Case 11	Basic	Subset	Full her	10.11	8.81	1.30	51
Case 12	Basic	Subset	Grid 1	10.11	9.34	0.77	51
Case 13	Basic	Subset	Grid 2	10.11	9.05	1.06	51
Case 14	Basic	Subset	Grid 3	10.11	9.58	0.53	51
Case 15	Basic	Subset	Grid 4	10.11	8.96	1.15	51
Case 16	Full	Subset	Full her	6.91	6.49	0.42	51
Case 17	Full	Subset	Grid 1	6.91	6.63	0.28	51
Case 18	Full	Subset	Grid 2	6.91	6.71	0.20	51
Case 19	Full	Subset	Grid 3	6.91	6.63	0.28	51
Case 20	Full	Subset	Grid 4	6.91	6.49	0.42	51

Table 5.4: Expected losses for immediate decision and candidate designs in a number of alternative set ups

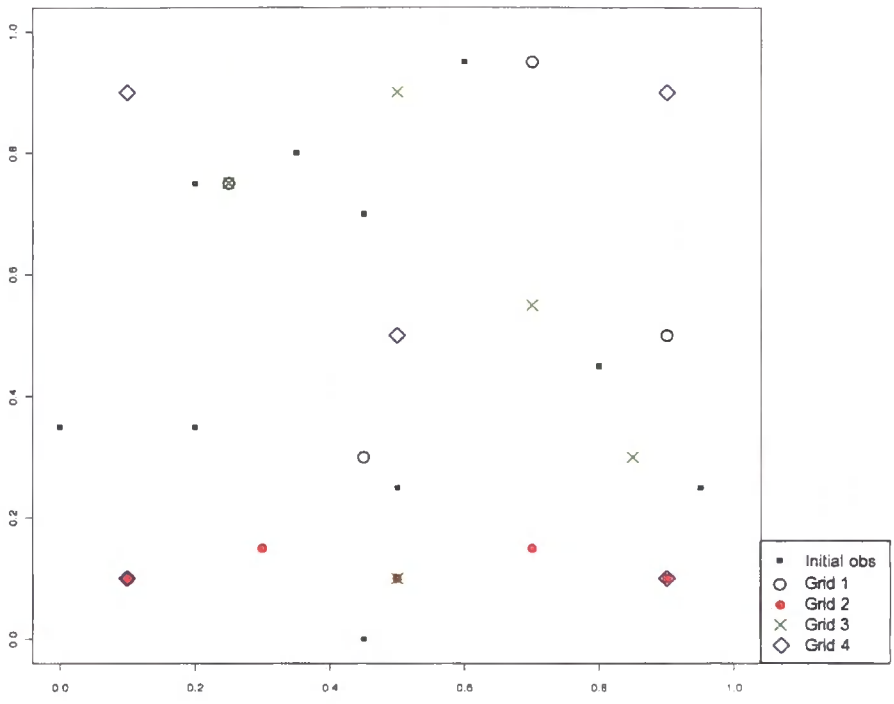


Figure 5.8: The locations of the four candidate grids as used in Example 5.5.1

ure 5.9, we see a larger improvement than for the equivalent cases with the full initial observation set. This reflects the fact that we have less information to begin with when we use the basic prior, and so we can learn more by taking samples. There will come a point at which we cannot obtain any extra information by continuing to add sample points.

- With all the initial observations included, grid 3 offers a very small improvement on the immediate decision loss compared with the alternative candidates. This is likely to be due to the fact that the five candidate points are in the blue, low uncertainty zone (as shown in Figure 5.7), and so we cannot learn a great deal more by sampling in these zones.

- When the points are spread out from each other and from the initial observations, and are in higher uncertainty zones (such as in grid 4), we see a better improvement.

- The full design beats or matches the other four grids for each set up, showing that there is still a considerable amount of information available in the site.

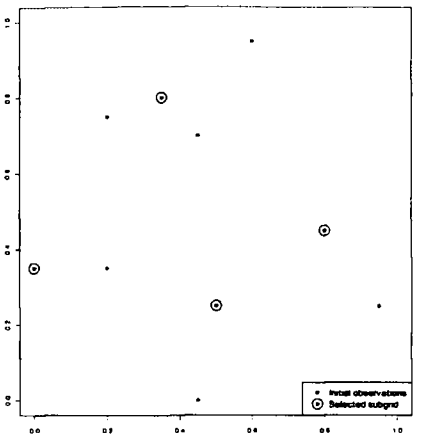


Figure 5.9: Locations of the initial dataset used for the example, with fewer observations example circled in green

However, sampling is quite cheap, if we were to increase the sampling cost then the four 5 point grids would become a better option. For example, if the sampling cost increased from 0.01 to 0.03 per sample in cases 1-5, then the improvement for case 1 is reduced to 0.21, and the improvement for case 2 is reduced to 0.25, thus becoming preferable.

We can report the proportion of times we change the decision from that of the immediate decision calculation as a result of the simulated values in order to assess where the decision is changing most frequently. This is depicted by Figure 5.10, for Case 1 of Table 5.4. We see that for 10 of the 25 inference locations, we have a decision which is frequently changing as a result of the simulated values we "observe". This shows the zones where we are most uncertain about the contamination status, intuitively these points are all on the boundary between "decision areas". We also give a breakdown of the decisions taken, in table form in the computer output, to allow for a detailed depiction of the results. For the DM to judge the actual benefit, these standardised loss values can be translated back into monetary terms (we shall consider this for the real case study of Chapter 7, when we have real figures).

If factors such as the correlation lengths, or prior beliefs were altered, then the expected losses calculated would be affected, showing that the careful consideration of model parameters is not only important at the predictive stage, but is carried

Location	Imm dec	Prop d_1	Prop d_2	Prop d_3
1	d_3	0	0.059	0.941
2	d_3	0	0	1
3	d_3	0	0	1
4	d_3	0	0	1
5	d_3	0	0	1
6	d_3	0	0.137	0.863
7	d_3	0	0.039	0.961
8	d_3	0	0	1
9	d_3	0	0	1
10	d_3	0	0	1
11	d_2	0.432	0.529	0.039
12	d_2	0.217	0.667	0.118
13	d_2	0.078	0.667	0.255
14	d_3	0.020	0.255	0.725
15	d_3	0	0.235	0.765
16	d_1	1	0	0
17	d_1	1	0	0
18	d_1	1	0	0
19	d_1	0.412	0.588	0
20	d_2	0.216	0.510	0.274
21	d_1	1	0	0
22	d_1	1	0	0
23	d_1	1	0	0
24	d_2	0.353	0.647	0
25	d_2	0.314	0.490	0.196

Table 5.5: Immediate decision, compared with the proportion of time each decision was made in the simulation calculation for Case 1

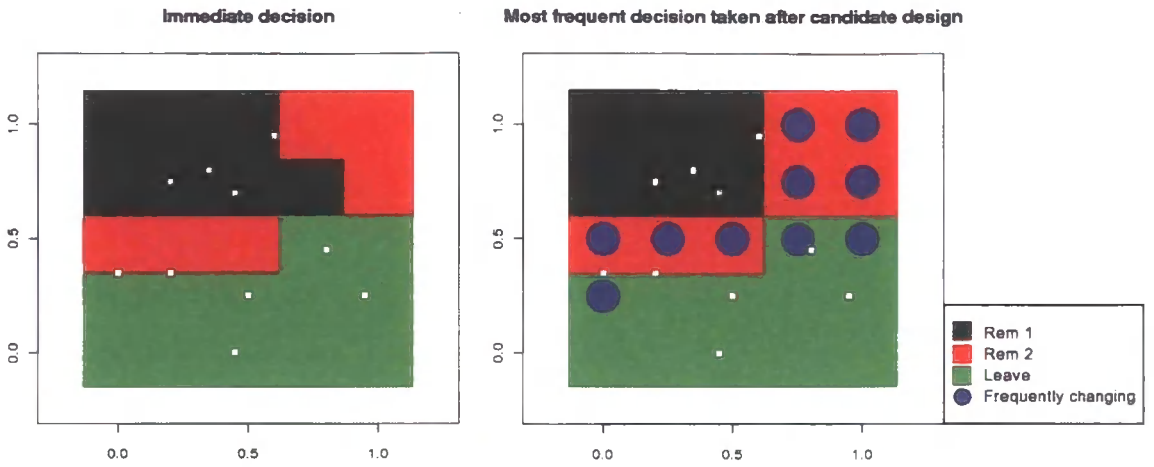


Figure 5.10: Depiction of the decisions taken for Case 1. Left image shows the decision chosen at each inference point for an immediate decision. The image on the right shows the most frequently taken decision when the candidate design is implemented. The blue locations represent those that take an alternative decision to the most frequently selected for more than 10% of the simulations

through to the decision and sample selection stages. Also if the cost of sampling were to increase, it would quickly become less beneficial to take more samples. For example, if we double the incremental cost of a sample in terms of our loss set up from 0.01 to 0.02, then for the 25 point herringbone grid the expected loss for the design increases by 0.25, and then the loss associated with taking an immediate decision is smaller for many of the cases in Table 5.4. This would lead to fewer samples being taken, as shall be demonstrated in the next chapter.

Possible approximate evaulation of expected loss

A fast alternative to the simulation approach is to use the current posterior predictive mean values only. That is, we update our beliefs with the initial data, and then use the posterior predictive mean for each candidate location as if it were the actual observation. This does not allow for extreme values which may arise when taking repeated draws from the posterior, but it will give an indication of the locations where uncertainty may be reduced, perhaps in order to select a small set of candidate designs for a full simulation calculation from a much larger original

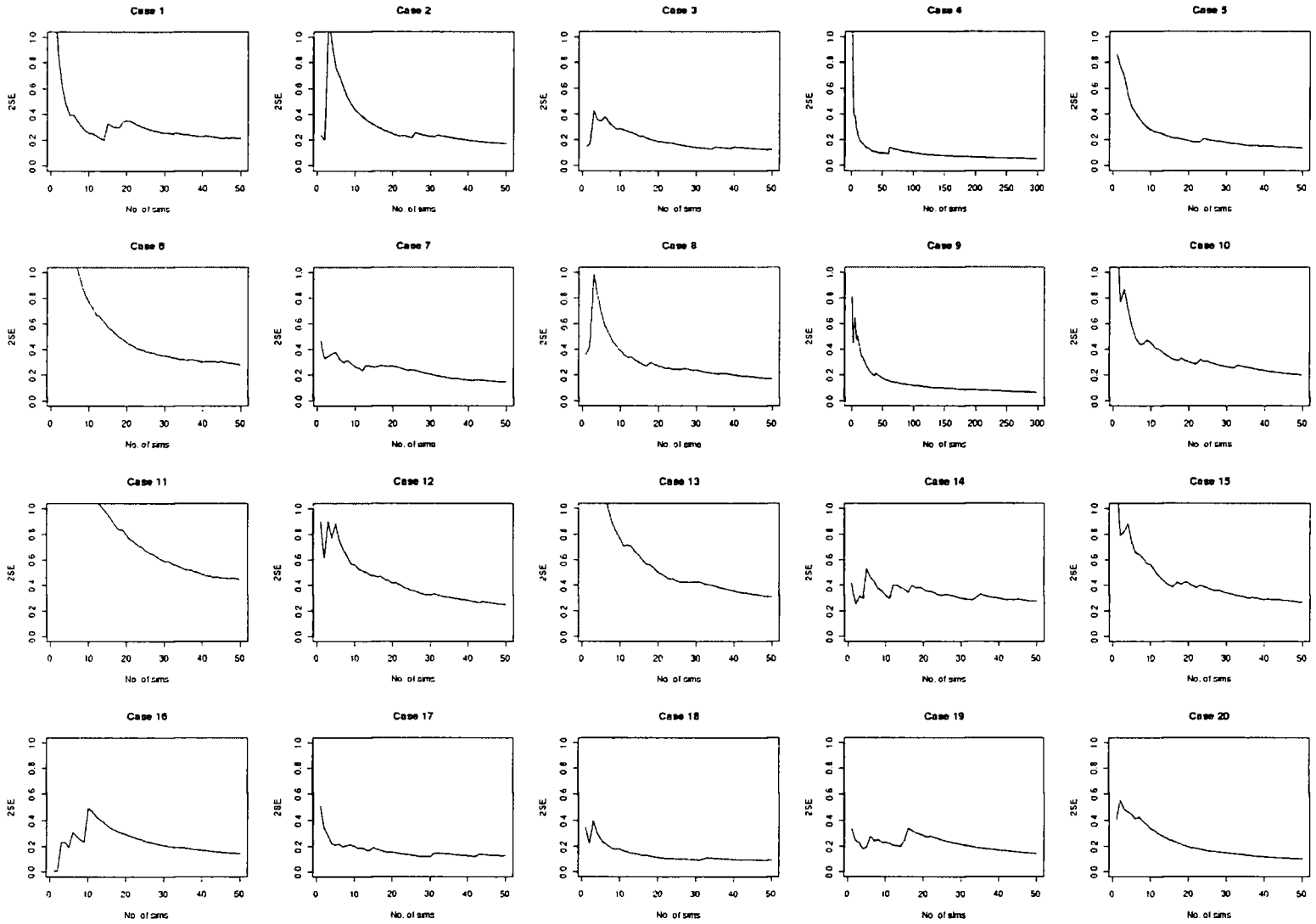


Figure 5.11: Value of twice the sample standard error for the expected loss calculation for each case covered in Example 5.5.1

Case	Cand des value, full method	Approx method value
1	6.65	5.67
2	7.01	6.48
3	7.19	6.66
4	7.36	6.79
5	6.80	6.41
6	7.39	5.52
7	7.56	6.71
8	7.59	6.71
9	7.83	6.95
10	7.45	6.52
11	8.81	6.23
12	9.34	7.93
13	9.05	7.42
14	9.58	7.88
15	8.96	7.37
16	6.49	6.00
17	6.63	6.14
18	6.71	6.24
19	6.63	6.16
20	6.49	6.09

Table 5.6: Candidate design value using the approximate method

number of candidates. Table 5.6 gives the results for the approximate calculation, in comparison to the values obtained in Table 5.4 using the full simulation method. While the approximation gives consistently lower values, the ordering of designs by lowest expected loss remains much the same, and could be used as a quick screening to remove poor candidates. However, this should be done with caution as it does not fully take into account the uncertainty levels at each location. The method is much faster than the full expected loss calculation as no simulations are required, it is simply assumed that the predictive mean at a location is the actual observed

value.

5.6 Expanding the basic set up

Now that we have considered a simple set up for any number of decisions, we expand the methodology to include more realistic loss functions. We also require an approach to deal with decision making when multiple contaminants are involved, whether they be independent or related.

5.6.1 More complex loss functions

If the expert would prefer to describe the loss as a function of the contaminant level, the problem becomes more complicated. We can no longer find the probability of a particular outcome and multiply it by the associated loss. Rather we are in a continuous situation where we will have to integrate over the range of contamination in order to find the expected loss associated with each decision d_i:

$$\mathrm{E}\left[L(O, d_i) \mid y\right] = \int_0^\infty p(y(\underline{x}^p) \mid y) L(y(\underline{x}^p), d_i) \mathrm{d}y(\underline{x}^p) \tag{5.11}$$

We can approximate this integral, using Monte Carlo integration methods [50]. We now simulate from the current posterior predictive distribution for the inference grid given the sample values on the candidate grid, as we require an estimation of what will occur at the inference points given the update using the candidate "observations".

The overall calculation of the expected loss for a design will follow the same procedure outlined in Section 5.5. However, at step 5 the calculation will require an approximation to determine the expected loss for decisions when the loss is a function of the contamination. The steps required for the approximation are:

1. Take K draws from the posterior predictive distribution for the inference grid

2. For each inference point, select only the k draws which lie in the boundaries of the integral we are approximating, i.e. between CV1 and CV2 for the integral shown below (Equation (5.12)).

3. For these draws, calculate the value of the loss function for each of the k values $L(y(\underline{x}^p)_k, d_i)$.

4. Sum these k values and divide by the total number of simulations K to obtain an estimator of the integral.

$$\int_{CV1}^{CV2} p(y(\underline{x}^p) \mid y) L(y(\underline{x}^p), d_i) \mathrm{d}y(\underline{x}^p) \approx \frac{1}{K} \sum_k L(y(\underline{x}^p)_k, d_i) \tag{5.12}$$

This procedure will very quickly become time consuming, as for each set of simulations for the candidate grid, we must draw a number of simulations from the inference grid in order to approximate these integrals. We can consider the efficiency of the method by monitoring the behaviour of the expected value and standard deviation of the approximation in order to ascertain when the approximation is "good enough". We further explain the method with the use of an example in the next subsection.

5.6.2 Example for multiattribute set up

If we wish to consider many factors for the decision analysis, then we can use a multiattribute approach as introduced in Section 5.2.2. For the hypothetical example, we shall look at a combination of two utility attributes, cost and human health impact; and three decisions, ex-situ (d_1), in-situ (d_2), or no action (d_3). There are two critical values in this case, as in the example of Section 5.5.1, although, as we are not using fixed losses, we require a third value, which we call the upper bound, UB. All loss functions must be bounded which is why we require the upper bound value to be specified as well as the relevant critical values. For the cost case we assume a discrete set up whereby there is a particular cost dependent on the classification of the contamination as low, medium or high. The highest loss value for the cost attribute is assigned to d_2, when the actual contamination level is above CV2. This is a worse scenario than if we take no action and the $y(\underline{x}^p)$ is above CV2 because we have to pay for two sets of remediation (the initial, unsuccessful method, and the alternative method to deal with the actual levels of contamination). The loss function for human health is given as a continuous function of the contamination level, this is why we require the approximation method described in the previous section.

	Cost	Human Health
d_1	0.6	0
d_2	$\begin{cases} 0.4 & \text{if } y(\underline{x}^p) < CV2 \\ 1 & \text{otherwise} \end{cases}$	$\begin{cases} 0 & \text{if } y(\underline{x}^p) < CV2 \\ 0.5\frac{y(\underline{x}^p)-CV2}{UB-CV2} & \text{if } CV2 < y(\underline{x}^p) < UB \\ 0.5 & \text{otherwise} \end{cases}$
d_3	$\begin{cases} 0 & \text{if } y(\underline{x}^p) < CV1 \\ 0.4 & \text{if } CV1 < y(\underline{x}^p) < CV2 \\ 0.6 & \text{otherwise} \end{cases}$	$\begin{cases} \frac{y(\underline{x}^p)}{UB} & \text{if } y(\underline{x}^p) < CV2 \\ 1 & \text{otherwise} \end{cases}$

Table 5.7: Loss table for three decision, two attribute set up

We have specified these partial loss functions on a [0,1] scale in order to ensure that we can combine them in a sensible way. We require a weighting on the importance of the two factors. The weights we assign for this example are cost $\frac{2}{3}$, and health $\frac{1}{3}$. These weights tell us that the DM considers the cost factor to be twice as important as health. This is likely to be the view of a site investigator who is keen to save as much money as possible, but at offers a compromise to the interests of the end users of the site and the EA by including a non-financial factor into the analysis. These loss values are assigned by giving a value of 1 to the worst consequence, 0 to the best, and scaling the values in-between to be representative of the desirability of the outcome.

Figure 5.12 shows the overall loss function when we combine the information given in Table 5.7 with the weights $\frac{2}{3}$ and $\frac{1}{3}$. We see that each decision is the optimal choice at some value of contamination. We could combine the loss attributes in any number of different ways. For example, Figure 5.13 shows how the loss function changes as the importance shifts from cost to health. We perform the calculation for the same 20 cases used in the example of Section 5.5.1 (using 500 simulations for the integration approximation), obtaining the results shown in Table 5.8.

We can also run the analysis for the alternative weights as shown in Figure 5.13, the results are shown in Table 5.9 for case 1. This shows the importance of carefully considering the assignment of loss functions, as the values obtained are very varied. The result for the second option, where only the health loss is taken into account,

Case	Type of prior	Init obs	Cand grid	Imm dec loss	Value od Cand design	Improvement	Num sims "required"
Case 1	Full	All	Full her	5.69	5.53	0.16	51
Case 2	Full	All	Grid 1	5.69	5.67	0.02	300
Case 3	Full	All	Grid 2	5.69	5.75	-0.06	51
Case 4	Full	All	Grid 3	5.69	5.78	-0.09	51
Case 5	Full	All	Grid 4	5.69	5.59	0.10	51
Case 6	Basic	All	Full her	5.54	5.51	0.03	226
Case 7	Basic	All	Grid 1	5.54	5.51	0.03	300
Case 8	Basic	All	Grid 2	5.54	5.62	-0.08	186
Case 9	Basic	All	Grid 3	5.54	5.56	-0.02	258
Case 10	Basic	All	Grid 4	5.54	5.42	0.12	105
Case 11	Basic	Subset	Full her	6.29	6.13	0.09	300
Case 12	Basic	Subset	Grid 1	6.29	6.02	0.27	51
Case 13	Basic	Subset	Grid 2	6.29	6.35	-0.06	300
Case 14	Basic	Subset	Grid 3	6.29	6.28	0.01	300
Case 15	Basic	Subset	Grid 4	6.29	6.20	0.09	65
Case 16	Full	Subset	Full her	5.63	5.61	0.02	59
Case 17	Full	Subset	Grid 1	5.63	5.57	0.06	53
Case 18	Full	Subset	Grid 2	5.63	5.64	-0.01	300
Case 19	Full	Subset	Grid 3	5.63	5.60	0.03	63
Case 20	Full	Subset	Grid 4	5.63	5.55	0.08	51

Table 5.8: Expected losses for immediate decision and candidate designs in a number of alternative set ups for the multiattribute loss set up

Cost weight	Health weight	Immediate dec.	Cand design
1	0	6.65	6.81
0	1	0	0.25
0.25	0.75	2.76	2.82
0.85	0.15	6.290	6.49

Table 5.9: Expected loss values for case 1, using different weightings of the two loss attributes

is intuitive and would not require a decision analysis. If there is no cost element we would always make the decision that gave no detrimental health impact regardless of the contamination levels.

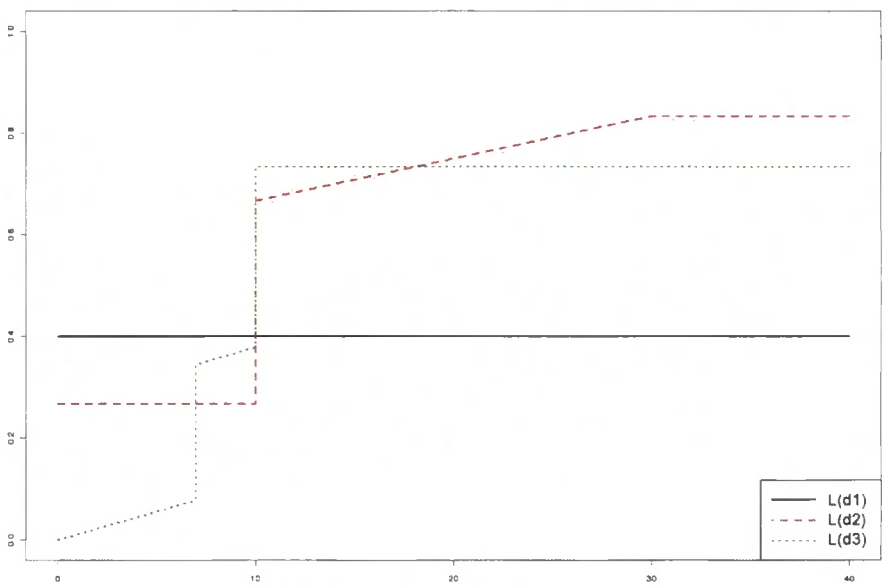


Figure 5.12: Combined loss functions for each of the three decision alternatives.

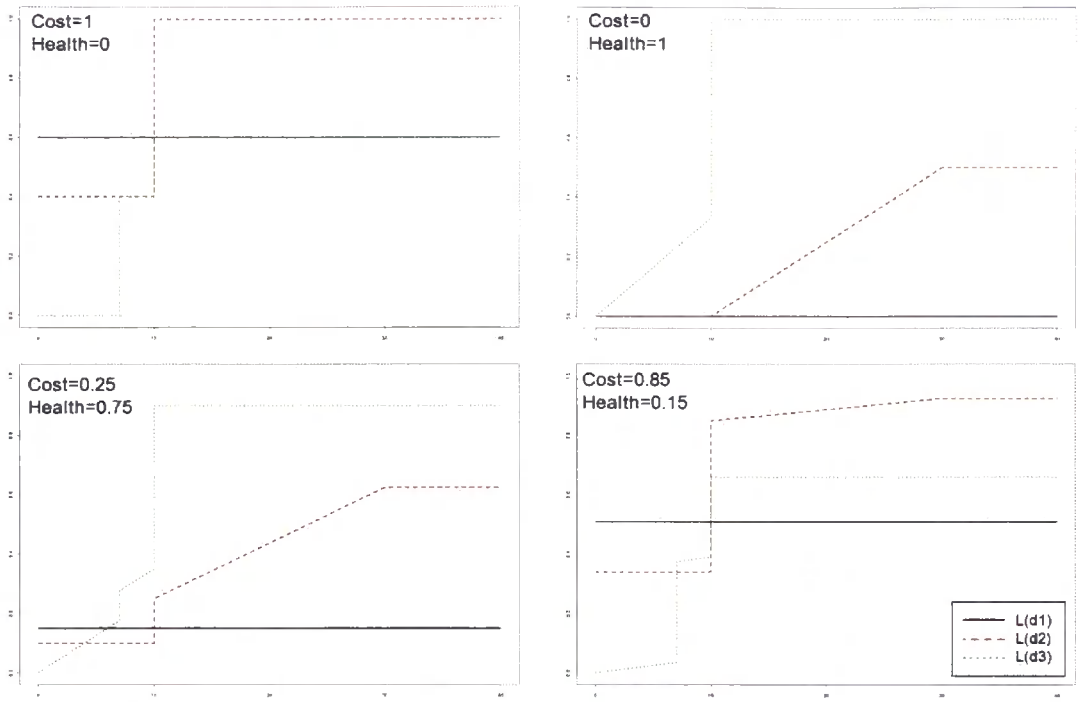


Figure 5.13: Effect of altering the weightings for the overall loss function

5.7 Decisions when multiple contaminants are involved

We have covered a number of loss set-ups when making remediation decisions for a single contaminant, but as in Chapter 3, it is likely that we will have more than one contaminant of interest on a site. We now consider an approach for the evaluation of the expected loss when multiple contaminants are to be considered.

5.7.1 Simple set up with numerical example

Initially, we will briefly consider the simplest possible set up with a numerical example. The simplest multiple contaminant decision is the two contaminant, two decision case. As in the set up of Section 5.3.2 we shall assume that each contaminant can be deemed "clean" or "contaminated" at each inference location, and that at each point we can choose to remediate or take no action. As we have two contaminants, and a discrete loss set up, we have a loss table as shown in Table 5.10.

A represents contaminant A exceeding the critical value (contaminated), and $\overline{A}$ denotes that contaminant A is below the critical value (clean). Figure 5.14 gives a depiction of the possible consequences, with losses and probabilities attached. We see from the decision tree how quickly branches will be added, increasing the computational load as decisions or contaminants are added. This demonstrates why a discrete approach will become quickly complicated. Even if we assume only two possible outcomes, contaminated or clean, then we must consider 2^n consequences for each decision, when we have n contaminants. For now, we will assume independence of contaminants, which allows us to calculate probabilities of outcomes as the product of the relevant probabilities for each contaminant.

	$A \cap C$	$A \cap \overline{C}$	$\overline{A} \cap C$	$\overline{A} \cap \overline{C}$
Rem	10	10	10	10
Leave	45	30	15	0

Table 5.10: Simple two contaminant, two decision loss table

The two contaminants we will use for this example are A (dataset A2) and C, as used in the modelling examples of Chapter 3. We shall use the prior specifications of Section 3.7.3.

We choose to assign a single remediation cost for this example, so remediation is always successful and doesn't depend on the actual state of the site and how contaminated it is. The choice of failure costs will be dependent on the particular contaminants and the views of the expert. In this case we assign a failure cost of 45 if both contaminants are above the critical value and we do not remediate. This is the sum of the individual failure costs (30 and 15), and the validity of such a set up which will have to be considered in some detail by the expert. The applicability of an additive loss criterion will depend on the cumulative effects of multiple failures. This will be determined both by the expert and the available guidance.

In this set up we assume that contaminant A has higher adverse effects at levels over its critical value than contaminant C does, and therefore a greater failure cost. This is a complex issue and would be determined subjectively by the expert based on the scientific information available for each contaminant of interest. We will consider

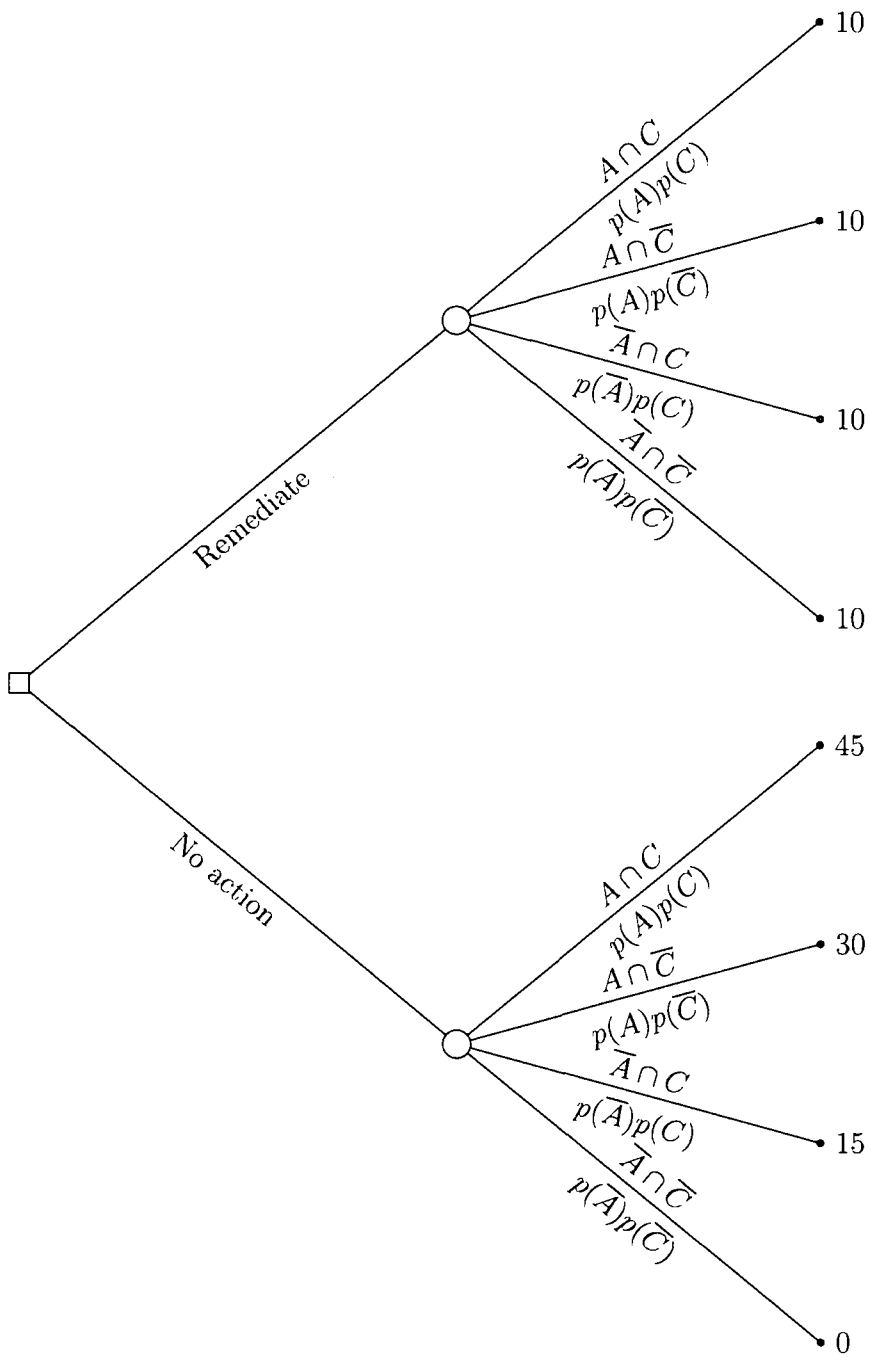


Figure 5.14: Decision tree for simple two contaminant problem, with costs and probabilities attached

a more detailed cost analysis when dealing with the real set up of Chapter 7.

We now perform the same steps as outlined in Section 5.5, but now with four outcomes to consider. Figure 5.15 shows the results of the calculation, for contaminant A only, contaminant C only, and the multiple decision calculation. We use the 25 point regular inference grid, and the 25 point herringbone grid as the candidate design. We see that the multiple contaminant calculation has picked the locations where we expect remediation to be necessary for either contaminant, or both. This is an intuitive outcome with an equally weighted loss set up. We will consider loss specifications which lead to different decisions for the three contaminant example of the next section. Table 5.11 gives the expected losses, for both the immediate decision, and the expected loss from implementing the 25 point candidate grid. The third row of the table shows the effect obtained from running the multiple decision analysis. For comparison, the fourth row shows the effect of performing the calculations individually and then combining the results, where we have taken into account the costs of remediation which are duplicated. That is, for the immediate decision value, the fourth row is the sum of row one and two, minus the cost of remediating the seven locations for C which are also remediated for contaminant A. The blue circles again show the zones where we changed our decision for more than 10% of the simulations.

Essentially the multiple contaminant set up is no more complicated than for a single contaminant, just more computationally intensive as we have to calculate more probabilities and losses, and combine them. The difficult part of a multiple contaminant analysis is at the modelling stage, and in Chapter 3 we found that in most cases it is computationally faster to perform individual single updates and combine the results. If we prefer to use the joint update and run the MCMC sampler, we can use the samples to calculate the losses and simulate for the candidate locations, and then use the closed update as a quick way to evaluate the value of the design. We wish to retain as much computational efficiency as possible in order to quickly search between designs in the next chapter.

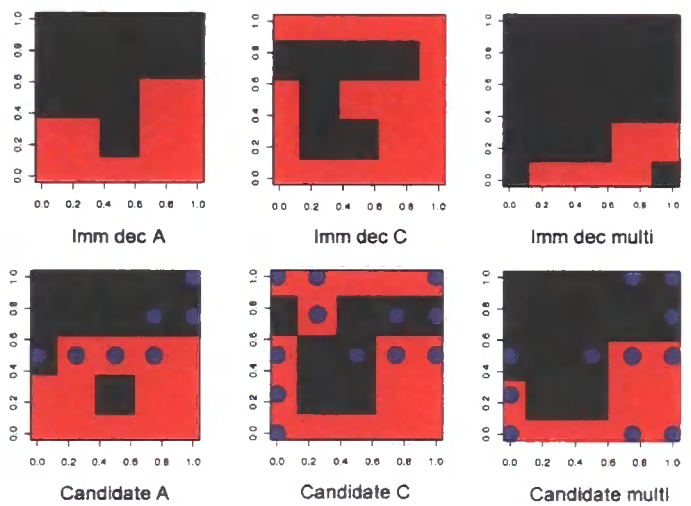


Figure 5.15: Results of the multiple contaminant expected loss calculation. The two left figures show the decision taken when we analyse for contaminant A only. The middle figures show contaminant C only, and the right figures show the result for the multiple calculation for both contaminants. The black area is where the optimal decision is remediation, and red is to leave. The bottom three figures display the decision that was selected most frequently at each location as a result of the simulations. The blue dots show those locations where a different decision to the displayed colour was taken for more than 10% of the simulations.

Case	Imm dec	Cand design
A	172.68	126.73
C	162.11	129.32
Multi	208.38	176.65
A∩C	274.79	209.05

Table 5.11: Expected loss values for the three calculations depicted in Figure 5.15. The A∩C values sum the expected loss obtained using the individual calculations for contaminant A and C, where we subtract the cost of remediation which is duplicated. This demonstrates the difference between a multiple contaminant decision calculation, and the summing of individual decision analyses.

5.7.2 Extension to more complicated loss functions

The discrete set up of the previous section will quickly become complicated as we increase the number of contaminants and decisions, and so we always prefer to work with continuous loss functions. While the use of continuous loss functions requires some more thought from the expert, it allows a much richer description of the consequences and so is preferred to a discrete set-up. We use a three contaminant set-up, as this can be extended to any number of contaminants as necessary. We will assume the three contaminants we consider are all metals or all organics, and so we may consider the cumulative effect of high contamination levels. We will consider a set up with three decisions. In this case we assume an additive, possibly weighted approach. In effect we are working in a similar way to the multiattribute approach, where the contaminants form the attributes of loss. An alternative method is maximum loss, whereby rather than adding the expected losses due to failure we select the largest of the three losses per location. We demonstrate both of these methods in an example.

Rather than using the approach of Section 5.7.1, we will consider an alternative method of building up the loss functions for each contaminant. As in Section 5.6 we use continuous loss functions, now for each contaminant. We have two relevant cut off points per contaminant, the critical value and the upper bound.

When we treat the contaminants as independent, as we will in order to obtain

probability calculations with computational ease, the evaluation of $\mathcal{M}_s(y)$ now involves the following, shown for a three contaminant, three decision example:

$$\begin{aligned}
\mathcal{M}_s(y) &= \min\left[E\left[L(A,B,C,d_1) \mid y\right], \right. \\
&\quad E\left[L(A,B,C,d_2) \mid y\right], \\
&\quad \left. E\left[L(A,B,C,d_3) \mid y\right]\right] \\
E\left[L(A,B,C,d_i) \mid y\right] &= w_A \int L(A,d_i)p(A \mid y)\mathrm{d}A \\
&+ w_B \int L(B,d_i)p(B \mid y)\mathrm{d}B \\
&+ w_C \int L(C,d_i)p(C \mid y)\mathrm{d}C \\
&+ L(d_i) \qquad (5.13)
\end{aligned}$$

where $L(d_i)$ is the fixed cost of decision i, i.e. the remediation cost, and $L(A,d_i)$ is the loss associated with failure for contaminant A when decision d_i is taken. w_j is a weight which may give higher influence to a particular contaminant. This weight specification may not be neccessary if the relative importance of each contaminant is described within the individual specification of loss functions for each contaminant, which is why working on a standardised loss scale is preferable, to allow comparibility.

Within this set up, we may still have loss specifications which have fixed values, or a combination of fixed values and functions. Table 5.12 gives an example of the form of the loss specifications required for a three contaminant, three decision set up.

	A	B	C	Fixed cost of decision
d_1	$F_{A1}(A)$	$F_{B1}(B)$	$F_{C1}(C)$	$L(d_1)$
d_2	$F_{A2}(A)$	$F_{B2}(B)$	$F_{C2}(C)$	$L(d_2)$
d_3	$F_{A3}(A)$	$F_{B3}(B)$	$F_{C3}(C)$	$L(d_3)$

Table 5.12: Form of the loss table for multiple contaminant case with continuous failure loss functions and an overall implementation cost per decision

Example

We now carry out a numerical example for the extended multiple contaminant decision analysis. We use the three hypothetical datasets as analysed in the examples of the modelling chapter, Sections 3.6 and 3.7.3. We have three separate loss specifications. We use the loss set up of Table 5.13, with critical values $CV_A = 10, UB_A = 20$, $CV_B = 150, UB_B = 250$ and $CV_C = 3, UB_C = 16$. The specification of the loss functions in the multiple case are particularly important here. We must consider the combination of the loss functions and weightings carefully. The candidate design is the full 25 point herringbone grid, and the inference grid is the 25 point regular grid.

	L(A)	L(B)	L(C)	Fixed cost of decision
d_1	0	0	0	0.5
d_2	$F_{A2}(A)$	$F_{B2}(B)$	$F_{C2}(C)$	0.25
d_3	$F_{A3}(A)$	$F_{B3}(B)$	$F_{C3}(C)$	0

Table 5.13: Three contaminant, three decision loss set up, with failure costs per contaminant and the overall loss associated with implementing the decision

The loss functions are given in Equations (5.14-5.19). For contaminant A we are assuming that both remediation methods 1 and 2 are successful for levels below the upper bound. For all three contaminants we give a failure cost of 0 for remediation method 1, which tells us that it is successful at any level of contamination, even above the upper bounds. The remaining loss functions incorporate a loss which increases with "actual" contamination levels, until the upper bound is surpassed, when a fixed value is assigned. The upper bound represents the point at which the contamination is deemed to be equally harmful at levels in excess of the UB. The failure costs also involve a cost of having to remediate when the wrong decision was

taken.

$$F_{A2}(A) = \begin{cases} 0 & \text{if } (A < 10) \\ 0 & \text{if } (10 < A < 20) \\ 0.75 & \text{otherwise} \end{cases} \tag{5.14}$$

$$F_{A3}(A) = \begin{cases} 0 & \text{if } (A < 10) \\ \frac{A}{40} + 0.25 & \text{if } (10 < A < 20) \\ 0.75 & \text{otherwise} \end{cases} \tag{5.15}$$

$$F_{B2}(B) = \begin{cases} 0 & \text{if } (B < 150) \\ \frac{B}{1000} & \text{if } (150 < B < 250) \\ 0.75 & \text{otherwise} \end{cases} \tag{5.16}$$

$$F_{B3}(B) = \begin{cases} 0 & \text{if } (B < 150) \\ \frac{B}{1000} + 0.25 & \text{if } (150 < B < 250) \\ 0.75 & \text{otherwise} \end{cases} \tag{5.17}$$

$$F_{C2}(C) = \begin{cases} 0 & \text{if } (C < 3) \\ \frac{C}{64} & \text{if } (3 < C < 16) \\ 0.75 & \text{otherwise} \end{cases} \tag{5.18}$$

$$F_{C3}(C) = \begin{cases} 0 & \text{if } (C < 3) \\ \frac{C}{64} + 0.25 & \text{if } (3 < C < 16) \\ 0.75 & \text{otherwise} \end{cases} \tag{5.19}$$

	Imm dec	Candidate
A	2.769	2.476
B	1.446	1.740
C	6	5.217
Multiple add	5.873	4.704
Multiple max	6.578	6.270

Table 5.14: Results for the multiple example

We calculate the expected loss using the same method as described by Method ??, and expanded upon in Section 5.5, where we now construct the loss elements in a

slightly different way. We individually calculate at the expected loss for each contaminant, as in Equation (5.13), and then perform one of two calculations. If we have decided on an additive method, then each expected loss will be the sum of the individuals, as shown in Equation (5.13). Alternatively if the maximum loss criterion is used we have

$$\begin{aligned} E\left[L(A,B,C,d_i) \mid y\right] &= \max\Big\{\int L(A,d_i)p(A \mid y)\mathrm{d}A, \\ &\quad \int L(B,d_i)p(B \mid y)\mathrm{d}B, \\ &\quad \int L(C,d_i)p(C \mid y)\mathrm{d}C\Big\} + L(d_i) \end{aligned}$$

The results show that the choice of an additive or maximum loss choice makes a clear difference for the decision analysis. With the additive loss set up we effectively get an averaging of the failure losses and can therefore remediate in fewer locations overall. When we consider the real case study in Chapter 7 we will need to carefully consider the implications of adding the losses, to ensure a poor choice of combined loss function doesn't lead to erroneous decisions being made. For the maximum loss criterion the multiple decision analysis generally takes the worst case scenario at each location. As we are using simulations, this is not the case all the time, and those locations with a blue dot show locations where the most frequently taken decision is chosen less than 90% of the time. We will look at more detailed multiple analyses, considering the effects of altering observation values, locations and loss functions in the next two chapters.

We see from Table 5.14 that the candidate grid gives a reduction in expected loss except for the case of Contaminant B. After a detailed inspection of the proportion of times decisions were taken for this case, for two of the frequently changing points of Figure 5.16 the decision taken as a result of the simulated values is d_2 for half the simulations and d_3 for the other half. This suggests that these points will be good candidates for investigation in the sample search algorithm. We will verify this once we have introduced the sample search methodology in the next chapter. Looking at the multiple results and the locations of frequently changing decisions, we see that

particularly in the additive case we should expect to see sampling points throughout the site being considering in a multiple sample search.

Computational time

We shall consider the computational burden of these decision calculations in the next chapter, as this will allow us to take account of the repeatability of the calculations. When we calculate the expected loss for a single candidate design we have to create several large objects which will be stored and used again and again for the required calculations. So when we consider the time taken for one calculation, it will not be reflective of how long several expected loss calculations will take, as we may be able to reuse some of the created objects. In the next chapter we will look for an algorithm to search over a number of potential designs and select the "best", so we will have to run the calculation once for each candidate design. It will be beneficial at the modelling stage to exclude contaminants which are very unlikely to pose a SPOSH to potential receptors in order to improve the computational efficiency of the decision and sampling stage of the analysis.

Conclusions

We have considered a number of ways in which we can set up a loss structure with which we may analyse the contamination levels and remedial options. Now we have introduced the decision set up, and described a way to calculate the expected utility of a particular candidate sampling design, we require a computationally tractable method of searching between candidate designs in order to select the most efficient and informative sampling locations given the current data and beliefs.

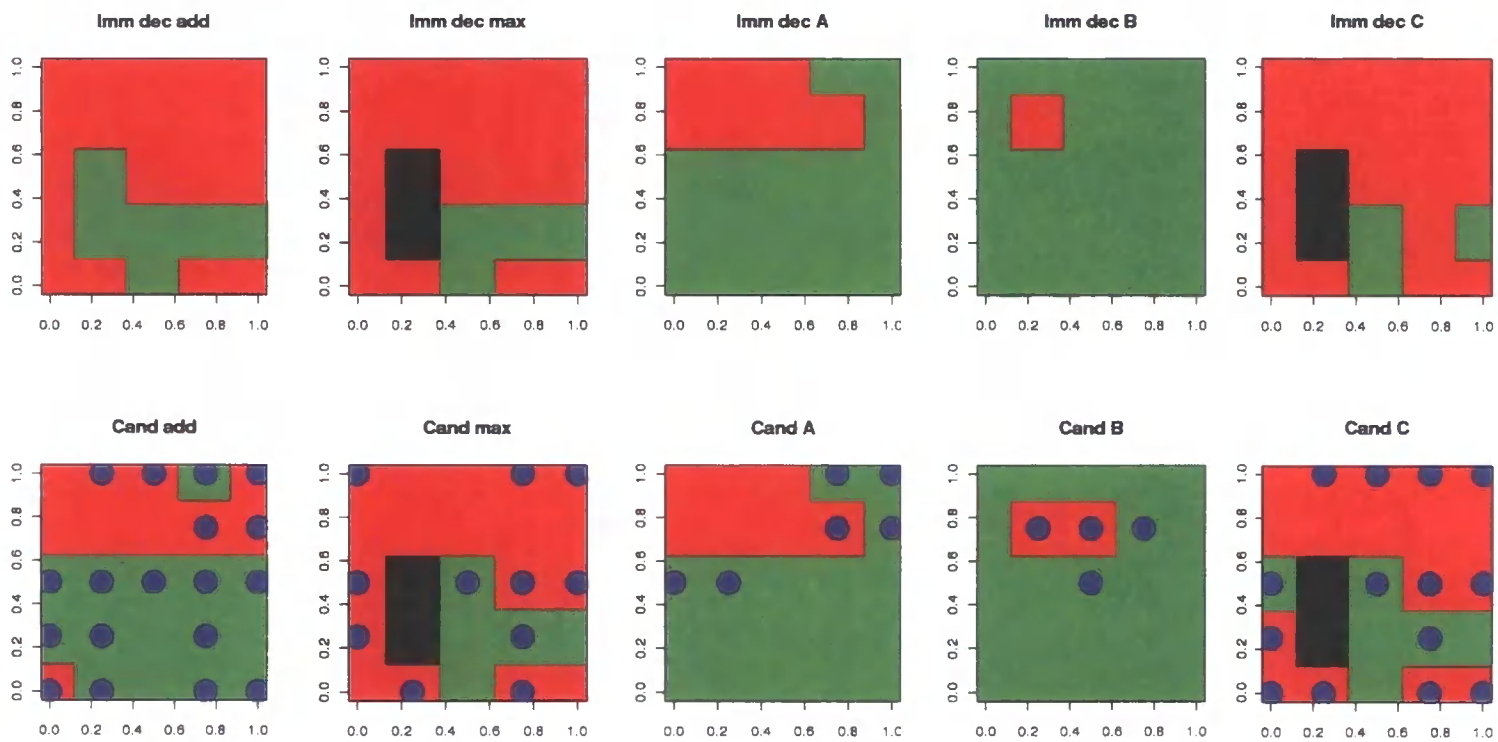


Figure 5.16: Multiple contaminant decision results. Top figures show the immediate decision to be taken for no further sampling in each case, and the lower images show the most frequently taken decision after implementing the candidate design, with frequently changing decisions in blue

Chapter 6

Selection of informative sampling designs

The final aspect to consider in our development of a decision tool for contaminated land investigation is that of selecting sample designs. We consider an approach for the selection of designs that are optimal or near optimal, based on the loss set up introduced in the decision theoretic approach of Chapter 5. In Chapters 3 and 5 we developed methodologies to allow for the inclusion of a great deal of subjective and quantitative information for the problem of site investigation. In principle (and in a world with no computational restraints), we would carry all this information through to this final stage, sample selection. In reality, the sample search algorithm we propose will not be feasible if all features are included. For example, to perform a full MCMC update at each step of a search algorithm would be very slow computationally. We will look at the elements we can retain whilst ensuring the algorithm remains practically viable.

First we look at potential methods of searching for optimal spatial sampling designs, before going on to describe the approach we have decided to implement in this thesis. We will consider this for single and multiple contaminants, and will look at several examples for the hypothetical datasets before going on to consider the case study in Chapter 7.

6.1 Current methods and computational limitations

Popular methods used to search for sampling schemes include simulated spatial annealing and sequential point selection, as discussed in [2], where these methods are compared to systematic and random sampling to demonstrate their benefit. A space filling design is developed in [77], while [16] gives an overview of the available options and recent developments. Cox [17], specifically considers sampling choice for contaminated land investigation, and considers a method of sequential addition of points based on delineating the high areas of contamination. Each of the methods in the literature has attractive features, depending on the complexity of the model, the loss criteria and computational capabilities.

Searching over a grid for the best possible configuration of points will always be a computationally intensive method, particularly when the modelling approach used for the calculation of probabilities is complex. The approach we suggest here attempts to balance the detail carried through from the modelling method and computational feasibility.

Clearly we cannot use a method whereby the expected value of every possible grid choice is calculated, as for n candidate locations this involves 2^n expected loss calculations. A possible alternative could be to use a quick method to determine a small number of candidate designs and search among these. Methods used for the selection of these designs could be a variance minimization criterion, or a method by which we search for points furthest away from existing observation locations in order to ensure coverage. An alternative to searching over designs of any number of points could be to fix the sampling budget, thereby determining the number of points we can afford to sample, and then search among all grids of this size. However, it is possible that this may also result in a large computational burden if the full candidate grid has many locations in relation to the number of points we can afford to investigate.

The method we have decided to implement is that of stepwise selection, with a carefully selected initial search grid. In principle this is akin to the ideas of

variable selection in regression analysis, as described by [53] and [42]. We will use the methods described in the previous chapter to evaluate the expected loss for a particular candidate design, where we prefer designs with a smaller expected loss value. The balance between the expected reduction in loss and the cost of each potential sampling point will drive the sample search algorithm.

6.2 Stepwise search algorithm

The use of a stepwise search algorithm reduces the number of expected losses we must evaluate from 2^n to at most $\sum_{i=2}^{n} i$ where n is the number of candidate locations. The method is a simple procedure, with three potential search options. We may sequentially add points to a design one point at a time, at each step selecting the point which gives the largest improvement. The second option is to delete points from the full design grid one at a time. Improvement will be measured by the decrease in expected loss. We continue with this stepwise algorithm until a stopping point is reached. This can happen in two ways. Either we have reached a design from which we can gain no further improvement, or a sampling budget has been reached.

While both methods mentioned, stepwise addition and deletion of points, give a reasonable method which will lead to effective designs, we discuss their limitations and then describe a combination of the two methods which will help overcome these limitations. This third option of search will add more points to a design than necessary, and delete back to some stopping point. The loss criterion developed in the previous chapter should lead us to select points which both resolve uncertainty, and help us make the correct decisions most often.

6.2.1 Stepwise add

In this case, we start with an empty grid and determine which point to add at each step. Extending the notation from the previous chapter, we are searching for $\delta^*_{fin} \in \Omega$, the "optimal" design. At each stage we will call δ^*_i the best design containing i locations. The final, "optimal" design will be denoted by δ^*. We use

the methods described in the previous chapter, specifically Section 5.5 which gives the method for calculating the expected value of a candidate design. We need to consider efficient ways of simulating from the relevant conditional distribution, as we are performing this calculation many times, rather than just once as in the previous chapter. The algorithm takes the following steps, assuming we have n_C candidate locations $\{\underline{x}_1, \ldots, \underline{x}_{n_C}\}$.

1. Find the initial best point to add to the design by calculating the expected loss for each one point design $\delta_1^1 = \{\underline{x}_1\}, \delta_1^j = \{\underline{x}_j\}, \ldots, \delta_1^{n_C} = \{\underline{x}_{n_C}\}$, and choosing that with the minimum value of $C(y, y_{\delta_1^j})$, and add this to the current best grid, δ_1^*. We make draws from the conditional distribution of $(y_{\delta_i} \mid y)$, and for each draw update the beliefs as if this was the actual observation set. So the initial location chosen will be that which minimises the total expected loss, i.e. so that

$$C(y, y_{\delta_1^*}) = \min_{\delta_1^j}(C(y, y_{\delta_1^1}), C(y, y_{\delta_1^2}), \ldots, C(y, y_{\delta_1^{n_C}}))$$

2. Compare this value to the expected loss associated with an immediate decision and if $C(y) < C(y, y_{\delta_1^*}) + L_{S_{\delta_1}}$ then we stop and do not add the point. Otherwise add the point $\underline{x}_{\delta_1^*}$ to the design δ_1^*, and continue to step 3.

3. At step i, from the remaining unsampled candidate locations, $\{\underline{x}_1, \ldots, \underline{x}_{n_C} \setminus \delta_{i-1}^*\}$, we add each location one at a time to the current best sampling design δ_{i-1}^*, and evaluate the expected loss for each. Then select the location to add to the current design which gives the minimum expected loss, i.e. so that

$$C(y, y_{\delta_i^*}) = \min_{\delta_i^j}(C(y, y_{\delta_i^1}), C(y, y_{\delta_i^2}), \ldots, C(y, y_{\delta_i^{n_C}}))$$

where $y_{\delta_i^j}$ are the observations for the previously selected points plus the j^{th} remaining candidate point.

4. If

$$C(y, y_{\delta_i^*}) + L_{S_{\delta_i}} < C(y, y_{\delta_{i-1}^*}) + L_{S_{\delta_{i-1}}}$$

add the point $\underline{x}_{\delta_i^*}$ to the design, go to step 3 and repeat, else stop with final design $\delta_{fin}^* = \delta_{i-1}^*$

So, at each step, we add one point to the design, such that the new i point design is that which yields the minimum expected loss. Step 4 says that if the expected loss for the best i point design is smaller than the best $i-1$ point design we should continue. However, if, due to the increasing cost of sampling, we can no gain from the addition of points, we stop and report the final design with expected loss value.

This method of stepwise search vastly reduces the number of designs that we must search over. While it is possible that we miss the true optimal design, the method presents a computationally viable way to obtain a sampling design which should perform well in terms of reducing uncertainty and aiding decision making. As we add each point to the design, we have one fewer candidate point to search over, which leads us to the figure of at most $\sum_{i=2}^{n_C} i$ calculations.

Making conditional draws in an efficient way

At step 3 of this algorithm, we need to make a draw for the currently selected locations, and the next candidate location. We can make these draws in several ways, some more computationally efficient than others. The simplest way would be to work out the joint distribution of each new candidate grid given the initial data at each step and update with simulations from this distribution. This would be time consuming and inefficient if we have a very fine candidate grid.

Alternatively, we generate from successive conditional distributions in order to reduce computational time. Having selected the best point to add at step i we have a set of observations for this location and so there is no need to make draws for this location again. For example, having selected the first point to add, we search for the second point to add to the design. Instead of generating the pair of observations for $y_{\delta_2^j}$ directly, we can exploit the fact that we already have simulations for the point y_{δ_1}. To do this we require the conditional distribution of $(y_{\delta_{2\setminus 1}} \mid y_{\delta_1}, y)$. With this distribution we then only need make a draw of the second value, as we already have the initial candidate point simulation. This may seem like unneccessary extra work, but if we have a large number of points to search over it will save time.

6.2.2 Stepwise delete

The stepwise delete procedure works in much the same way as the stepwise add procedure, but we start from the full n_C point design, and delete the points one by one such that the remaining points give the minimum expected loss.

1. Calculate the expected loss associated with the full design, $C(y, y_{\delta_{n_C}}) + L_{S_{\delta_{n_C}}}$.

2. Find the initial best point to delete from the full design by calculating the expected loss for each $n_C - 1$ point design $\delta^1_{n_C-1} = \left\{\underline{x}_2, \ldots, \underline{x}_{n_C}\right\}, \delta^j_{n_C-1} = \left\{\underline{x}_1, \ldots, \underline{x}_{j-1}, \underline{x}_{j+1}, \ldots, \underline{x}_{n_C}\right\}, \ldots, \delta^{n_C}_{n_C-1} = \left\{\underline{x}_1, \ldots, \underline{x}_{n_C-1}\right\}$, and choose the design which gives the minimum value of $C(y, y_{\delta^j_{n_C-1}})$. We delete the point for which we obtain this minimum, and have a new candidate grid, $\delta^*_{n_C-1}$; i.e. we require the design:

$$C(y, y_{\delta^*_{n_C-1}}) = \min_{\delta^j_{n_C-1}} (C(y, y_{\delta^1_{n_C-1}}), C(y, y_{\delta^2_{n_C-1}}), \ldots, C(y, y_{\delta^{n_C}_{n_C-1}}))$$

3. Compare this value to the expected loss associated with the full design grid and if $C(y, y_{\delta_{n_C}}) + L_{S_{\delta_{n_C}}} < C(y, y_{\delta^*_{n_C-1}}) + L_{S_{\delta_{n_C-1}}}$ then we stop and do not delete any points. Otherwise delete the "best" point $\underline{x}_{\delta^*_{n_C-1}}$ from the design δ^*_i, and continue to step 4.

4. At step i, from the remaining candidate locations δ^*_{i-1}, we delete each point in turn and evaluate the expected loss in each case. Then we select the location $\underline{x}_{\delta^*}$ to delete from the current design such that the expected loss is minimised:

$$C(y, y_{\delta^*_i}) = \min_{\delta^j_i}(C(y, y_{\delta^*_i} \setminus y_{\delta^1_{n_C-i}}), C(y, y_{\delta^*_i} \setminus y_{\delta^2_{n_C-i}}), \ldots, C(y, y_{\delta^*_i} \setminus y_{\delta^{n_C}_{n_C-i}}))$$

5. If

$$C(y, y_{\delta^*_i}) + L_{S_{\delta_i}} < C(y, y_{\delta^*_{i-1}}) + L_{S_{\delta_{i-1}}}$$

go to step 3 and repeat, else stop with final design $\delta^*_{fin} = \delta^*_{i-1}$

6.2.3 Combination of the add and delete algorithms

If we use the stepwise add search, we are not able to fully address the relationship between all of the locations, as this is dealt with in the stepwise delete method.

However, in practice the use of the stepwise delete scheme may be infeasible if we wish to start from a very dense candidate grid. We can combine the two methods to account for this, by using the stepwise add, adding more points than necessary and then deleting back to find an optimal grid. This allows for fewer calculations than deleting from a coarse grid, and will also help consider the joint effect of groups of points.

Another alternative could be to place a coarse grid over the site and select the areas of interest, then we can place finer grids over these areas and determine more specific locations of interest. As previously mentioned, in practice site investigators often work on an ad hoc basis, placing samples where accessible, and so we should not worry about placing arbitrarily fine grids.

When implementing the combined stepwise search, we have two options for a stopping point. We can add extra points and then delete back to the same number of points at which we found no further benefit. Alternatively, we could continue deleting points until it is no longer efficient to do so. This raises some questions about the point at which we stop. For example, we may have added points, and then deleted but stopped before we reach the number of points contained in the "optimal" grid as found by the stepwise add. The discrepancy may be due to the fact that we are using simulations, or that a better grid can be found by taking into account the joint distribution, and so we do not need to delete back to as few points as chosen by the original stepwise add search.

6.2.4 Calculation in practice

While we could theoretically run these algorithms for all the modelling set ups of Chapter 3, i.e. the closed update, the conditional conjugate approach and the multiple joint update, in the remainder of this work we stick to the use of individual closed updates. In the future we would like to consider ways of making the algorithm and MCMC samplers quick enough such that we could perform a full search using the full conditional approach. We feel that, given the "current" practice, using the full closed model with all the expert zonal information gives a viable methodology and considerable improvement whilst remaining practical. The procedure is no different

for multiple contaminants, as we just use the multiple decision analyses developed in the previous chapter when evaluating the expected losses.

The initial and final steps of the algorithm could involve a full MCMC update, for a single or multiple approach as required. We then use these updated parameters within the closed update to enable a computationally viable search, finishing with an MCMC run to show the final results. It should be remembered when using this method that the closed update uses different modelling assumptions (i.e. dependent β and σ^2 parameters), and so the results will be different to those if we ran the MCMC sampler at every iteration of the search algorithm. In Chapter 4 we compared the closed update with the MCMC approach and found that both methods yielded similar results, both for the hypothetical example and the real case study. We also looked at cross validation to assess the predictive performance of the methods, and found that the differences between the analyses were not considerable enough to warrant the extra computational load associated with the MCMC steps within the search algorithm.

6.3 Selection of starting grid

This algorithm performs well in terms of searching over the grid effectively to determine the best points to add to or delete from a design. However, if we start with a very fine candidate grid it may take a long time, computationally, before an optimal grid is found. We can look at ways to select a sensible starting grid for the search. Sensible points will be those where we have high uncertainty and the potential to change our decision due to the observation at that point. If we consider hypothetical contaminant A from the example of Section 3.6.2 (with dataset A2), then we can update using the initial data and consider the posterior predictive distribution in order to select points of interest. Table 6.1 shows the predictive mean values at each of the candidate grid points (the full 25 point herringbone design), along with plus and minus two predictive standard deviations. Where the values cross a decision boundary (in this case CV=10, UB=20, highlighted in red), there is a potential decision change and therefore direct information to be gained by investigating these

points. These 10 locations are shown as the purple points in Figure 6.1. The remaining 15 points are far enough from a decision boundary, or have low enough uncertainty (or both), that we can discard them as not giving enough information to warrant taking the sample. We then need to decide which of these purple points are the "best" to add to a candidate grid for full exploration. Initially, we select the point which has the potential to cross the decision boundary the furthest. That is the location where the predictive distribution suggests that the "true" contamination levels could lead to a change of decision, and by the largest margin. We sum the distance by which the points exceed the boundaries. Figure 6.2 gives a visual depiction of this. It shows in red the values which we sum to determine the best point to add, as plus and minus two standard deviations gives an indicator of the range of values that the true contamination value is likely to take. We could use any number of standard deviations that the DM wishes.

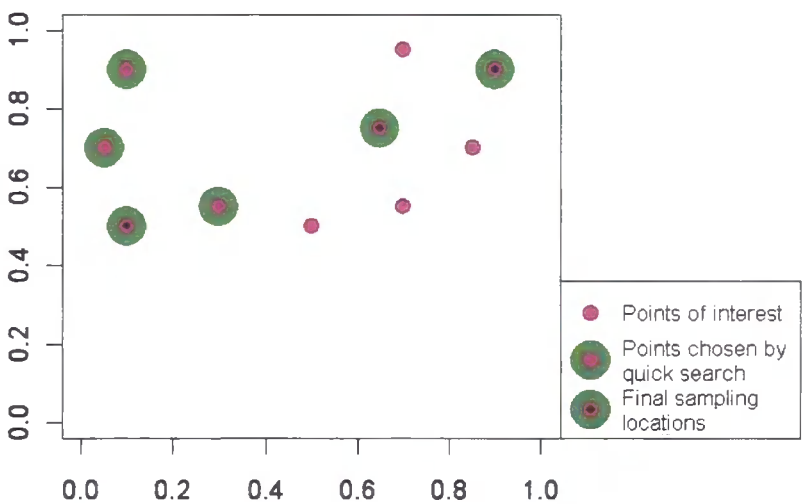


Figure 6.1: Demonstration of the initial grid search. Shown in purple are those points which have a potential decision change given the current beliefs. Then the best six points as chosen by the quick search are in green. The black points show those that were chosen by the full stepwise add, showing that these are a subset of the points chosen by the initial quick search.

We can improve this search method further by considering the impact of crossing

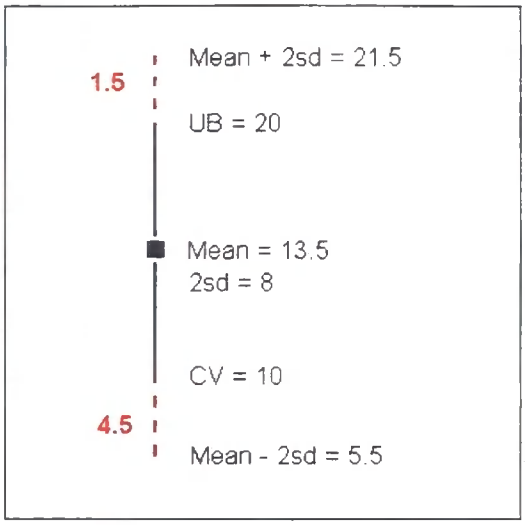


Figure 6.2: "Distance" method for determining points to include in candidate design

each decision boundary. For example, if we have two locations, at one the actual contamination value potentially exceeds the upper bound, and at the other location the predictive mean and variance gives a potential value range which contains the critical value. Which of these potential decision boundary crossovers is more important depends on the losses associated with the erroneous classification. The effect of crossing each boundary can be weighted in order to ensure that erroneous decisions with a "worse" consequence have more influence in the selection of points to be investigated. This idea drives the selection of points in the actual stepwise search, as we calculate the expected loss and select those points which minimise this value.

Using the simple criterion of considering the value of the predictive mean plus or minus two predictive standard deviations, we add the "best" of these points to a candidate grid, and then use a quicker stepwise search method by updating the predictive variance only and identifying further potentially informative points. As this method will use a very simple variance calculation, the ordinary kriging variance (for which we only need the locations not observations, removing the necessity of full updates and simulations, see Section 3.3.3), we can use a much finer grid for an initial "sweep". When we have identified the number of points we wish to investigate

Location	Mean	Plus 2sd	Minus 2sd
1	3.57	8.1	0
2	3.59	7.2	0
3	2.62	5.92	0
4	2.18	5.75	0
5	0.57	4.77	0
6	5.42	8.39	2.46
7	5.12	7.1	3.15
8	4.44	7.12	1.77
9	4.18	7.27	1.08
10	2.95	5.82	0.08
11	8.9	12.75	5.06
12	9.4	13.28	5.52
13	7.49	11.01	3.98
14	7.25	10.73	3.77
15	5.44	8.75	2.13
16	13.38	17	9.75
17	14.09	15.38	12.79
18	11.97	13.33	10.61
19	11.1	14.68	7.52
20	8.75	12.88	4.63
21	16.18	20.47	11.9
22	14.5	17.18	11.81
23	13.56	15.75	11.37
24	9.63	12.1	7.16
25	9.7	14.59	4.8

Table 6.1: Comparison of predictive mean using contaminant A, described in Section 6.3, with predictive standard deviations. Locations where there is a potential crossing of a decision boundary are highlighted in red.

fully, we run the careful stepwise search using our full modelling assumptions and this reduced grid. The steps taken in the initial search are:

1. Perform a full update with the initial dataset and calculate the predictive mean and variance, (see Equations 3.33-3.36), at all candidate locations on the full grid.

2. For each point, determine the absolute difference between the predictive mean and the decision boundaries (i.e. the critical value and upper bound). We then subtract these values from two times the predictive standard deviation to calculate a "distance".

$$\begin{aligned} dist_1 &= 2\sqrt{\text{Var}(\hat{y}(\underline{\mathbf{x}}^p) \mid \hat{y}_\Delta(\underline{\mathbf{x}}))} - \mid \lambda^* - UB \mid \\ dist_2 &= 2\sqrt{\text{Var}(\hat{y}(\underline{\mathbf{x}}^p) \mid \hat{y}_\Delta(\underline{\mathbf{x}}))} - \mid \lambda^* - CV \mid \end{aligned}$$

3. Sum these "distances" (with weighting for importance if necessary), to obtain a measure of the potential to change our decision at a point. $decpot = w_1 dist_1 + w_2 dist_2$.

4. Add the point with the maximum potential to change our decision to the final candidate grid and remove this point from the candidate grid we are searching over.

5. We then update the predictive variance at the remaining search points using the kriging variance as in Equation (3.39).

6. Repeat steps 2 to 5 using the kriging variance rather than the predictive variance as in Equation 3.36 until enough points have been selected for the careful search.

As points are selected, the predictive variance is updated at the remaining candidate locations. Then new points may become "important" in terms of their relative information value. In many cases, the points will be close on the grid, as these will remain the regions of highest uncertainty. From Table 6.1, points with no red values are least likely to be added to the sample, as the predictive standard deviation

at remaining points cannot increase using this approach, so these points will not suddenly become interesting. If any of the values minus two standard deviations fall below zero, we can truncate to zero. We discussed in the modelling chapter (Chapter 3) ways of dealing with negative predictions, by using scaling factors on the variance parameter.

The points chosen as the best to use for a careful search are the green locations in Figure 6.1. The points actually chosen in the stepwise add search using the full 25 point grid are shown in black. We see that the points which were chosen by the full stepwise add search are a subset of those chosen by the initial search to determine a good starting grid, showing that the method has chosen sensible points.

The points selected were the same when the candidate grid used was the 25 point herringbone, and the green points of Figure 6.1. The time saved by discarding points using the quick search method is considerable. For this example it was about 4.5 times faster to run the quick search and the stepwise add on the chosen 6 points than it was to perform the stepwise add search over the 25 points. This makes sense, as the 25 point grid has 4 times as many points to investigate.

6.4 Examples

Before considering the loss set-up and sample search for the real case study, we demonstrate the methodology for the hypothetical datasets. We will look at all three approaches, add, delete and the combined method. We will do this for a single contaminant (hypothetical contaminant A, with dataset A2), and then for three contaminants (A,B and C).

6.4.1 Single contaminant search

First we consider contaminant A only and use the dataset A2, with prior beliefs used in Sections 3.6 and 3.7.3. The loss set up used is from Table 5.13 and Equations (5.14-5.15). We give a CV of 10 and a UB of 20. The sampling cost for this example was 0 initial cost and 0.02 for each incremental sampling point added.

In order to account for the uncertainty which occurs at the edges of the site boundary, due to the lack of samples there, we decided to expand the herringbone starting grid. We now have a 33 point grid as shown in Figure 6.3. We used the initial search algorithm to remove $\frac{2}{3}$ of these points, and searched carefully over the remaining 11. The points selected are all in the north of the site, which is where the values are higher, and closer to decision boundaries, and the initial uncertainty is relatively high in the NE of the site. These points are close together, but in this example that should be expected due to the correlation structure. The fact that the uncertainty levels and critical levels combine to result in the optimal decision in the south of the site being very unlikely to change also directs sampling to the north.

Figures 6.4-6.6 show the results of implementing the add, delete and combined search algorithms respectively, for the same prior and loss structures in each case. The expected losses we obtain are shown in Table 6.2, where the candidate design value in the combined case is for the final grid after further addition and deletion of points from the initially selected stopping point. When we perform the combined search, we should check that the final design has a lower expected loss than the initial point at which we stopped adding, otherwise we would stick with the first result. In this case we have chosen to add 50% more points and then delete back

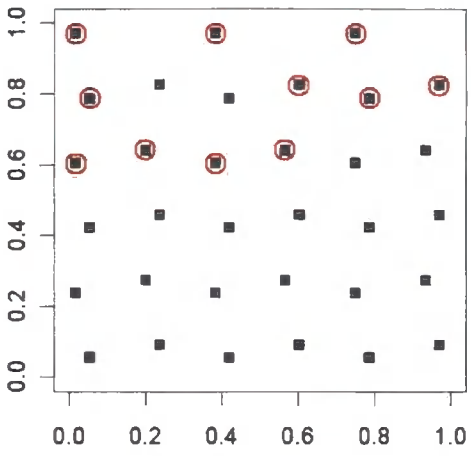


Figure 6.3: Larger, 33 point herringbone grid, with 11 points used for careful search highlighted

	Imm dec	Cand design
Add	2.978	2.713
Delete	2.978	2.708
Combined	2.978	2.682

Table 6.2: Results of the stepwise search algorithms for contaminant A, where we see the expected loss associated with implementing an immediate decision, and the expected loss for each of the "best" designs using the three methods.

to the same number at which we stopped adding. We could also have continued to delete until it was no longer worthwhile. In this case we see that the stepwise combined search results in the best expected loss, all the designs select five points. It is not unexpected that we see different results for each search method, for the reasons of joint relationships mentioned in Section 6.2.3. We should make checks on the standard error of the expected loss calculations, and ensure the value is small enough to allow us to make sampling choices with a reasonable level of confidence. Here we need to run in the region of 3000 simulations to obtain acceptable results.

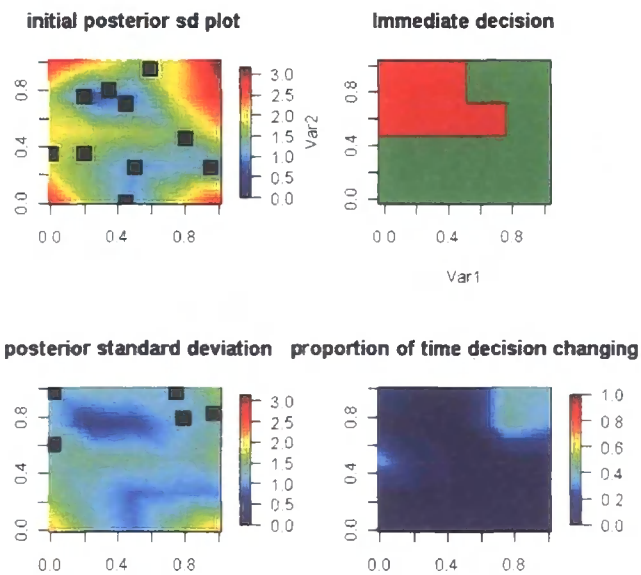


Figure 6.4: Results for the single contaminant stepwise add search routine, contaminant A. The top left image shows the predictive standard deviation after the initial data is observed, the top right shows the immediate decision taken. The bottom left image shows the final locations selected by the stepwise search, and the resulting uncertainty if this design were to be implemented. The bottom right image shows the proportion of times the decision taken using simulations differed from the immediate decision in the final design.

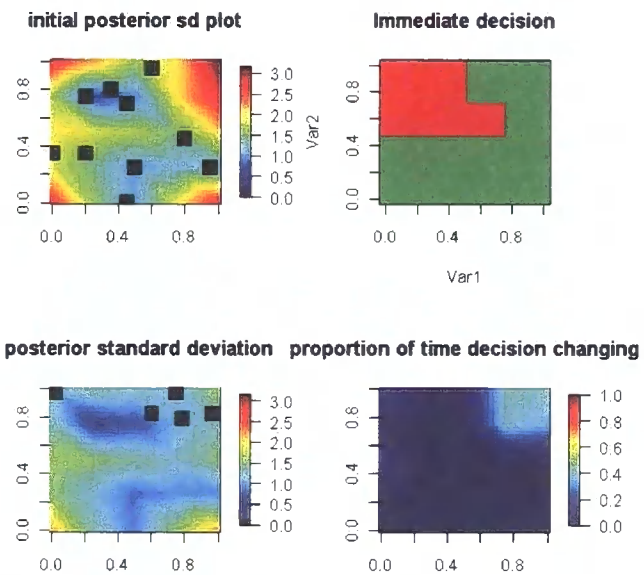


Figure 6.5: Results for the single contaminant stepwise delete search, contaminant A. The top left image shows the predictive standard deviation after the initial data is observed, the top right shows the immediate decision taken. The bottom left image shows the final locations selected by the stepwise search, and the resulting uncertainty if this design were to be implemented. The bottom right image shows the proportion of times the decision taken using simulations differed from the immediate decision in the final design.

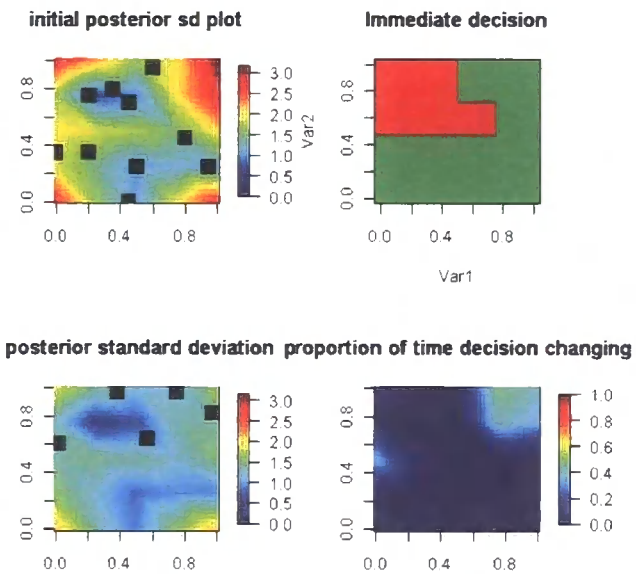


Figure 6.6: Results for the single contaminant stepwise combined search routine, contaminant A. The top left image shows the predictive standard deviation after the initial data is observed, the top right shows the immediate decision taken. The bottom left image shows the final locations selected by the stepwise search, and the resulting uncertainty if this design were to be implemented. The bottom right image shows the proportion of times the decision taken using simulations differed from the immediate decision in the final design.

In Appendix B.2 we also consider single contaminant searches for hypothetical contaminants B and C.

6.4.2 Multiple contaminant search

Selection of search grid

In order to combine the quick search for multiple contaminants, we run the algorithm for each contaminant, select the required number of locations and then merge these grids to find a "good" starting grid. This is where using the initial search algorithm may benefit from a more detailed criterion, as when contaminants are distributed very differently through the site we are likely to retain a large number of initial points, leading to heavier computational loads.

When looking for sampling locations in practice, it is likely that the site investigator will want one sampling scheme where all contaminants are tested for, regardless of the different distributions of contaminants across the site. It would be inefficient and unrealistic to go to different locations and only test for the particular contaminant of interest. In reality, every sample taken is analysed for a full suite of contaminants, and so it is pointless to have a sampling scheme for each separate contaminant. This could lead to new contaminants becoming interesting to the investigator in terms of SPOSH. If this is the case a prior belief elicitation could be carried out for the new contaminant and added to the analysis. We can combine unrelated contaminants in this set up to determine the best overall sampling scheme. However, we may decide that we wish to keep the two search schemes completely separate. For this example we will consider A,B and C together, but firstly look at the pair of contaminants A and B, as these appear to be distributed similarly through the site.

Contaminants A and B joint search

For this example we use the additive multiple decision calculation (rather than the maximum, as discussed in Section 5.7.2). Table 6.3 shows the resulting expected losses. In this example the minimal expected loss is obtained using the combined

search, where four points were selected, as in the other cases which performed similarly. However, it is difficult to ascertain the difference between results when using scaled utilities, in the next chapter we will work in terms of financial loss as utility to demonstrate how communication with the DM can be made as simple as possible. We will look at the effect of weightings in the next example. The search scheme again concentrates on the north of the site, where we saw sampling locations selected for contaminant A. The decision in the north east region of the site has changed from the single search for contaminant A only to include an extra point requiring remediation, even though no remediation was required for contaminant B (as seen in Appendix B.2). This effect is due to the cumulative expected failure cost of taking no action for two contaminants becoming higher than the cost of remedial action, which is now one cost which deals with both contaminants.

While the proportion of time the decision changes, as shown in each figure, is generally low, this should not be a sign of uninformative sampling. As we use the predictive distributions to draw samples, we will not regularly observe extreme values in the simulation exercise. If we were to actually implement the sampling as a stepwise routine in practice, we would obtain real data and possible changes of decision. This algorithm uses all the current information to determine a best course of action using simulated values.

	Imm dec	Cand design
Add	3.590	3.400
Delete	3.590	3.440
Combined	3.590	3.361

Table 6.3: Multiple contaminant search results, in this case for contaminant A and B. Again we see it is worth implementing a sampling design, in this case the combined search finds the sampling scheme with minimum expected loss.

All three hypothetical contaminants

Finally we consider sample selection for all three contaminants A,B and C to determine the best course of action in the hypothetical case. We perform the analysis

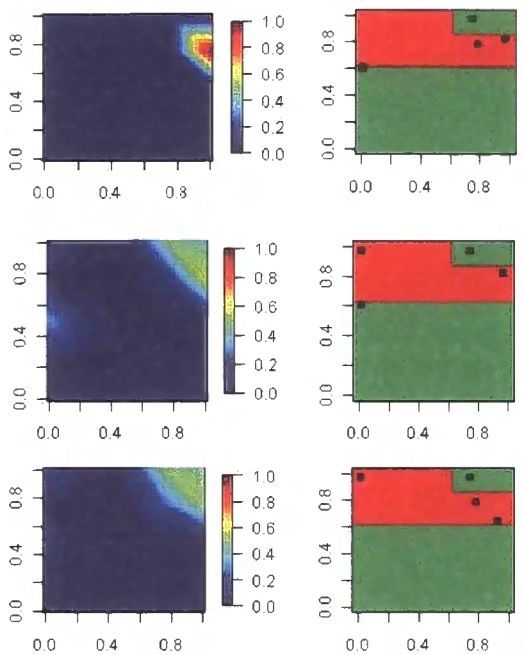


Figure 6.7: Search results for joint stepwise search, contaminants A and B. From top to bottom the figures show the results of the add, delete and combined searches respectively. Figures on the left showing the proportion of time the decision is changing as a result of the simulation draws, and the figures on the right show the immediate decision with no further sampling, along with the locations selected by the search algorithm.

with two different loss weightings. First we use equal weights of 1, and then $\left(\frac{1}{2}, \frac{1}{2}, 2\right)$. It may be deemed that for "similar" contaminants (in terms of effects and dealing with failure), that an averaging of losses incurred would be more reasonable. This does not affect the cost of implementing each remediation method.

In Figure 6.8 we see the results of the search algorithms, and Table 6.4 gives the expected losses in each case. We see similar sampling results for all three cases, with an extra point added in the delete search. We have demonstrated here that all three designs perform similarly, and the combined search is effectively a computational compromise between the add and delete methods. It allows us to add points and then take account of relationships between points by adding extra and then deleting

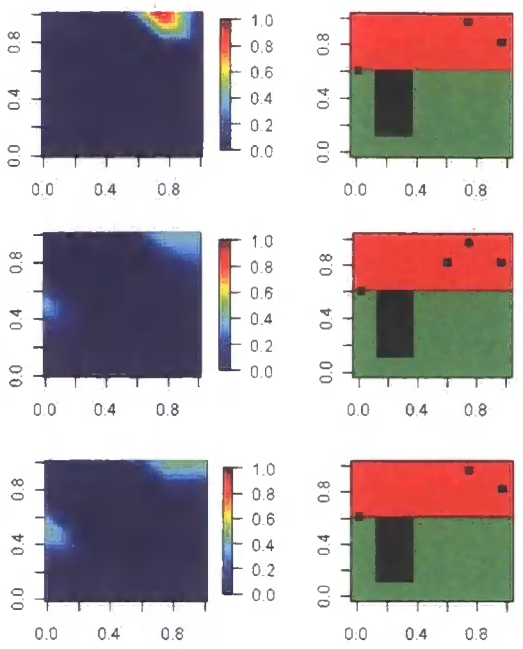


Figure 6.8: Results of the three contaminant search routines. From top to bottom: add, delete, combined. Left hand figures display proportion of time decision changing from immediate decision, right hand images show the immediate decision and locations selected by the search algorithm.

back. However, the most efficient method will depend on the final number of points selected and the density of the starting grid. For example, if we start with 100 candidate points and the final designs have 10 points, then it will be faster to use the stepwise add routine.

To compare the times taken for the multiple example, even when three individual searches are performed and then combined to find an initial grid, the method of performing an initial sweep then careful search is almost twice as fast as performing a careful search on the full grid.

For the alternative weights we do not see a great difference, except again in the north east of the site where the reduced weighting of the contaminants A and B allows us to remediate two fewer locations. This highlights the need to be as realistic and clear as possible when assigning loss functions. We will further consider the effect of altering loss structures in the next chapter.

	Imm dec	Cand des
Add	4.880	4.700
Delete	4.880	4.782
Combined	4.880	4.759

Table 6.4: Results for the multiple, three contaminant example, with weights (1,1,1) and the additive form used for the combination of losses.

	Imm dec	Cand des
Add	4.746	4.662
Delete	4.746	4.683
Combined	4.746	4.627

Table 6.5: Results for the multiple example, weights $(A = \frac{1}{2}, B = \frac{1}{2}, C = 2)$

We have introduced and demonstrated a sample search algorithm, which we will implement for the real case study in the next chapter. We have used a method which is computationally faster than a full search over all possible grids, and which offers several options as well as an initial sweep through of points to determine those which are of most interest in terms of uncertainty reduction and decision changing potential.

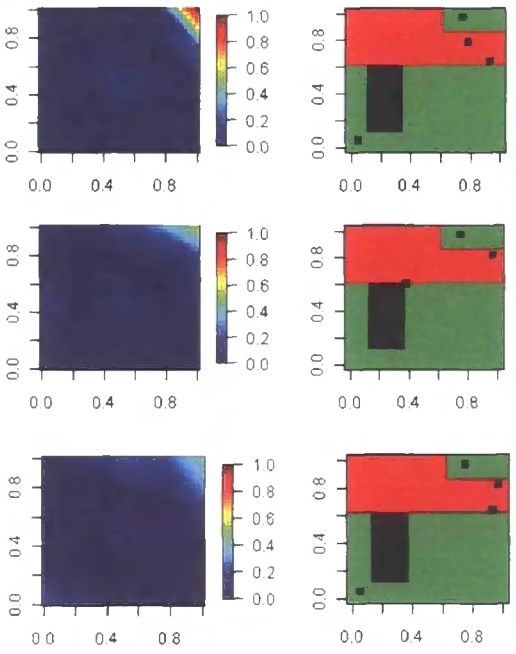


Figure 6.9: Results of the three contaminant stepwise searches with alternative weightings. From top to bottom we see the results of the add, delete and combined stepwise searches.

Chapter 7

Decision and sampling analysis for a real site

In order to demonstrate the practical application of the methods introduced in this thesis, we will now carry out a decision analysis and sampling search for the real case study. We have considered the prior elicitation and modelling for the site in Chapter 4. It now remains to select a set of optimal sampling locations for a second stage investigation.

We will first set up a realistic cost structure, and consider the decisions we can make, and then we will implement the methods of the previous two chapters. Throughout this example we will treat money as utility. This tends to be common practice in practical applications, it allows for a clearer description of results to DMs. Much of the hard work required was dealt with in the modelling presented in Chapter 4, reminding us of the importance of a thorough elicitation and model diagnostics.

7.1 Cost set up

We assign costs in the way described in the previous chapters. We require sampling costs, remediation costs and failure costs. These are chosen through discussion with the relevant experts. We bear in mind that the values we specify are directly related to the size of the area that each inference point represents. For this purpose, a

square regular inference grid is preferable, as it allows ease of calculation.

The dimensions of the site are approximately $300m \times 120m$, and we have chosen to divide the inference grid into $25m \times 25m$ regular squares. So, using an approximate value of $\frac{1}{1.6}$ tonne$\approx 1m^3$, we will have 390.625 tonnes of soil to deal with per remediation unit. These remediation units are quite large, in that we will be making a decision for each $625m^3$ of soil. However, they show how the methodology can be used to guide the DM by showing the preferred course of action for smaller sections of the site. The ideal scenario would be to determine a course of action at every single location, but this is not computationally feasible.

7.1.1 Remediation options and costs

For this example we consider that there are three options per type of contaminant. That is, we look at remedial options for the metals, Arsenic, Lead and Zinc, and then for BaP separately. The metal remediation method need only be paid for once and deals with all three of the contaminants. In many cases a different set of remedial options will be required for this contaminant. In a similar manner to the hypothetical loss structure, we will assume that there are three possible remedial actions.

- The first remediation method d_1 can be assumed to be successful for any level of contamination. This will be an ex-situ method, that of hazardous landfill. This method will be suitable for the metals and the BaP, as the soil is being taken away from the site, dealing with both types of contaminant. On discussion with site investigation experts, we assign a cost of £250 per tonne to this method.

- The second remediation method d_2 will be less expensive, but also potentially less effective at high levels of contamination. This will be an in-situ method, such as windrowing [1] for the BaP, and stabilisation [2] for the metals. The cost

[1] Windrowing uses aeration and mixing of soil, addition of nutrients and control of moisture content to reduce contaminant levels

[2] Stabilisation methods reduce the solubility, mobility and toxicity of metals.

will be £75 per tonne for each method, which is a considerable reduction on landfill, but also a more time consuming method. We will assume in this loss set up that d_2 is appropriate until the contamination levels pass the upper bound. If it was deemed that both the metals and the BaP required in-situ remediation, we use both methods and one tonne would cost £150.

- The final decision d_3 is monitored natural attenuation. This will cost £15 per tonne, which is cheap compared to the other two methods, but we will have higher failure costs to deal with for this option.

Table 7.1 shows the cost of each possible decision, at several scales. It shows the cost of implementing each option for the whole site, and each inference unit, showing that we can make improvements on cost even by taking 60 separate inference location decisions instead of making a site-wide decision.

Remediation method	Cost per m^3	Cost per inf "point"	Whole site cost
d_1 Hazardous landfill	156.25	97,656.25	5,625,000
d_2 In-situ metal/organic	46.875	29,296.88	1,687,500
d_3 No action (MNA)	9.375	5,859.38	337,500

Table 7.1: Costs of possible remediation options, per m^3, per inference location and for the whole site

7.1.2 Failure costs

Attaching realistic failure costs is more difficult than a remediation cost, as there are many ways in which loss can be incurred, as discussed in Section 5.2.2. We assume that having "failed" at a location, we have to spend the relevant amount to rectify the decision, as well as an extra penalty, which represents losses incurred through factors such as: damage to human health, damage to reputation, legal costs, loss of land value, time wasted. We alter these costs to see the effect of different failure losses. We require upper bound levels to determine the failure losses. While these don't exist in the current guidance, we use 3× the relevant critical value, and for BaP 10×. These are used in this example to determine the point at which failure

becomes "worst" when we take no action (represented here by MNA), and also to determine the levels above which d_2 is no longer effective. This is not the only way in which such figures may be used. Our framework is such that the DM can implement loss functions as they see fit for any particular site.

As well as using the SGVs, we briefly consider site specific target levels (SSTLs), which were used in the original site report. This allows for more meaningful comparisons to be made. Table 7.2 gives the loss set up, and Equations 7.1-7.8 give the

	L(As)	L(Pb)	L(Zn)	L(BaP)	Cost of method, £
d_1	0	0	0	0	97,656.25
d_2 metal	F_{As2}	F_{Pb2}	F_{Zn2}	F_{BaP3}	29,296.88
d_2 organic	F_{As3}	F_{Pb3}	F_{Zn3}	F_{BaP2}	29,296.88
d_3	F_{As3}	F_{Pb3}	F_{Zn3}	F_{BaP3}	5,859.38
Sampling	-	-	-	-	1500

Table 7.2: Loss table for real case study, where the cost of the method is per square on the inference grid

functions. We vary the values of the c costs in these equations in order to observe the impact on the decision of different failure costs. This value is the most difficult to judge, and is a topic that requires further investigation in the future to allow for effective decision analyses. We give a smoothly increasing function for each failure cost, increasing from 0 up to the "maximum" cost. We see a similar structure for each loss function, as we assume an increasing loss with contamination, up to our "worst case scenario" at the upper bound.

The rows of the table relating to the d_2 loss show that the failure loss for the metals when using the organic method are the same as taking no action and vice versa. This is because the two methods are not appropriate for the other type of contamination. We combine these decisions when we analyse the multiple example.

$$F_{As2}(As) = \begin{cases} 0 & \text{if } (As < 60) \\ 97,656.25 + c_{As2} & \text{otherwise} \end{cases} \tag{7.1}$$

$$F_{As3}(As) = \begin{cases} 0 & \text{if } (As < 20) \\ 29,296.88 + c_{As2}\left[\exp\left(\ln 2 \times \frac{(As-20)}{40}\right) - 1\right] & \text{if } (20 < As < 60) \\ 97,656.25 + c_{As3} & \text{otherwise} \end{cases} \tag{7.2}$$

$$F_{Pb2}(Pb) = \begin{cases} 0 & \text{if } (Pb < 1350) \\ 97,656.25 + c_{Pb2} & \text{otherwise} \end{cases} \tag{7.3}$$

$$F_{Pb3}(Pb) = \begin{cases} 0 & \text{if } (Pb < 450) \\ 29,296.88 + c_{Pb2}\left[\exp\left(\ln 2 \times \frac{(Pb-450)}{900}\right) - 1\right] & \text{if } (450 < Pb < 1350) \\ 97,656.25 + c_{Pb3} & \text{otherwise} \end{cases} \tag{7.4}$$

$$F_{Zn2}(Zn) = \begin{cases} 0 & \text{if } (Zn < 1944) \\ 97,656.25 + c_{Zn2} & \text{otherwise} \end{cases} \tag{7.5}$$

$$F_{Zn3}(Zn) = \begin{cases} 0 & \text{if } (Zn < 648) \\ 29,296.88 + c_{Zn2}\left[\exp\left(\ln 2 \times \frac{(Zn-648)}{1296}\right) - 1\right] & \text{if } (648 < Zn < 1944) \\ 97,656.25 + c_{Zn3} & \text{otherwise} \end{cases} \tag{7.6}$$

$$F_{BaP2}(BaP) = \begin{cases} 0 & \text{if } (BaP < 10.4) \\ 97,656.25 + c_{BaP2} & \text{otherwise} \end{cases} \tag{7.7}$$

$$F_{BaP3}(BaP) = \begin{cases} 0 & \text{if } (BaP < 1.04) \\ 29,296.88 + c_{BaP2}\left[\exp\left(\ln 2 \times \frac{(BaP-1.04)}{9.36}\right) - 1\right] & \text{if } (1.04 < BaP < 10.4) \\ 97,656.25 + c_{BaP3} & \text{otherwise} \end{cases} \tag{7.8}$$

Sampling costs

For this study we will assign a zero initial sampling cost, and only have an incremental cost per borehole drilled. When we take a sample we assume it will be taken at the centre of the grid location. The cost per borehole is £1500, which includes analyses for all contaminants of interest, as having taken a sample, we suppose that it would be analysed for everything, which is why searching for all contaminants together is beneficial.

The sampling cost of £1500 is small in relation to the costs associated with the remediation methods. As such we may have computational issues; if the code cannot be run until the standard error of the loss calculation is small enough, we can add a "tolerance" level to our calculation. This tolerance level would be a number multiplied by the sampling cost. If we find a very small difference between the current and previous "best" minimum expected loss, we may force the sampler to look ahead a step to see if the addition or deletion of a point may improve the expected loss and potentially save us money. For example the joint learning ability of two points may be significantly more than adding one. However, it may not be worth adding one, but by looking ahead we see that it is worth adding the two points using this tolerance level. This could be particularly useful in the case of the stepwise delete, as in reality the site investigator would like to go to as few locations as possible simply for convenience. Of course if we see a very small difference between two candidate designs, then we should probably choose to go to that with the larger number of points, as this will allow us to obtain as much real data as possible, and therefore an increased learning potential. By using the look ahead, we increase the chances of selecting a "good" sampling configuration without having to run the algorithm for a huge number of simulations.

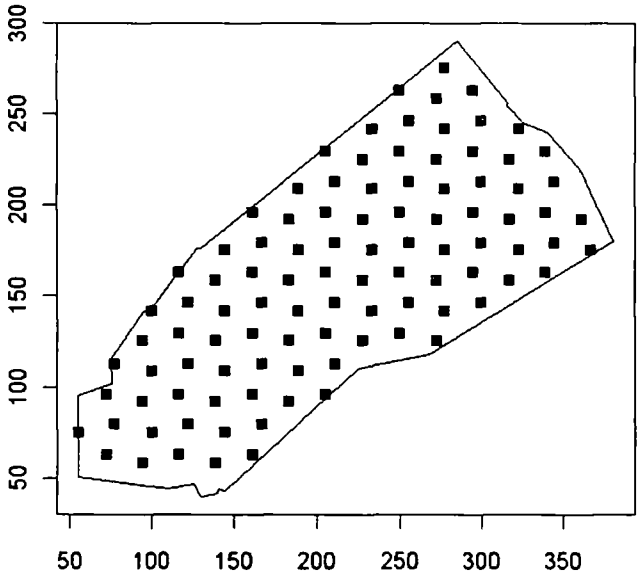


Figure 7.1: Initial herringbone grid placed over the site

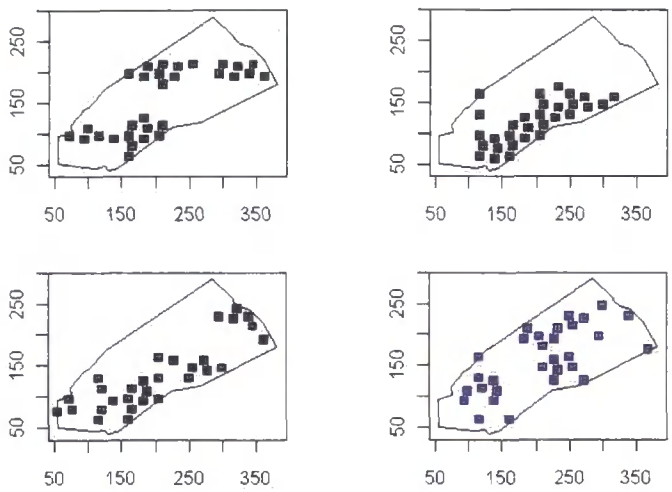


Figure 7.2: Points selected by the search algorithm as "good" starting grids for the careful search. Top left, Arsenic, top right, Lead, bottom left, Zinc, bottom right BaP.

7.1.3 Selection of initial grid

We will first place a fine herringbone grid over the site, as shown in Figure 7.1, and use the initial quick selection method to determine those points which will be informative in order to run the sampling search algorithm. We bear in mind that points that are "close" to each other on the grid may not be close in terms of the specified correlation length, and as such we may select points next to each other on the grid if they are in a region where the decision is expected to change.

For this example we will use the method outlined in Section 6.3, with a slight change. In the methodology described, we allowed for a weighting to account for the relative "importance" of crossing each decision boundary. In this example, each time a point is added in the quick search, the weighting associated with crossing the two decision boundaries is switched from (0,1) to (1,0), and then back again for the next point. By doing this we can consider those points which may cross one boundary or the other, rather than considering the cumulative effect, which may not pick up the most informative points as effectively as summing the losses. So, every other point selected by the quick search algorithm will consider a different change

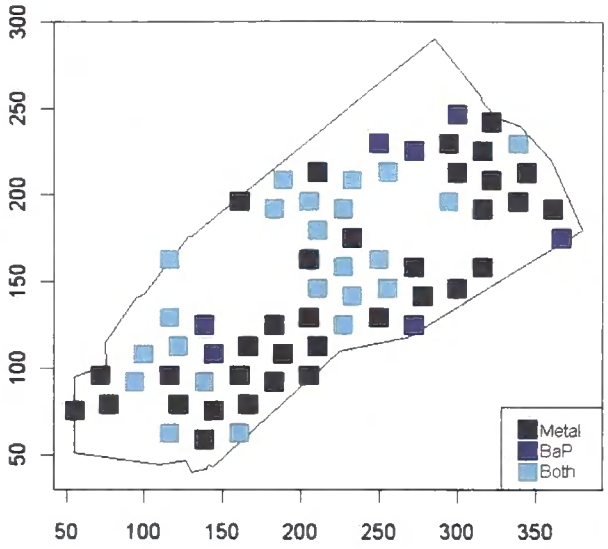


Figure 7.3: Combined points from initial search algorithm for all four contaminants, showing whether the points were initially selected for metals, BaP, or both.

of decision, as both of these costs are important as they may save us money.

We run the search algorithm for each of the contaminants in turn. We decided to select the 30 most informative points, the resulting grids are shown in Figure 7.2 as the best points to investigate. These are points where the decision may change. When we merge these grids to find the overall search grid for the multiple search we get points as shown in Figure 7.3.

7.2 Single contaminant analyses

For the single contaminant analyses we consider at a number of failure cost structures for the stepwise add searches to determine the effect of altering this very uncertain cost. We will look at a stepwise add and delete search for the four contaminants and then go onto the multiple contaminant search.

7.2.1 Decision analysis and sample search for Arsenic

For Arsenic we give a fixed loss when the upper bound is crossed and d_1 has not been implemented, as this is the point where failure becomes the "worst". Up until

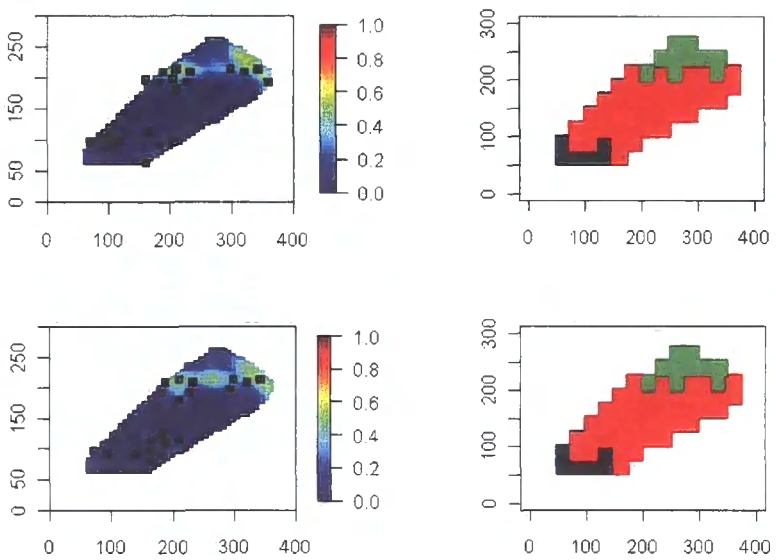


Figure 7.4: Results of the stepwise add and delete search for Arsenic, where the top images show the sampling design for add, along with the proportion of time the decision is changing, and also the immediate decision if no more sampling were to be undertaken. The same is shown at the bottom for the delete algorithm.

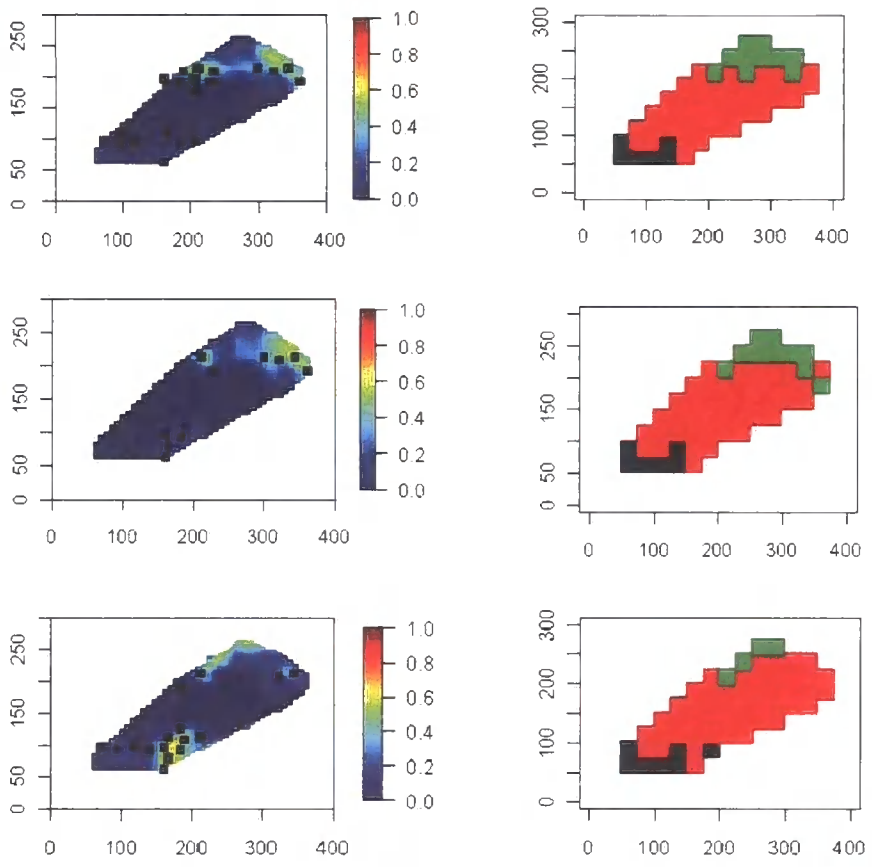


Figure 7.5: Results of the stepwise add search for Arsenic with the initial grid shown in Figure 7.2. From top to bottom, case 1, 2, 3 from Table 7.3, with left hand figures showing selected locations and proportion of simulations where decision changed from immediate decision, which is shown in the right figures.

Case	c_{As2}	c_{As3}	Imm dec	Final add
1	20,000	40,000	1,705,753	1,672,114
2	60,000	120,000	1,792,245	1,736,967
3	400,000	600,000	2,309,214	2,072,121

Table 7.3: Effect of varying failure costs on the expected loss for Arsenic contamination in the stepwise add algorithm. All values are in £.

this point we will have failure increasing as a function of actual contamination levels, from the critical value to the upper bound.

In this case the optimum number of points is 20 when using a 30 point starting grid as selected in Section 7.1.3. Table 7.3 gives the results of the stepwise add, and for comparison in Case 1 we obtain a expected loss of £1,663,597 for the stepwise delete search. This gives a slightly lower expected loss, this is to be expected in general as the stepwise delete takes into account joint relationships and so potentially gives us more information. Figure 7.4 shows the sampling designs chosen by the add and delete schemes respectively, with the cost structure of Case 1. The top image in Figure 7.6 shows the points chosen by the add search using the full initial grid, and so we see that using the initial sweep through followed by the careful search gives good results while saving on computational effort. We require approximately 5000 simulations in order to ensure the standard error of our expected loss calculation is at an acceptable level. In practice the algorithm should be run for as long as is possible to ensure the best results (for a few days in an ideal world). It took some time to obtain results with 5000 simulations, and it may not be feasible to run the search for this long in routine use. This is where the tolerance level may come in useful, as it allows us to continue searching even when the standard error is not smaller than the difference between the expected loss values that are being calculated. We may wish to run the search several times, but for fewer simulations each run in order to obtain an idea of whether a stepwise add or delete will reach a stopping point first (e.g. if we have a candidate grid of 50 and 7 points are selected then we would prefer the stepwise add). Then we can run the search algorithm longer to obtain better results. In Table 7.3 we see the increasing costs when we

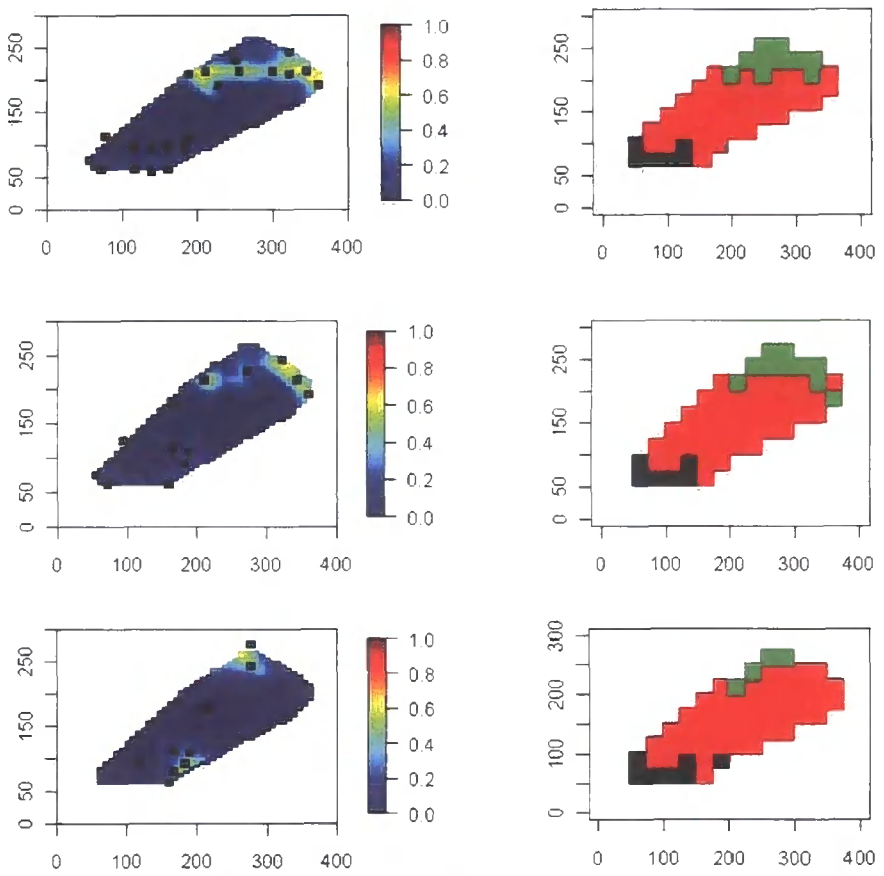


Figure 7.6: Results of the stepwise add search for Arsenic, for full search grid without initial sweep through, again for cases 1,2 and 3 from Table 7.3

Case	c_{Pb2}	c_{Pb3}	Imm dec	Final add
1	20,000	40,000	1,351,502	1,294,402
2	60,000	120,000	1,374,196	1,316,119
3	400,000	600,000	1,429,916	1,392,725

Table 7.4: Effect of varying failure costs on the expected loss for Lead contamination

increase the failure cost for the case of the stepwise add search. This is intuitive, as the failure cost increases we will incur higher costs, whether through changing our decision in order to avoid the higher failure cost, or the increase in potential loss from taking no remedial action (i.e. d_3). Figure 7.5 shows the changing immediate decision with the increasing failure costs, where the area in the north east of the site is turning to red (d_2) from green (d_3) as the expected loss associated with failure becomes greater than the cost of d_2.

To compare the results of our modelling with the recommendations of the site report we note that SSTLs are used, and using these values we found that it is not worthwhile to take any more sampling for Arsenic as no remediation is required. This shows that in reality our model backs up the findings of the site report. However, we chose to use the SGVs to demonstrate the potential of the search algorithm. We will see for BaP that further sampling is required in the site report, and so it is useful to see the results for that case.

7.2.2 Decision analysis and sample search for Lead and Zinc

The search strategy and set up for Lead and Zinc are very similar to that for Arsenic, we briefly give the results of the search for the stepwise add and delete methods in the original cost set ups, and then the stepwise add method for two further failure costs.

In this case (Table 7.4), the stepwise add performs slightly better than the stepwise delete, for Case 1 we get £1303506 for the delete search. A difference of £9000 between the two methods is not too large, the cost of six samples. We would expect differences in these results, and in comparison to the overall expected loss this is not considerable. Again, the losses increase as expected with increasing failure

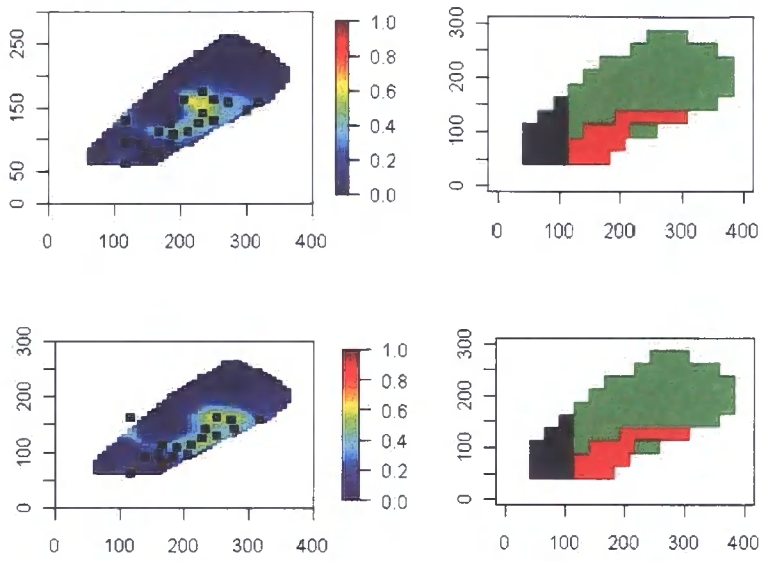


Figure 7.7: Result of the stepwise add and delete search routine for Lead contamination. Top shows add, bottom shows delete.

Case	c_{Zn2}	c_{Zn3}	Imm dec	Final add
1	20,000	40,000	1,494,521	1,389,413
2	60,000	120,000	1,563,937	1,430,543
3	400,000	600,000	1,870,132	1,527,136

Table 7.5: Effect of varying failure costs on the expected loss for Zinc contamination

cost. Figure 7.7 shows the chosen sampling designs for the stepwise add and delete methods respectively. For Zinc, Table 7.5 shows the results of the stepwise add, and for comparison we obtain £1389219 for the Case 1 delete. These values are very close, suggesting that in practice we should use the method which is fastest computationally as both perform similarly. The points selected by the algorithm do differ, this may be due to several factors. If the points are close in terms of the correlation length then there will be similar information to be gained from two locations not very far apart, so either could be chosen. Again we note that this algorithm uses a simulation method, and so if we are obtaining marked changes we would suggest further investigating the standard error of the expected loss calculation as a performance indicator.

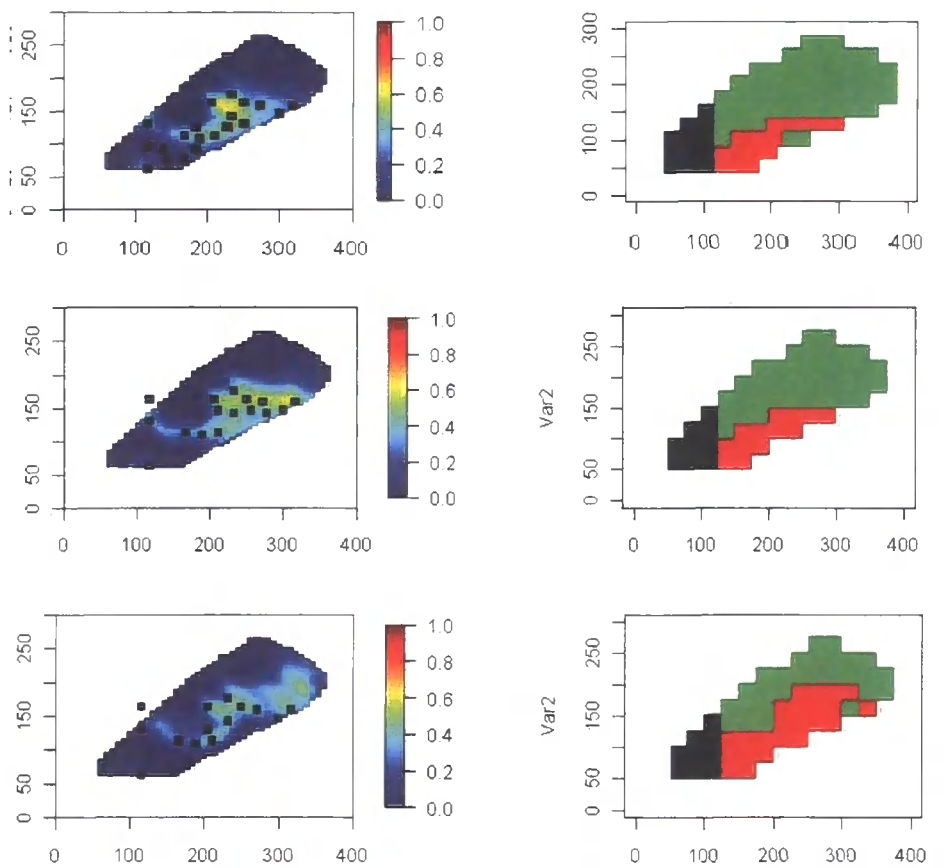


Figure 7.8: Result of the stepwise add search routine for Lead contamination, from top to bottom for Cases 1,2,3 in Table 7.4

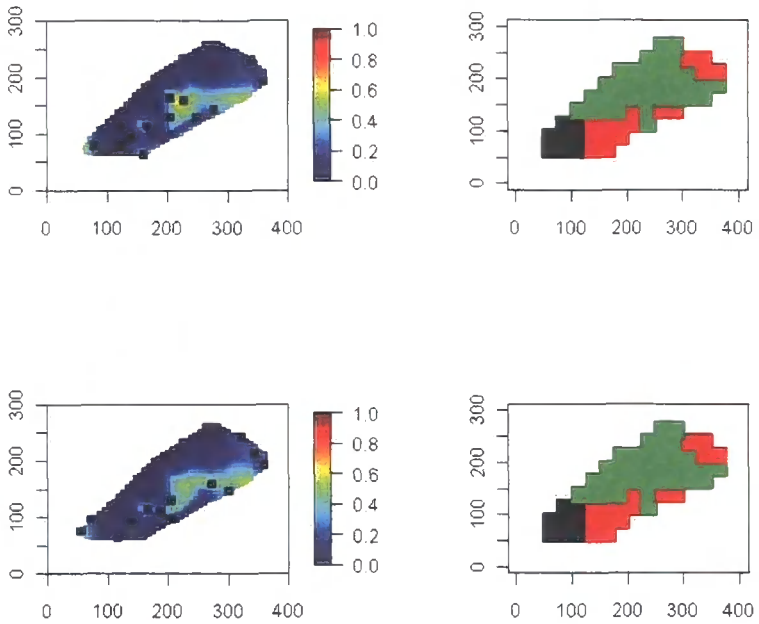


Figure 7.9: Results of the stepwise search for Zinc, stepwise add at the top and stepwise delete at the bottom.

7.2.3 Decision analysis and sample search for BaP

For the analysis of BaP, the only real difference in the set up was the value of the UB used. For the metals we gave a value of 3 times the critical value, and here we give 10 times. As seen in the earlier chapter, and also the site report, BaP is a difficult contaminant to work with. The critical value suggested by the EA is so low that often the natural underlying levels in soil exceed the critical value, causing problems for site investigators, in that the whole site may be determined as contaminated when in fact the soil is healthy and has naturally occuring levels present. However, in this site there were clear physical signs of organic contamination present at the riverside and also seeping from the soil. We see here a more clear cut difference between the decision made, in that the decision was either ex situ remediation, or no action. This is because of the given critical values, and the very small range of contamination values which would suggest the in-situ remediation method. The expert may wish to further consider the extreme value observed by the riverside. This is where the removal of a point may be a reasonable option, as the high BaP

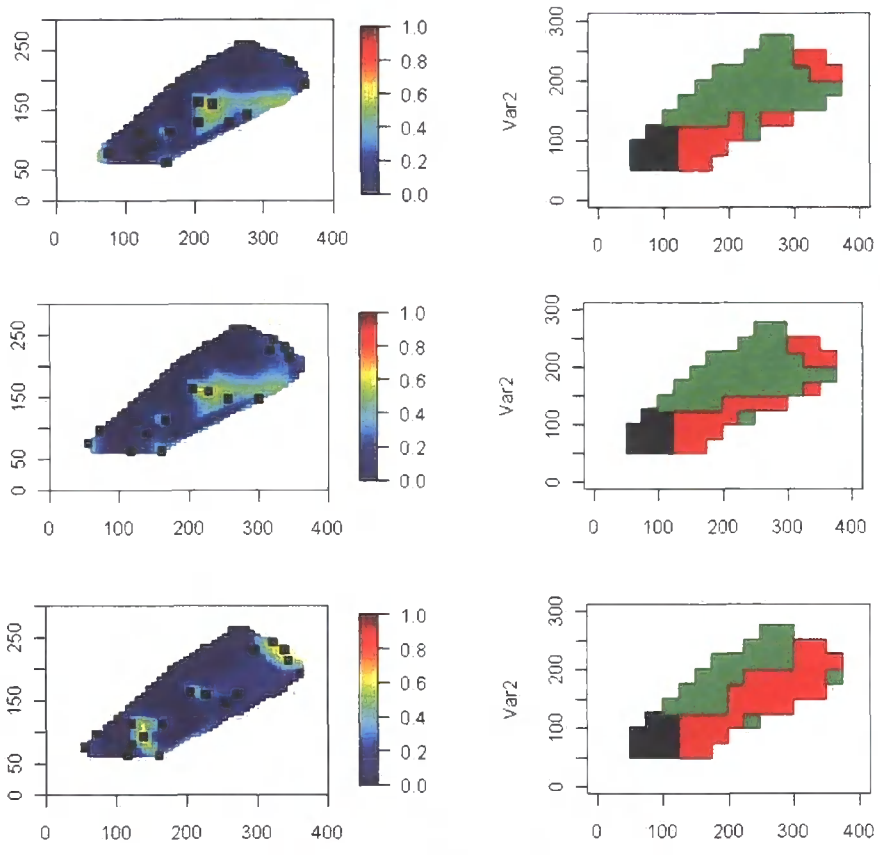


Figure 7.10: Result of the stepwise add search routine for Zinc contamination, for Cases 1,2,3 in Table 7.5

	Imm dec	Final add	Final del
BaP	4,680,523	4,514,396	4,490,041

Table 7.6: BaP results, stepwise add and delete

is influencing a large area of the site.

We see a much higher expected loss associated with the BaP contamination. This is to be expected, and reflects the findings of the site report. We see quite a difference between the expected losses associated with the stepwise add and delete methods. However, the same number of points are chosen, all be it at different locations. This occurs due to the different simulations and possibly from the joint relationships of candidate points which is taken to account in the delete method. As there was more uncertainty in the BaP analysis, we require more simulations than for the Arsenic. We can combine a smaller number of simulations with the tolerance level mentioned previously in order to obtain a faster result if necessary. In reality, the time between stages of sampling in a site investigation can be anything from a week to six months, so while we would like the sample search algorithm to be as fast as possible for convenience, it does not present an immediate stumbling block for real life implementation if the search takes a day or two.

7.3 Multiple decision analysis

After considering the exploratory findings of the elicitation and modelling undertaken in Chapter 4, we perform a multiple decision set up for all four contaminants of interest together. However, we will use individual updates as this is computationally more efficient and we do not improve our learning ability enough to justify the increased computational time associated with a joint update. This increase in computational time is associated with the updating part of the analysis, and we showed in Chapter 4 the time saved by performing individual updates. While we see a considerably different distribution of BaP to the metals, in reality the site investigator aims to delineate the two types of contamination in one set of sampling. This will allow them to determine the overall remediation strategy required in the

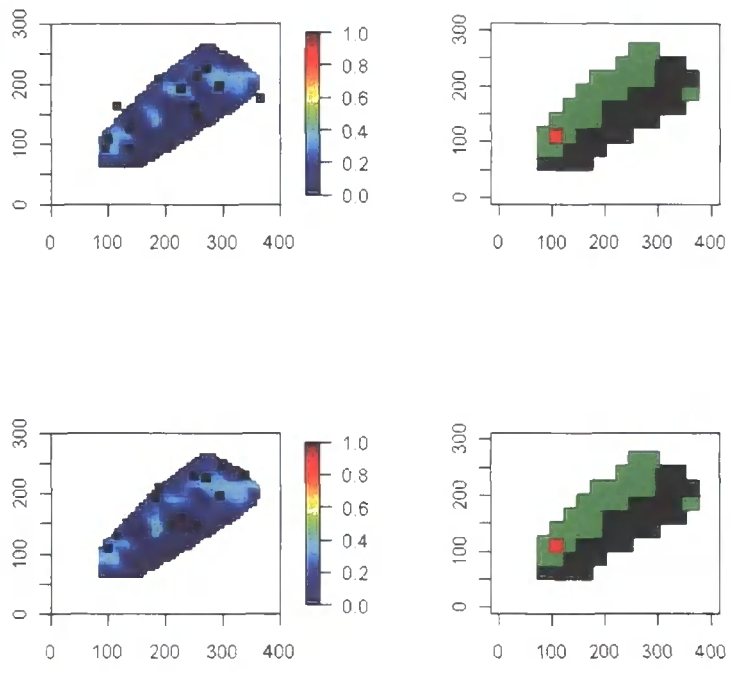


Figure 7.11: Result of the stepwise add and delete search routines for BaP contamination.

most cost effective manner. While some technologies can be used to deal with a wide range of contamination (such as barrier methods, or removal of the affected soil), we envision in this case study that different remediation methods would be required for the in-situ treatment of metals and organics.

When we analyse all four contaminants together we must consider the method of remediation selected. We now effectively have five possible decisions at each inference point: Ex situ remediation (deals with both metals and BaP), in-situ remediation (for metals only), in-situ remediation (for BaP only), in-situ remediation (for both types of contaminant), or the "no remediation" option (MNA). This is not more complicated than a set-up with fewer decisions, it simply involves more decisions to be entered in the algorithm. In general we will need to run the algorithm for longer in this case as we expect the standard error of the expected loss calculation to increase with the number of contaminants. This is where analysing the contaminants using a joint multiple approach could be beneficial, in allowing us to run fewer simulations to achieve the desired accuracy. However, the compu-

tational time associated with an MCMC run at each stage of the search algorithm would still be far longer than running a closed update per contaminant. This is a key issue when considering the practical implementation of this methodology, and a limitation at present. Further work would involve the investigation of a compromise between the methods, or further tweaking of the sample search code to speed the process up for day to day use.

The results of the stepwise add and delete methods are shown in Table 7.7 and Figure 7.12. We see that the immediate decisions taken after updating with the initial data are either the expensive remediation method d_1, or "no remediation" (MNA), d_3. Comparing this image with the figures for the single contaminants shows that this effectively combines the results seen there. The areas where the decisions are changing, as shown on the left images of Figure 7.12, generally cover the regions of uncertainty in the single cases. The sampling locations chosen, from a starting grid of 57 (Figure 7.3), cover much of the site, as we see if we combine the results of the four individual searches. The locations chosen cover the decision boundaries and areas of high combined uncertainty. The combined effect of failure and the cost of implementing two in-situ remediation options leads us to make only the two "extreme" decisions d_1 and d_3. The savings made by the 25 point designs suggested compared to the immediate decision are considerable, £180,464 and £208,110 in the add and delete case respectively. We bear in mind that while we specified remediation costs carefully chosen by an expert, the failure costs were less set in stone, and these figures take into account non-financial factors which have been given a financial cost for representation in this loss set up. These figures are not directly comparable with the single cases, as we used weights of $\frac{1}{3}$ on each metal contaminant, and so have a smaller additive contribution. However, as a rough guide, adding the results of the individual stepwise add calculations, using weights of $\frac{1}{3}$ on the metals, we obtain a value of £6,246,937. This suggests that it is beneficial to perform the multiple search algorithm, even with the computational worries associated with it.

	Imm dec	Final design
Add	4,975,425	4,794,961
Delete	4,975,425	4,767,315

Table 7.7: Results of the add and delete search for all four contaminants

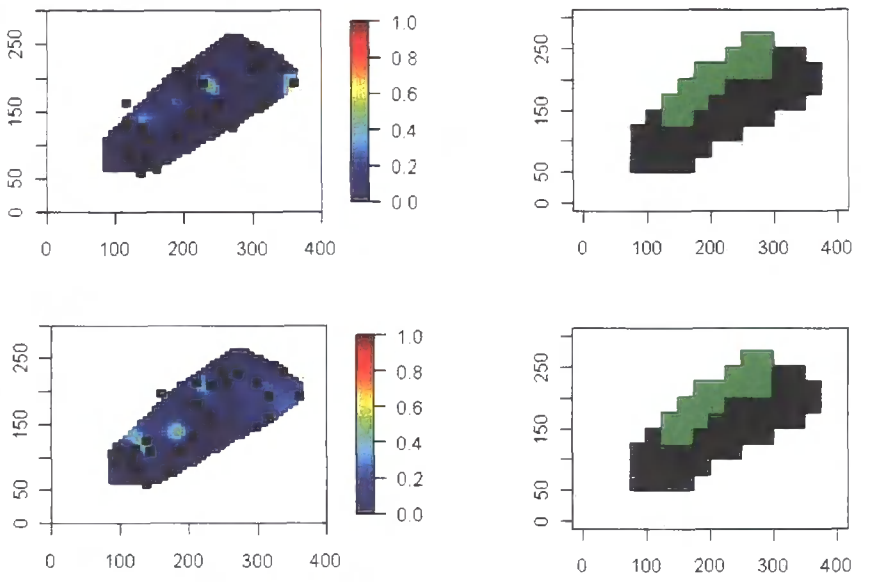


Figure 7.12: Results of the stepwise add and delete search for all four contaminants.

7.4 Comparison to recommendations from original study

The original study does not consider depth or location, just how many values are above the guideline or SSTL. At the time of writing this thesis the actual situation at the site was unresolved. The local authority was undertaking further sampling in order to determine the extent of the contamination at this site, and the neighbouring site to the west, particularly for the Lead levels. The conclusions and recommendations from the site report were:

- Additional desk study will be required, along with further sampling to determine the extent of the contamination.

- All four contaminants of interest failed the MVT in certain regions of the site, and were determined to require further investigation. SSTLs were derived and passed except in the case of Lead and BaP. The Lead contamination was deemed to consist of localised hotspots.

- It is envisaged that remediation will be required to deal with a range of identified SPR linkages. The remediation will likely comprise a number of elements, dependent on the contaminant and location of the contamination. The potential remediation methods discussed included excavation, capping, barrier methods and natural attenuation.

The analyses made in this thesis result in some different conclusions drawn by the site investigation. This is for several reasons. Clearly we are implementing a detailed probabilistic model with which we make statements regarding contamination, whereas the site investigation looks at the suggested statistical tests of [21]. Also, we use the relevant SGV or available critical value as an indicator that remediation is required, while the actual report calculates SSTLs. If we run our analyses with these values, the recommendations regarding immediate decisions become much closer. The SSTLs tend to be less conservative than the SGVs and take into account further site specific aspects about SPR chains. Clearly we can change the critical values and costs in the model with ease, and so this model will have no problem dealing with changing legal definitions in the future. We chose to use the SGVs in this example in order to demonstrate the potential of the methods described. Our analysis offers a second stage sampling design for the investigation of all four contaminants, or just for BaP if the investigators stuck with the SSTL values for the metals. We also give a recommendation for an immediate remediation strategy if no more sampling were to be undertaken, and would aid the DMs in the next stage of the site investigation. Unfortunately, the site in question has been put on hold and so it was not possible to implement this sampling scheme to look at the real outcome.

Some of the main issues arising from this example were the small sampling costs in comparison to the remediation costs and ensuring the standard error was small enough to be confident about the sampling designs selected by the algorithm. If we

could improve the computational speed of the search algorithm in order to allow for a greater number of simulations to be feasible then we could be even more confident about the results.

As we have seen, it is beneficial to learn about the contaminants together in terms of sampling and decision making as it allows us to allocate the budget for the remediation method in one go. The BaP contamination is somewhat dominant in terms of driving the remediation, due to the completely different distribution, and higher occurence of the contaminant compared to the critical value. It would be up to the needs and views of the expert to decide on a site specific basis to determine which contaminants to group together. Given the large amounts of money which are being considered in this example (which were all obtained from discussion with the relevant experts, excluding the largely unknown failure penalties), there is a clear need for a methodology, such as that described within this thesis, which is able to improve the quantification of risk and therefore improve decision making.

Chapter 8

Conclusions

The aim of this thesis was to provide a Bayesian decision framework, in which site investigation experts and decision makers can be included at all relevant stages of the process, in order to guide sampling selection and decision making for remediation options.

The process of site investigation is complex and comprises several, often overlapping, elements. As well as the understanding of several geochemical and engineering processes, along with the collection of detailed information required to build a realistic model accurately representing the spread of contamination through a site, there are several legal and moral issues to deal with. The EA and DEFRA are currently moving through a period of change in terms of the definition and classification of contaminated land, therefore any methodology presented for the problem of site investigation must be flexible and site specific in nature. Also, there is often disagreement and uncertainty in practice regarding who is responsible for the clean up of sites which are deemed contaminated.

We must be sure when specifying loss functions that all consequences are taken into account. As we have mentioned, the site investigator and land owner are clearly looking to spend as little as possible in order to get the land into a safe state for development or resale, whereas the EA and HPA place primary importance on protection of human and environmental health. We have stressed the importance of a balanced approach when determining the loss functions, and feel that this is a key area for development in the future.

We wanted three key features to include in the methodology which would build towards the ultimate aim of efficient sample selection. These are modelling, a decision set up and a sampling search. The methodology developed in this thesis could in theory be applied to a variety of practical problems whereby decisions and further stages of sampling are likely to be required in a spatial setting. It offers a general spatial decision making framework, which we have geared towards site investigation. It has been developed with the analysis of very specific cases in mind, and may be made as complex or basic as desired.

8.1 Modelling for site investigation

The initial stage of the framework involved the development of a probabilistic model for the description of contamination levels over a site. We implemented a full Bayesian approach which allowed for a detailed description of the qualitative and quantitative prior information available, through the expert information, desk study and preliminary sampling. We used a closed model for ease of computation, but also offered a conditional conjugate approach to allow for independence of the mean and variance parameters, as well as multiple contaminant analyses. The multiple analyses proved to be very similar to the individual updates in the cases we considered. This may in part be due to the fact that the data observed matched the prior beliefs well, or that we were obtaining "enough" information from the main contaminant of interest, and the secondary contaminants were not required in the update.

To make the best use of the information we required an effective elicitation procedure. On discussion with the site investigation experts we decided to use a method whereby we asked as few complex questions as possible. The method of colouring maps to gain a depiction of the expert's opinions helped follow the thought process, in order to drive the questioning to further delineate boundaries and obtain detailed correlation structures. This process became further complicated when considering interrelationships between contaminants, although again the use of simple questioning helped the expert. This process is something to be further developed in the future, if the methodology were to be implemented in practice, the

experiences of several elicitation sessions would give further insight into the areas for improvement.

After considering hypothetical examples we looked at a real case study, for which we investigated the distribution of Arsenic, Lead, Zinc, total PAH and BaP. This allowed us to consider the effect of multiple contaminants by analysing the metals together, as well as modelling for an organic contaminant. We considered BaP, as it is a particular contaminant of concern, and it is very different from the metals. As such, it would be expected to occur independently from the metals, as well as spreading in a different way through the site.

Looking at two types of contaminants allowed us to assess the ability of the modelling approach to deal with different contaminant transport methods. To further improve this model, we would envisage two elements of immediate interest. First, to deal with the negative predictions which may arise when we have a small predicted mean and a relatively large standard deviation, we could use a log normal transformation on the data. This could easily be performed in the analysis if the expert believed it would be a more effective way of modelling the contamination at a site. We suggested in the thesis a possible weighting on the variance in order to remove these negative predictions. As an example, the weighting could be based on the value of the predicted mean at a particular location, scaling the variance down to a reasonable level. The model will often include regions of high contamination (and high uncertainty), next to regions where contamination values are much lower, as such it becomes difficult to fully resolve uncertainty even when low values are predicted.

Another improvement would be to find a way to account for hotspots of contamination which we see arising in the case of BaP for example. We could consider adding an extreme value distribution to the model, as a way to include hotspots, and give a more realistic description of the uncertainty. We could have a completely separate hotspot model, and hope that the hotspots are picked up in the main by the initial site walkover and investigation. However, the modelling of hotspots is a complex topic of its own and would require further discussion with experts.

8.2 Decision and sampling for site investigation

The next two, closely linked, stages of the thesis used the model developed in Chapter 3 in order to combine a loss structure and decision theoretic approach to drive a sample selection algorithm. The decision set up allows us to calculate the expected loss associated with a particular sampling design, as well as the loss from implementing an immediate decision where no further sampling is to be undertaken.

We decided to use a stepwise search algorithm to select the "best" set of locations at which to take the next stage of sampling. We looked at stepwise add, delete and combined search algorithms to account for relationships between locations. We developed a quick screening method to determine which locations are likely to be the most informative, so that the stepwise search would be more effective. Removing points that have little or no informative value can be screened out as early as possible, meaning that more computational time can be spent deciding between locations likely to have suggest a change of decision.

We have offered a framework which will assist effective decision making and guide sampling selection. However, at present the methodology presented requires further refinement and development to be realistically feasible for routine use, and to be fully accessible to non-specialists.

8.3 Further work

The spatial distribution of a site is extremely complex, and modelling it raises many issues. In this thesis we have offered a considerable improvement on the previous statistical tests offered by [21], with a model which allows for spatially related observations. However, we would also envisage the inclusion of depth in the model in the future, to further investigate the contamination levels on the site. We could do this with the current model by considering "slices" of the site at depth and considering a number of models concurrently. The inclusion of groundwater modelling would further inform us about the potential spread of contamination. This is currently taken into account implicitly by the expert when specifying the expected directions of trends. We could build on this by including detailed descriptions of the

groundwater flow and inclusion of other geological information.

For this methodology to be useful in everyday site investigations, it has to be offered as a user-friendly piece of software. This could comprise an elicitation tool for the non-specialist to input their beliefs without requiring a statistician on hand for every site investigation.

We have demonstrated the key features of a methodology for the investigation of potentially contaminated land. However, this model is still computationally slow, and quick alternatives need to be implemented to allow a realistically feasible methodology. Ideally a method would allow for the inclusion of an MCMC run at each stage of the iteration, to allow the fullest description of the site and parameters of interest, although as shown this may not be necessary given the similarity of the results obtained. Even the improvement of the computational burden for the closed case would allow for far more simulations to be run, improving the accuracy of results. Also, we could look at extending the functionality of the code, to allow for missing data or observations at different locations for each contaminant.

A key aspect would be to further investigate the actual allocation of loss functions. The loss set-up described within this thesis is extremely flexible and would allow for the specification of loss functions as complex or simple as desired. However, the practicalities of assigning such numbers is a detailed topic which needs further exploration to allow this methodology to be implementable in a meaningful way. As the goalposts move in the subject of contaminated land, and the consequences of making bad decisions change, the allocation of realistic loss functions will further become the key element to determining a best course of action.

Notation

Modelling

$\underline{x}$	Single location p37
x_e, x_n	Location co-ordinates, in east and north direction p62
$\underline{\mathbf{x}}$	Several locations p39
$y(\underline{x})$	Observation at a location p37
$\mathbf{D}$	Correlation matrix of sampling locations, basic model p39
$\epsilon(\underline{x})$	Error term p37
$\mathbf{X}$	Matrix of observation locations p37
θ_i	Correlation length (in direction i) p37
κ	Correlation power p37
$\mathrm{E}[y(\underline{\mathbf{x}})]$	Expert judgement, expected value p46
$\hat{y}(\underline{x})$	Adjusted observations, expert expected value subtracted p46
$\mathbf{G}(\underline{x})$	Expert judgement, function of relative uncertainty p46
$\mathbf{D_G}$	Correlation structure, modified for expert uncertainty judgement p52
$\psi(\underline{x}, \underline{x}', \underline{\theta}, \kappa)$	Correlation structure p38
a_i	Zone weighting for correlation p48
$\underline{\eta}$	Parameter set p51
$\underline{\mathbf{b}}$	Correlation between prediction and sampling locations p55
p	(As a superscript) referring to predictive locations p39
Δ	(As a subscript) referring to observed values, i.e. actual data p39

$y(\underline{x}) = \mathbf{X}\beta + \epsilon(\underline{x})$	Model p37
β	Model mean parameter p38
β_i	Mean value for direction i
Σ	Between contaminant variance
	matrix for multiple contaminant set up p77
σ^2	Model variance p37
a	Inverse Gamma scale p53
d	Inverse Gamma shape p53
$\mathbf{m}$	Normal mean parameter p53
$\mathbf{V}$	Normal variance parameter p53
$e(\underline{x})$	Model nugget variance p44
τ^2	Model nugget variance parameter p44
ν	Relative nugget variance p45
λ	Predictive mean p55
Λ	Predictive variance p55
$(\mathbf{D}_{\mathbf{G}}^{p} + \nu\mathrm{I})$	Prior correlation matrix for the prediction locations p55

Decision and sampling

d_i	Decision i p141
$\mathcal{D}$	Set of all decisions p142
δ	Sampling design p151
Ω	Set of all designs p151
y_δ	Simulated observations for design p152
O_k	Outcome k of a decision d_i p141
u	Utility p141
L	Loss p??
L_R	Cost of remediation method p146
L_F	Failure loss p146
L_{Init}	Initial sampling cost p146
L_{Incr}	Incremental sampling cost p146
n_I	Number of inference locations p139
n_C	Number of candidate locations p146
n_{sim}	Number of simulations p152
$\mathcal{M}$	Expected loss associated with optimal decision p142
$\mathcal{C}$	Total expected loss over inference locations p142
CV	Critical contaminant value p150
UB	Upper bound p167

Glossary of technical terms and abbreviations

EA	Environment Agency
DEFRA	Department for Environment Food and Rural Affairs
HPA	Health Protection Agency
SPOSH	significant possibility of significant harm
MNA	Monitored Natural Attenuation
LPA	Local Planning Authority
SGV	Soil Guideline Values
SSTL	Site specific target level
SSSI	Site of Special Scientific Interest
ICRCL	Interdepartmental Committee on the Redevelopment of Contaminated Land
CLEA	Contaminated Land Exposure Assessment (model)
CLR	Contaminated Land Report
CL:AIRE	Contaminated Land: Applications in Real Environments
CIEH	Chartered Institute of Environmental Health
US EPA	United States Environmental Protection Agency
WFD	(EU) Water Framework Directive
IG	Inverse Gamma distribution
NIG	Normal Inverse Gamma distribution
DM	Decision Maker

Bibliography

[1] S Banerjee, B P Carlin and A E Gelfand, *Hierarchical Modeling and Analysis for Spatial Data*, 2004, Chapman and Hall

[2] M Banjevic and P Switzer, *Optimal Network Designs in Spatial Statistics*, Royal Statistical Society 2001 Conference on Spatial Modelling, http://www.banjevic.com/statistics/networkpaper.pdf

[3] R P Bardos, C Mariotti, F Marot and T Sullivan, *Framework for Decision Support used in Contaminated Land Management in Europe and North America*, Land Contamination and Reclamation, Vol. 9, No. 1, pp. 149-163

[4] J O Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer, 1988

[5] R Boulding, *EPA Environmental Engineering Sourcebook*, United States Environmental Protection Agency, 1996

[6] British Standards Institute, *BS 10175:2001 Investigation of potentially contaminated sites*, 2001

[7] D Bunn, *Applied Decision Analysis*, McGraw-Hill, 1984

[8] G Casella and E I George, *Explaining the Gibbs Sampler*, The American Statistician, Vol. 46, No. 3, 1992, pp. 167-174

[9] S Chib and E Greenberg, *Understanding the Metropolis-Hastings Algorithm*, The American Statistician, Vol. 49, No. 4 (1995), pp. 327-335

[10] CIRIA, *Special Publication 103 - Remedial Treatment for Contaminated Land - Volume III: Site Investigation and Assessment*, 1995

[11] CL:AIRE and CIEH, *Guidance on Comparing Soil Contamination Data with a Critical Concentration*, 2008, www.claire.co.uk

[12] CL:AIRE, *Statistical Assessment of Contaminated Land: Some Implications of the 'Mean Value Test'*, Technical Bulletin 12, www.claire.co.uk

[13] CL:AIRE, *Improving the Reliability of Contaminated Land Assessment using Statistical Methods: Part 1 - Basic Principles and Concepts*, Technical Bulletin 7, www.claire.co.uk

[14] I Clark, *Practical Geostatistics*, Applied Science Publishers, 1979

[15] *Clean-up and regeneration bulletin*, Land contamination and Reclamation, Vol. 14, No. 1, February 2006

[16] D D Cox, L Cox and K B Ensor, *Spatial sampling and the environment:some issues and directions*, Environmental and Ecological Statistics 1997, Vol. 4, pp. 219-233

[17] L A Cox Jr., *Adaptive Spatial Sampling of Contaminated Soil*, Risk Analysis, 1999, Vol. 19, No. 6, pp. 1059-1069

[18] P S Craig, M Goldstein, A H Seheult, J A Smith, *Constructing Partial Prior Specifications for Models of Complex Physical Systems*, The Statistician, 1998, Vol. 47, No. 1, pp. 37-53

[19] N A C Cressie, *Statistics for Spatial Data*, Wiley, 1993

[20] DEFRA and the Environment Agency, *Sampling Strategies for Contaminated Land*, 1994

[21] DEFRA and the Environment Agency, *An Overview of the Development of SGV's and Related Research, R&D Publication CLR7*, 2002, www.environment-agency.gov.uk

[22] DEFRA and the Environment Agency, *Potential contaminants for the assessment of land, R&D Publication CLR8*, 2002, www.environment-agency.gov.uk

[23] DEFRA and the Environment Agency, *Contaminants in soil: Collation of toxicological data and intake values for humans, R&D Publication CLR9*, 2002, www.environment-agency.gov.uk

[24] DEFRA and the Environment Agency, *The Contaminated Land Exposure Assessment (CLEA) model: Technical Basis and Algorithms, R&D Publication CLR10*, 2002, www.environment-agency.gov.uk

[25] DEFRA and the Environment Agency, *Model procedures for the management of Land Contamination*, 2004, www.environment-agency.co.uk

[26] DEFRA and the Environment Agency, *Defra Circular 01/2006 Environmental Protection Act 1990: Part 2A Contaminated Land*

[27] DEFRA website *www.defra.gov.uk*

[28] DEFRA, *Guidance on the Legal Definition of Contaminated Land*, 2008, www.defra.gov.uk

[29] DEFRA, *Guidelines for Environmental Risk Assessment and Management*, 2000, www.defra.gov.uk

[30] DEFRA, *Improvements to Contaminated Land Guidance: Outcome of the "Way Forward" exercise on soil guideline values*, 2008, www.defra.gov.uk

[31] DEFRA landfill guidance *http://www.defra.gov.uk/environment/waste/topics /landfill-dir/pdf/landfilldir.pdf*

[32] M H De Groot, *Optimal Statistical Decisions*, Wiley, 2004

[33] P J Diggle and P J Ribeiro Jr, *Model based geostatistics*, Springer, 2006

[34] P J Diggle and S Lophaven, *Bayesian Geostatistical Design*, Scandinavian Journal of Statistics

[35] The Environmental Act of 1995, Part IIA Contaminated Land, *www.opsi.gov.uk/acts/acts1995/ukpga_19950025_en_6#pt2-l1g57*

[36] The Environment Agency, *Environment Agency Guidance on Requirements for Land Contamination Reports*, July 2005

[37] The Environment Agency website *www.environment-agency.gov.uk*

[38] Environment Agency, *Guidance on the use of soil screening values in ecological risk assessment* Science report, 2008

[39] EU Water Framework Directive, *http://ec.europa.eu/environment/water/water-framework/index_en.html*

[40] Proposed EU Soil Framework Directive, *http://ec.europa.eu/environment/soil/three_en.htm*

[41] EU Landfill directive, *http://ec.europa.eu/environment/waste/landfill_index.htm*

[42] M A Efroymson, *Multiple regression analysis*, Mathematical methods for digital computers, Editors A Ralston and H S Wilf pp. 191-203, Wiley, 1960

[43] M Evans and H Moshonovy, *Checking for Prior-Data Conflict*, Bayesian Analysis Vol. 1, No. 4, 2006, pp. 893-914

[44] C W Fetter, *Contaminant hydrogeology*, Prentice-Hall, 1999

[45] R A Freeze, J Massmann, L Smith, T Sperling and B James, *Hydrogeological decision analysis: 1. A framework*, Ground Water 1990, Vol. 28, pp. 738-766

[46] R A Freeze, B James, J Massmann, T Sperling, L Smith, *Hydrogeological decision analysis: 4. The concept of data worth and its use in the development of site investigation strategy*, Ground Water 1992, Vol. 30, pp. 574-588

[47] J R Gay and A Korre. *A spatially-evaluated methodology for assessing risk to a population from contaminated land*, 2006, Environmental Pollution Vol. 142, pp. 227-234

[48] A Gelman, J B Carlin, H S Stern and D B Rubin, *Bayesian Data Analysis*, Chapman and Hall, 1995

[49] P Goovaerts, *Geostatistics for Natural Resources Evaluation*, Oxford, 1997

[50] J M Hammersley and D C Handscomb, *Monte Carlo Methods*, Wiley, 1964

[51] G A Hardman, *Bayesian Inspection Planning for Large Industrial Systems*, PhD Thesis 2007, Durham University

[52] R Haylock and A O'Hagan, *On Inference for Outputs of Computationally Expensive Algorithms with Uncertainty on the Inputs*, In Bayesian Statistics 5 (eds. J.M Bernardo et al), pp. 629-637, Oxford University Press

[53] R R Hocking, *The analysis and selection of variables in linear regression*, Biometrics 1976, Vol. 32, No. 1, pp. 1-49

[54] J B Kadane and L J Wolfson, *Experiences in Elicitation*, The Statistician, 1998, Vol. 47, No. 1, pp. 3-19

[55] M Kamon, *Remediation of Contaminated Land*, Journal of Disaster Research, 2007, Vol. 2, No. 3 pp. 173-189

[56] R L Keeney and H Raiffa, *Decisions with multiple objectives: preferences and value tradeoffs*, 1993, Cambridge University Press

[57] D G Krige, *A statistical approach to some basic mine valuation problems on the Witwatersrand*, Journal of the Chemical, Metallurgical and Mining Soceity of South Africa, Vol. 52, No. 6, pp. 119-139

[58] M D LaGrega et al, *Hazardous Waste Management*, 1994, McGraw-Hill

[59] D V Lindley, *Making Decisions*, 1985, Wiley

[60] J A Lyn, M H Ramsey, A P Damant and R Wood, *Two-stage application of the optimised uncertainty method: a practical assessment*, Analyst, 2005, Vol. 130, pp. 1271-1279

[61] I Maqsood, J Li, G Huang and Y Huang, *Simulation-based risk assessment of contaminated sites under remediation scenarios, planning periods, and land-use patterns-a Canadian case study*, Stochastic Environmental Research and Risk Assessment, 2005, Vol. 19, pp. 146-157

[62] K V Mardia, J Kent and J Bibby, *Multivariate Analysis*, Academic Press, 1979

[63] J Massmann, R A Freeze, L Smith, T Sperling and B James, *Hydrogeological decision analysis: 2. Applications to ground-water contamination*, Ground Water 1991, Vol. 29, pp. 536-548

[64] G Matheron, *Principles of geostatistics*, Economic Geology, 1963, Vol.58, pp. 1246-1266

[65] C P Nathanail, *The use and misuse of CLR 7 acceptance tests for assessment of risks to human health from contaminated land*, Quarterly Journal of Engineering Geology and Hydrogeology, 2004, Vol. 37, No. 5, pp. 361-367

[66] T Öberg and B Bergbäck, *A Review of Probabilistic Risk Assessment of Contaminated Land*, Journal of Soils and Sediments, 2001, Vol. 5, No. 4, pp. 213-224

[67] K Paleologos and I Lerche, *Multiple decision-making criteria in the transport and burial of hazardous and radioactive wastes*, Stochastic Environmental Research and Risk Assessment, 1999, Vol. 13, pp. 381-395

[68] A O'Hagan, *Eliciting Expert Beliefs in Substantial Practical Applications*, The Statistician, 1998, Vol. 47, No. 1, pp. 21-35

[69] A O'Hagan, *Kendall's Advanced Theory of Statistics, Vol 2B: Bayesian Inference*, Edward Arnold, 1994

[70] H Omre, *Bayesian kriging - merging observations and qualified guesses in kriging*, Mathematical Geology, 1987, Vol.19, pp. 25-38

[71] H Omre, *A Bayesian approach to kriging*, in Geostatistics, Vol. I, pp. 109-126, 1989

[72] R Development Core Team, *R: A language and environment for statistical computing*, http://www.R-project.org, 2007

[73] M H Ramsey and S L R Ellison (eds.), *Eurachem/ EUROLAB/ CITAC/ Nordtest/ AMC Guide: Measurement uncertainty arising from sampling: a guide to methods and approaches*, Eurachem, 2007

[74] P J Ribeiro Jr., *Technical Report ST-99-08, Bayesian inference in Gaussian model-based geostatistics*, Lancaster University

[75] P J Ribeiro Jr., *Model based geostatistics, applications and computational implementation*, PhD thesis, Lancaster University, 2002

[76] B D Ripley, *Spatial Statistics*, Wiley, 1981

[77] J A Royle and D Nychka, *An Algorithm for the Construction of Spatial Coverage Designs with Implementation in SPLUS*, Computers and Geosciences, Vol. 24, No. 5, pp. 479-488

[78] R W Scholz and U Schnabel, *Decision making under uncertainty in case of soil remediation*, Journal of Environmental Management 2006, Vol. 80, pp. 132-147

[79] T Sperling, R A Freeze, J Massmann, L Smith and B James, *Hydrogeological decision analysis: 3. Application to dewatering design at an open pit mine*, Ground Water, 1992, Vol. 30, pp. 376-389

[80] M L Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, 1999, Springer

[81] P D Taylor et al, *Assessing the reliability and the cost-effectiveness of routine site investigation*, Ground Engineering, August 2007, pp. 34-37

[82] US EPA website, *http://www.epa.gov/oswer/riskassessment/risk_superfund.htm*

[83] US EPA, *Guidance for Performing Preliininary Assessments Under CERCLA*, 1991

[84] US EPA, *Guidance for Performing Site Inspections Under CERCLA*, 1992

[85] US EPA, *A Citizen's Guide to Monitored Natural Attenuation*, http://www.clu-in.org/download/citizens/mna.pdf

[86] US EPA, *Visual Sampling Plan*

[87] J W van Groenigen, W Siderius and A Stein, *Contrained optimisation of soil sampling for minimisation of the kriging variance*, Geoderma, 1999, Vol. 87, pp. 239-259

[88] R Webster and M A Oliver, *Statistical Methods in Soil and Land Resource Survey*, Oxford University Press, 1990

Appendix A

Basic and Auxiliary Results

A.1 Standard Results

A.1.1 Bayes Theorem

The underlying statement which forms the basis for the probabilistic modelling and updating of beliefs is the statement of Bayes Theorem.

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)} \tag{A.1}$$

where

- $p(A)$ is the prior probability for A, i.e the quantification of the experts prior beliefs.
- $p(B \mid A)$ is the likelihood, i.e. the probability of observing B given the distribution for A.
- $p(A \mid B)$ is the posterior probability for A, conditioned on the value of B that was observed.
- $p(B)$ is the marginal probability for B

A.1.2 Closed Update

Conjugate analysis of the normal linear model with generalized error variance, from [69], Normal Inverse Gamma $(a, d, \mathbf{m}, \mathbf{V})$ joint prior specification

$$\begin{aligned}
\mathbf{V}^* &= (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1} \\
\mathbf{m}^* &= (\mathbf{V}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{V}^{-1}\mathbf{m} + \mathbf{X}'y) \\
a^* &= a + \mathbf{m}'\mathbf{V}^{-1}\mathbf{m} + y'y - (\mathbf{m}^*)'(\mathbf{V}^*)^{-1}\mathbf{m}^* \\
d^* &= d + m
\end{aligned}$$

Adjusting to allow for the extra notation and the correlated error structure,

$$\text{Posterior} \propto (\sigma^2)^{-(d+m+q+2)/2} \exp\left(-\frac{1}{2\sigma^2}\right.$$

$$\left.\times \left(\underbrace{(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta})'(\mathbf{D_G} + \nu\mathrm{I})^{-1}(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}) + (\underline{\beta} - \mathbf{m})'\mathbf{V}^{-1}(\underline{\beta} - \mathbf{m}) + a}_{\mathbf{Q}}\right)\right)$$

$$\begin{aligned}
\text{Let } \mathbf{R} &= (\mathbf{D_G} + \nu\mathrm{I}) \\
\text{Then } \mathbf{Q} &= \hat{y}_\Delta(\underline{\mathbf{x}})'\mathbf{R}^{-1}\hat{y}_\Delta(\underline{\mathbf{x}}) - \underline{\beta}'\mathbf{X}'\mathbf{R}^{-1}\hat{y}_\Delta(\underline{\mathbf{x}}) - \hat{y}_\Delta(\underline{\mathbf{x}})'\mathbf{R}^{-1}\mathbf{X}\underline{\beta} + \underline{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\underline{\beta} \\
&+ \underline{\beta}'\mathbf{V}^{-1}\underline{\beta} - \mathbf{m}'\mathbf{V}^{-1}\underline{\beta} - \underline{\beta}'\mathbf{V}^{-1}\mathbf{m} + \mathbf{m}'\mathbf{V}^{-1}\mathbf{m} + a \\
&= \underline{\beta}'\left(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} + \mathbf{V}^{-1}\right)\underline{\beta} - \underline{\beta}'\left(\mathbf{X}'\mathbf{R}^{-1}\hat{y}_\Delta(\underline{\mathbf{x}}) + \mathbf{V}^{-1}\mathbf{m}\right) \\
&- \left(\hat{y}_\Delta(\underline{\mathbf{x}})'\mathbf{R}^{-1}\mathbf{X} + \mathbf{m}'\mathbf{V}^{-1}\right)\underline{\beta} + \mathbf{m}'\mathbf{V}^{-1}\mathbf{m} + \hat{y}_\Delta(\underline{\mathbf{x}})'\mathbf{R}^{-1}\hat{y}_\Delta(\underline{\mathbf{x}}) + a \\
&= (\underline{\beta} - \mathbf{m}^*)'(\mathbf{V}^*)^{-1}(\underline{\beta} - \mathbf{m}^*) + a^*
\end{aligned}$$

A.1.3 Standard results for the multivariate normal distrribution

If we have $\mathbf{x} \sim \mathrm{N}(\mu, \Sigma)$ and we partition $\mathbf{x}$ into two pieces $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2})$ then

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right)$$

Marginalisation

$$\mathbf{x_1} \sim \mathrm{N}(\mu_1, \Sigma_{11})$$

Conditioning

$$\mathbf{x_2} \mid \mathbf{x_1} \sim \mathrm{N}\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathrm{x}_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right) \tag{A.2}$$

A.2 Calculation of D

We give an explicit calculation for the **D** matrix presented in Equation (3.20). As we have three regions specified, we require the calculation of three intermediate "**D**" matrices. This are found using Equations (3.4, 3.5) as

$$\mathbf{D}_1 = \begin{pmatrix} 1 & 0.486 & 0.493 \\ 0.486 & 1 & 0.729 \\ 0.493 & 0.729 & 1 \end{pmatrix}$$

$$\mathbf{D}_2 = \begin{pmatrix} 1 & 0.236 & 0.243 \\ 0.236 & 1 & 0.531 \\ 0.243 & 0.531 & 1 \end{pmatrix}$$

$$\mathbf{D}_3 = \begin{pmatrix} 1 & 0.001 & 0.001 \\ 0.001 & 1 & 0.042 \\ 0.001 & 0.042 & 1 \end{pmatrix}$$

We combine these to give **D** using Equation (3.18). We show the calculation for the $[1,2]^{th}$ element of the matrix here, where the weights come from Table 3.1

$$\begin{aligned}
\mathbf{D}[1,2] &= \sum_{k=1}^{3} a_{r_k}(\underline{x}_1)a_{r_k}(\underline{x}_2)\mathbf{D}_{r_k}[1,2] \\
&= a_{r_1}(\underline{x}_1)a_{r_1}(\underline{x}_2)\mathbf{D}_{r_1}[1,2] \\
&+ a_{r_2}(\underline{x}_1)a_{r_2}(\underline{x}_2)\mathbf{D}_{r_2}[1,2] \\
&+ a_{r_3}(\underline{x}_1)a_{r_3}(\underline{x}_2)\mathbf{D}_{r_3}[1,2] \\
&= 0.975 \times 0.15 \times 0.486 + 0.2 \times 0.15 \times 0.236 + 0.1 \times 0.977 \times 0.0007 = 0.07825
\end{aligned}$$

A.3 Conditional distributions for the MCMC algorithm

A.3.1 Single contaminant

Independent prior specifications, and likelihood

$$\begin{aligned} p(\underline{\beta}) &\sim \mathrm{N}(\mathbf{m}, \mathbf{V}) \\ p(\sigma^2) &\sim \mathrm{IG}\left(\frac{a}{2}, \frac{d}{2}\right) \\ p(\nu) &\sim \mathrm{IG}(a_\nu, d_\nu) \\ p(\hat{y}_\Delta(\underline{\mathbf{x}}) \mid \underline{\beta}, \sigma^2, \nu) &\sim \mathrm{N}\left(\mathbf{X}\underline{\beta}, \sigma^2(\mathbf{D_G} + \nu\mathrm{I})\right) \end{aligned}$$

We combine these to give the form of the posterior, and then individually condition on one parameter at a time, assuming all other parameters are known:

$$\begin{aligned} p(\underline{\beta}, \sigma^2, \nu \mid \hat{y}_\Delta(\underline{\mathbf{x}})) &\propto \frac{\frac{a}{2}^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \sigma^{2\left(\frac{-d+2}{2}\right)} \exp\left(\frac{-a}{2\sigma^2}\right) \frac{a_\nu^{d_\nu}}{\Gamma(d_\nu)} \nu^{-d_\nu+1} \exp\left(\frac{-a_\nu}{\nu}\right) \\ &\times \frac{1}{2\pi} |\mathbf{V}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\underline{\beta} - \mathbf{m}\right)' \mathbf{V}^{-1} \left(\underline{\beta} - \mathbf{m}\right)\right) \\ &\times \frac{1}{(2\pi)^{m/2}} |\sigma^2(\mathbf{D_G} + \nu\mathrm{I})|^{-\frac{1}{2}} \\ &\times \exp\left(-\frac{1}{2}\left(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}\right)' \sigma^{2(-1)} (\mathbf{D_G} + \nu\mathrm{I})^{-1} \left(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}\right)\right) \end{aligned}$$

$$\begin{aligned} p(\underline{\beta} \mid \sigma^2, \nu, \hat{y}_\Delta(\underline{\mathbf{x}})) &\propto \exp\left(-\frac{1}{2}\left(\underline{\beta} - \mathbf{m}\right)' \mathbf{V}^{-1} \left(\underline{\beta} - \mathbf{m}\right)\right) \\ &\times \exp\left(-\frac{1}{2}\left(\left(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}\right)' \sigma^{2(-1)} \left(\mathbf{D_G} + \nu\mathrm{I}\right)^{-1} \left(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}\right)\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\underline{\beta}'\mathbf{V}^{-1}\underline{\beta} + \underline{\beta}'\mathbf{X}'\frac{(\mathbf{D_G} + \nu\mathrm{I})^{-1}}{\sigma^2}\mathbf{X}\underline{\beta} - \mathbf{m}'\mathbf{V}^{-1}\underline{\beta}\right.\right. \\ &- \hat{y}_\Delta(\underline{\mathbf{x}})'\frac{(\mathbf{D_G} + \nu\mathrm{I})^{-1}}{\sigma^2}\mathbf{X}\underline{\beta} - \underline{\beta}'\mathbf{V}^{-1}\mathbf{m} \\ &\left.\left. - \underline{\beta}'\mathbf{X}'\frac{(\mathbf{D_G} + \nu\mathrm{I})^{-1}}{\sigma^2}\hat{y}_\Delta(\underline{\mathbf{x}}) + \cancel{\mathbf{m}'\mathbf{V}^{-1}\mathbf{m}} + \cancel{\hat{y}_\Delta(\underline{\mathbf{x}})'(\mathbf{D_G} + \nu\mathrm{I})^{-1}\hat{y}_\Delta(\underline{\mathbf{x}})}\right)\right) \\ &\sim \mathrm{N}\left(\mathbf{m}^*, \mathbf{V}^*\right) \end{aligned}$$

where

$$\begin{aligned}\mathbf{V}^* &= \left(\mathbf{V}^{-1} + \mathbf{X}'\frac{(\mathbf{D_G}+\nu\mathrm{I})^{-1}}{\sigma^2}\mathbf{X}\right)^{-1} \\ \mathbf{m}^* &= \mathbf{V}^*\left(\mathbf{V}^{-1}\mathbf{m} + \mathbf{X}'\frac{\mathbf{D_G}+\nu\mathrm{I}}{\sigma^2}\hat{y}_\Delta(\underline{\mathbf{x}})\right)\end{aligned}$$

$$\begin{aligned}p(\sigma^2 \mid \underline{\beta}, \nu, \hat{y}_\Delta(\underline{\mathbf{x}})) &\propto \sigma^{2(-\frac{d+2}{2})}\exp\left(-\frac{a}{2\sigma^2}\right) \\ &\times \sigma^{2(-\frac{n_1}{2})}\exp\left(-\frac{1}{2\sigma^2}\left(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}\right)'\left(\mathbf{D_G}+\nu\mathrm{I}\right)^{-1}\left(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}\right)\right) \\ &\sim \mathrm{IG}\left(\frac{a^*}{2}, \frac{d^*}{2}\right)\end{aligned}$$

where

$$\begin{aligned}a^* &= a + \left(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}\right)'\left(\mathbf{D_G}+\nu\mathrm{I}\right)^{-1}\left(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}\right) \\ d^* &= d + m\end{aligned}$$

Finally we require the conditional distribution for ν given the other parameters. Here we find that we cannot express this as a closed form distribution and so the Metropolis step is required.

$$\begin{aligned}p(\nu \mid \underline{\beta}, \sigma^2, \hat{y}_\Delta(\underline{\mathbf{x}})) &\propto \nu^{-(d_\nu+1)}\exp\left(\frac{a_\nu}{\nu}\right)\left|\mathbf{D_G}+\nu\mathrm{I}\right|^{-\frac{1}{2}} \\ &\times \exp\left(-\frac{1}{2}\left(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}\right)'\sigma^{2(-1)}\left(\mathbf{D_G}+\nu\mathrm{I}\right)^{-1}\left(\hat{y}_\Delta(\underline{\mathbf{x}}) - \mathbf{X}\underline{\beta}\right)\right)\end{aligned}$$

A.3.2 Multiple contaminants

In Equation (3.48) we gave the form of the prior distribution for the multiple contaminant case. Here, we look at the $\underline{\beta}$ specification and give the form of the conditional which is required in order to calculate the posterior $\underline{\beta}_i$ distributions. We use the results of A.1.3.

$$\begin{aligned}p(\underline{\beta}_i \mid \underline{\beta}_j, \underline{\beta}_k) &\sim \mathrm{N}\left(\mu_i^*, \Sigma_i^*\right) \\ \mu_i^* &= \mathbf{m}_i + \begin{pmatrix}\Sigma_{ij} & \Sigma_{ik}\end{pmatrix}\begin{pmatrix}\Sigma_{jj} & \Sigma_{jk} \\ \Sigma_{kj} & \Sigma_{kk}\end{pmatrix}^{-1}\left[\begin{pmatrix}\underline{\beta}_j \\ \underline{\beta}_k\end{pmatrix} - \begin{pmatrix}\mathbf{m}_j \\ \mathbf{m}_k\end{pmatrix}\right] \\ \Sigma_i^* &= \Sigma_{ii} - \begin{pmatrix}\Sigma_{ij} & \Sigma_{ik}\end{pmatrix}\begin{pmatrix}\Sigma_{jj} & \Sigma_{jk} \\ \Sigma_{kj} & \Sigma_{kk}\end{pmatrix}^{-1}\begin{pmatrix}\Sigma_{ji} \\ \Sigma_{ki}\end{pmatrix}\end{aligned}$$

We then derive the forms of each conditional posterior distribution as shown in the previous section. However, where we see $\mathbf{m}$ and $\mathbf{V}$ on page 250, we use μ_i^* and $\mathbf{W}_i^*$ respectively. This leads us to the conditional posterior shown on page 79.

We may generalise this to any number of contaminants by partitioning the vector $\underline{\beta}_1, \underline{\beta}_2, \ldots, \underline{\beta}_p$ and the associated variance matrix as follows (where $2 < i < p-1$)

$$\underline{\beta}_{\backslash i} = \begin{pmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \\ \vdots \\ \underline{\beta}_{i-1} \\ \underline{\beta}_{i+1} \\ \vdots \\ \underline{\beta}_p \end{pmatrix} \mathbf{m}_{\backslash i} = \begin{pmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \vdots \\ \mathbf{m}_{i-1} \\ \mathbf{m}_{i+1} \\ \vdots \\ \mathbf{m}_p \end{pmatrix} \Sigma = \left(\begin{array}{c|c} \Sigma_{ii} & \Sigma_{i\backslash i} \\ \hline \Sigma_{\backslash ii} & \Sigma_{\backslash i\backslash i} \end{array}\right) \mathbf{W}_i = \Sigma_{ii} - \Sigma_{i\backslash i}\Sigma_{\backslash i\backslash i}^{-1}\Sigma_{\backslash ii}$$

where we define

$$\begin{aligned}
\Sigma_{ii} &= \Sigma_{ii} \\
\Sigma_{i\backslash i} &= \begin{pmatrix} \Sigma_{i1} & \Sigma_{i2} & \ldots & \Sigma_{i(i-1)} & \Sigma_{i(i+1)} & \ldots & \Sigma_{ip} \end{pmatrix} \\
\Sigma_{\backslash ii} &= \begin{pmatrix} \Sigma_{1i} \\ \Sigma_{2i} \\ \vdots \\ \Sigma_{(i-1)i} \\ \Sigma_{(i+1)i} \\ \ldots \\ \Sigma_{pi} \end{pmatrix} \\
\Sigma_{\backslash i\backslash i} &= \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \ldots & \Sigma_{1(i-1)} & \Sigma_{1(i+1)} & \ldots & \Sigma_{1p} \\ \Sigma_{21} & \Sigma_{22} & \ldots & \Sigma_{2(i-1)} & \Sigma_{2(i+1)} & \ldots & \Sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \Sigma_{(i-1)1} & \Sigma_{(i-1)2} & \ldots & \Sigma_{(i-1)(i-1)} & \Sigma_{(i-1)(i+1)} & \ldots & \Sigma_{(i-1)p} \\ \Sigma_{(i+1)1} & \Sigma_{(i+1)2} & \ldots & \Sigma_{(i+1)(i-1)} & \Sigma_{(i+1)(i+1)} & \ldots & \Sigma_{(i+1)p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \Sigma_{p1} & \Sigma_{p2} & \ldots & \Sigma_{p(i-1)} & \Sigma_{p(i+1)} & \ldots & \Sigma_{pp} \end{pmatrix}
\end{aligned}$$

and the posterior conditional $\underline{\beta}_i$ distribution becomes

$$
\begin{aligned}
p(\underline{\beta}_i \mid \underline{\beta}_{\setminus i}, \sigma^2, \nu) &\sim \mathrm{N}(\mathbf{m}_i^*, \mathbf{V}_i^*) \\
\mathbf{m}_i^* &= \mathbf{V}_i^* \left(\frac{\mathbf{X}'(\mathbf{D_G} + \nu_i \mathrm{I})^{-1} \hat{y}_{\Delta i}(\underline{\mathbf{x}})}{\sigma_i^2} + \mathbf{W}_i^{-1} \mathbf{m}_i \right. \\
&+ \left. \mathbf{W}_i^{-1} \left[\Sigma_{i\setminus i} \Sigma_{\setminus i \setminus i}^{-1} \underline{\beta}_{\setminus i} \right] - \mathbf{W}_i^{-1} \left[\Sigma_{i\setminus i} \Sigma_{\setminus i \setminus i}^{-1} \mathbf{m}_{\setminus i} \right] \right) \\
\mathbf{V}_i^* &= \left(\frac{\mathbf{X}'(\mathbf{D_G} + \nu_i \mathrm{I})^{-1} \mathbf{X}}{\sigma_i^2} + \mathbf{W}_i^{-1} \right)^{-1}
\end{aligned}
$$

Appendix B

Additional examples

B.1 Lead contamination

The elicitation and update demonstration for Lead is included in this appendix as the methods are very similar to the Zinc case included in the main text, Section 4.2.3. We will use this information when considering the design of sampling schemes in Chapter 7.

B.1.1 Prior specification

Due to the presence of the leadworks previously to the left of the site, the expert was fairly confident that Lead would be present in elevated quantities to the west. However, unlike the Arsenic, she didn't think there would be any to the east, so the site was only divided into four regions. After further discussion about her uncertainty regarding the blue section, the expert stated that she expected a soft boundary between the blue and green regions. We describe this by adding a transitional region, as discused in Section 3.4.2 with expected value ranging from the mean for the blue region in the west, through to the green mean in the east. After discussion with the expert, we opt for a region approximately equal in width to the blue region as shown in Figure B.1.

Based on the uncertainty the expert expressed we assigned a prior variance of $\sigma^2 = 200^2$ and a zero prior linear $\underline{\beta}$ as we have a detailed regional specification, so $\underline{\beta}$ learns about the departures from this description. Tables B.1 and B.2 gives the

	a	d	$\mathbf{m}$	$\mathbf{V}$
Closed prior	40000	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{pmatrix}$
Closed posterior	428199.1	17	$\begin{pmatrix} 286.01 \\ -0.35 \\ -1.80 \end{pmatrix}$	$\begin{pmatrix} 0.906 & -0.002 & -0.002 \\ -0.002 & 0.000 & -0.000 \\ 0.002 & -0.000 & 0.000 \end{pmatrix}$
Cond conj prior	40000	3	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 200000 & 0 & 0 \\ 0 & 200000 & 0 \\ 0 & 0 & 200000 \end{pmatrix}$
Cond conj posterior	504723.1	17	$\begin{pmatrix} 300.94 \\ -0.39 \\ -1.84 \end{pmatrix}$	$\begin{pmatrix} 31399.50 & -68.67 & -80.23 \\ -68.67 & 0.62 & -0.36 \\ -80.24 & -0.36 & 0.87 \end{pmatrix}$

Table B.1: Prior and posterior hyperparameters in the closed and conditional conjugate case for Lead

means, uncertainties and correlation lengths and weights used in the update, where

Region	E($y(\underline{\mathbf{x}})$)	G	$\underline{\theta}$	κ	$\underline{x} \in$	1	2	3	4	5
1	45000	5	2	5	$a_1(\underline{x})$	0.975	0.220	0	0	0
2	4500	3	2	100	$a_2(\underline{x})$	0.222	0.95	0.192	0.055	0
3	*smooth*	2	2	100	$a_3(\underline{x})$	0	0222	0.975	0.153	0.222
4	300	0.5	2	5	$a_4(\underline{x})$	0	0	0.05	0.975	0.220
5	450	1	2	20	$a_5(\underline{x})$	0	0	0.1	0.153	0.95

Table B.2: Prior specification for Lead contamination at Site R

smooth is a function decreasing linearly from left to right, from 4500 to 450.

$$smooth = 4500 - 4500\left(\frac{x_e - 75}{45}\right) \tag{B.1}$$

Updating, cross validation and diagnostics

We carry out the same procedure as for the Arsenic example to determine how the model is performing. Figure B.3 shows the predicted values and standard devia-

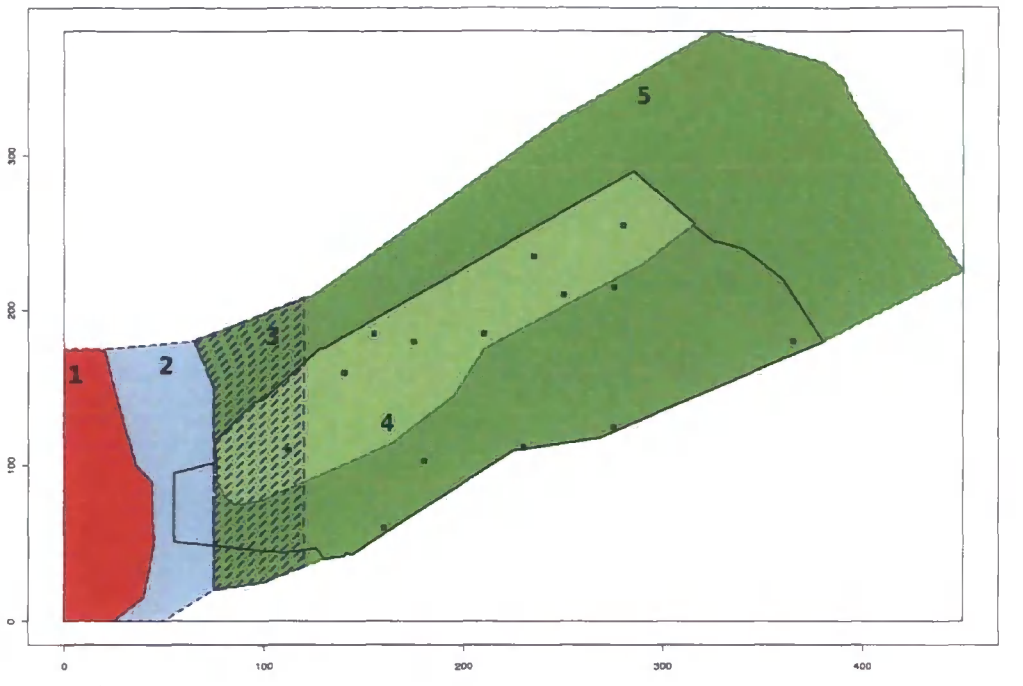


Figure B.1: Extent of transitional region to account for expert uncertainty regarding possible spread of Lead contamination

tions for the same herringbone grid as previously. Also included in Figure B.2 is a depiction of the contamination levels with the large values cut off so the predicted behaviour throughout the rest of the site is clearer. In the closed case σ^2 was reduced from 40000 to 28546, (standard deviation from 200 to 169) and to 33648 (standard deviation of 183) in the MCMC update. The cross validation results in Table B.3 and Figure B.4 shows that most the predicted values are within two standard deviations, and many within one. While not as close as the Arsenic cross validated values, there is no indicator of the prior being unreasonable. The values that do poorest again have few influential locations around them, or are close to boundary regions, which suggests that we are uncertain in these regions, the sampler should choose to look here for further investigation given a suitable decision criterion. Again there is no discernable difference between the closed and MCMC approach, except the time taken to perform the updates.

The problem of large uncertainty and small means is more noticeable in this case due to the larger numbers being considered. However, we feel the model accurately

describes the contamination in the site as a whole

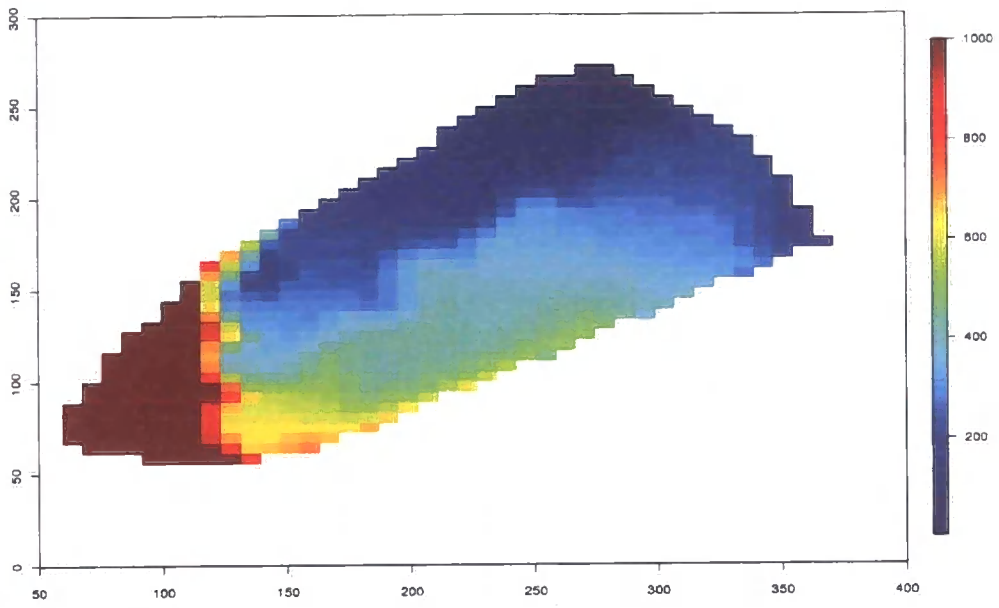


Figure B.2: Prediction for Lead contamination with all values over 1000 cut off at 1000 to better depict the prediction levels in the rest of the site.

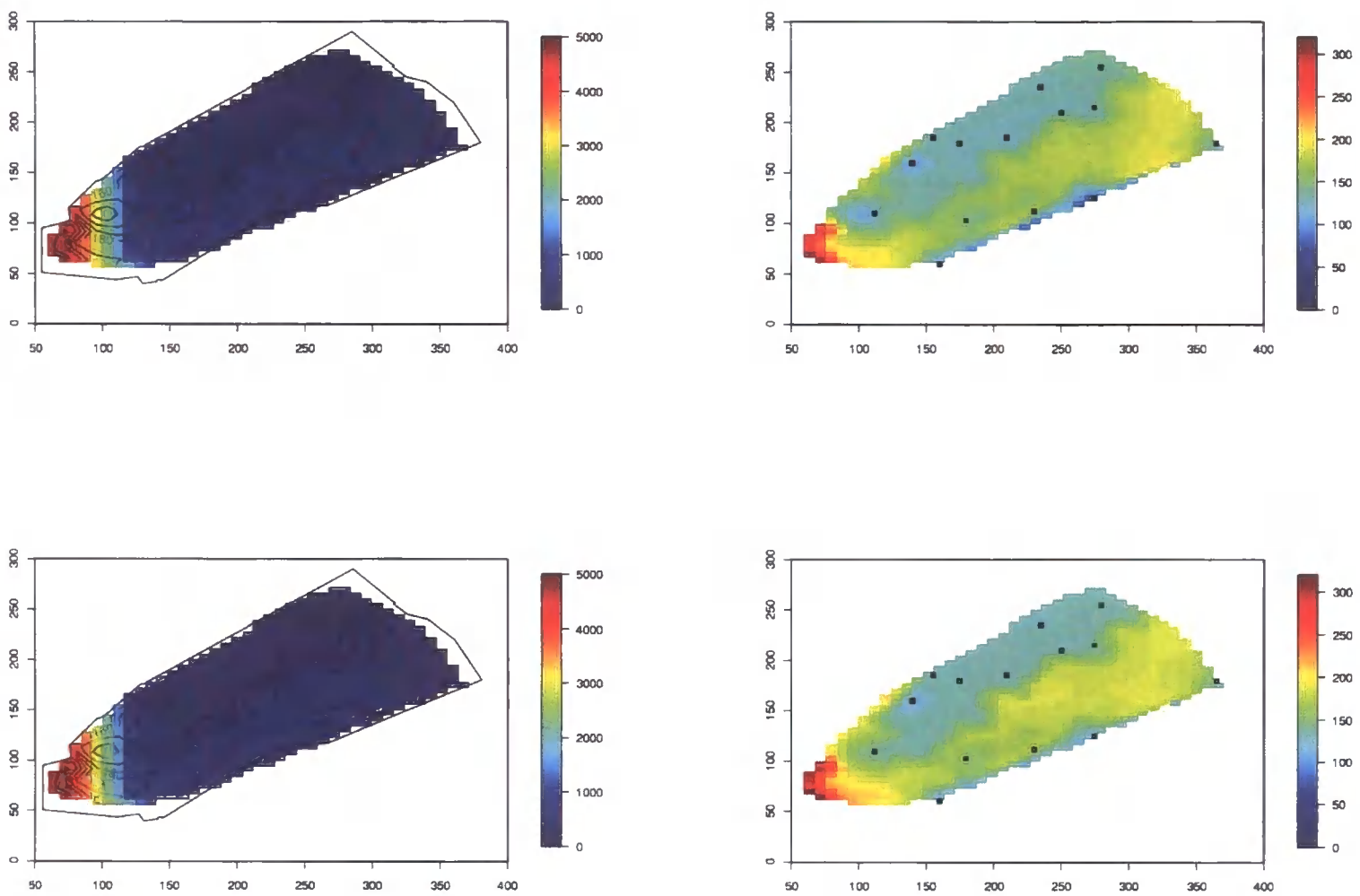


Figure B.3: Lead posterior predictive distributions, comparison of methods. Closed update, mean with standard deviation overlaid in top left image, top right image shows standard deviation. Single MCMC approach, mean with standard deviation overlaid in bottom left image, bottom right image shows standard deviation.

		Closed			Cond conj		
Location	Value	Pred mean	Pred sd	Standardised	Pred mean	Pred sd	Standardised
(160,60)	770	523.0	197.7	1.25	525.67	185.34	1.32
(230,112)	390	484.5	192.6	-0.49	483.69	191.52	-0.49
(275,125)	525	381.7	198.8	0.72	384.83	191.83	0.73
(365,180)	67.5	440.1	212.2	-1.76	437.82	178.48	-2.07
(112,110)	1600	1184.7	227.7	1.82	1197.15	247.23	1.63
(140,160)	51	329.6	122.8	-2.27	338.5	112.8	-2.55
(180,103)	370	549.0	181.9	-0.98	542.93	185.68	-0.93
(275,215)	210	268.9	184.4	-0.32	270.35	194.1	-0.31
(155,185)	150	226.8	141.3	-0.54	226.78	136.17	-0.56
(175,180)	300	193.0	133.0	0.80	193.49	135.96	0.78
(210,185)	270	175.5	130.2	0.73	177.23	136.81	0.68
(235,23)	35	98.7	141.1	-0.45	98.01	138.09	-0.46
(250,210)	290	95.3	124.7	1.56	94.66	126.96	1.54
(280,255)	43	22.4	152.3	0.14	17.79	137.67	0.18

Table B.3: Cross validation for Lead observations

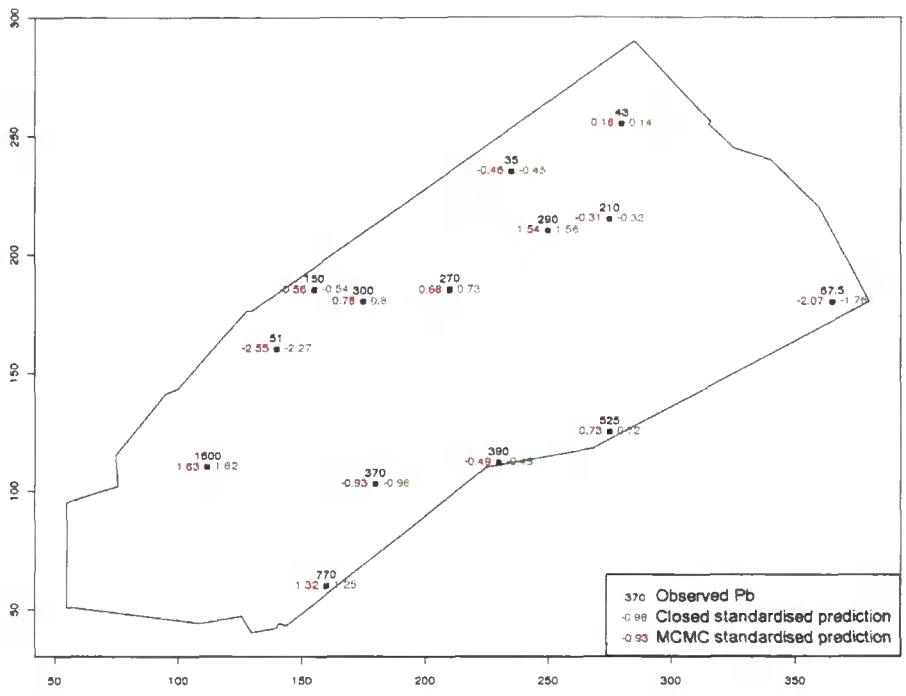


Figure B.4: Cross validation results: Standardised prediction values for Lead at observation locations, for closed update in green and MCMC in red

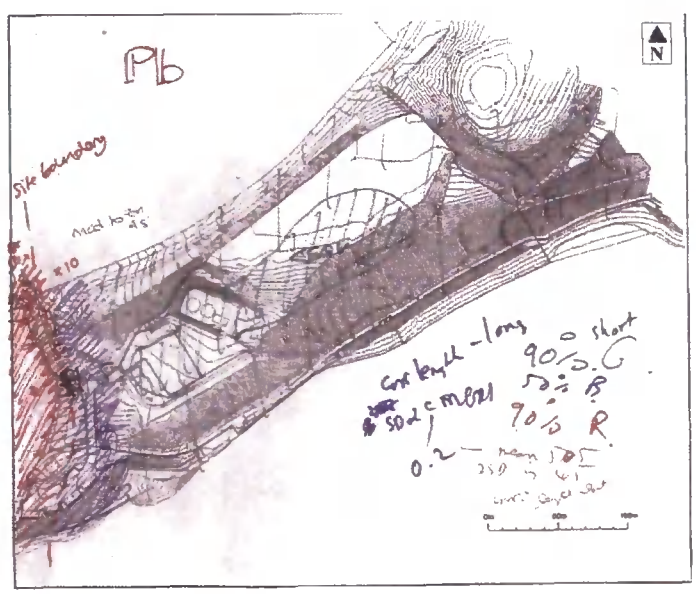


Figure B.5: Expert belief elicitation maps, Lead

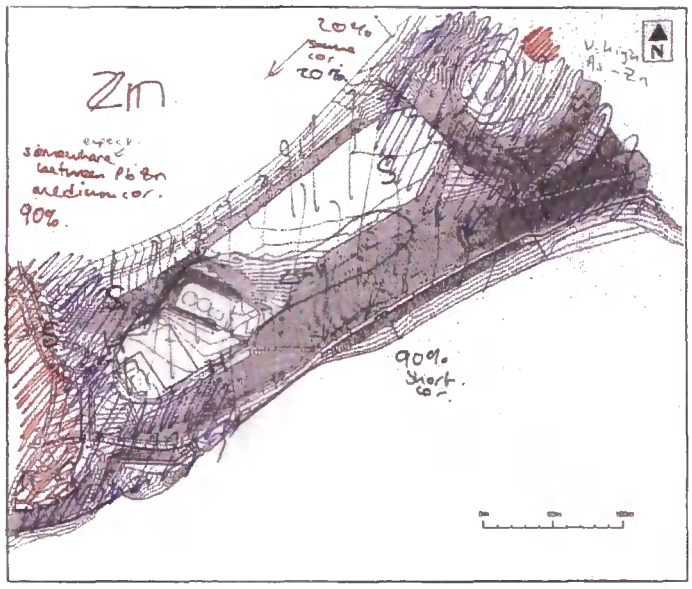


Figure B.6: Expert belief elicitation maps, Zinc

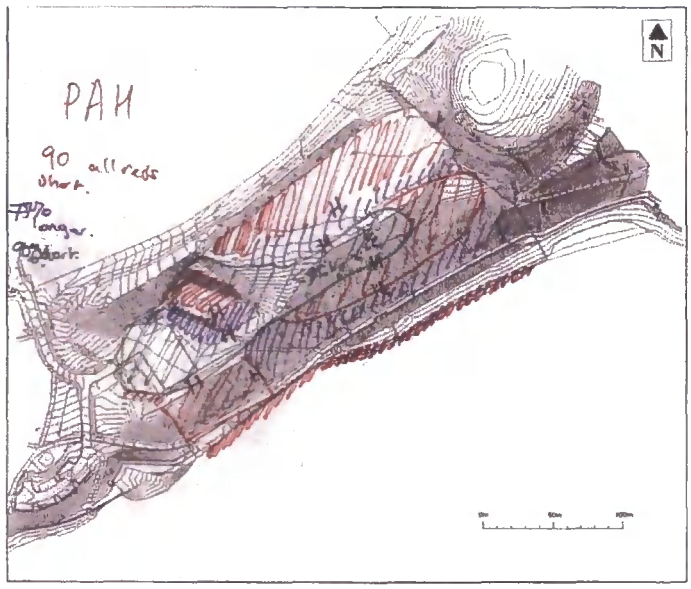


Figure B.7: Expert belief elicitation maps, PAH

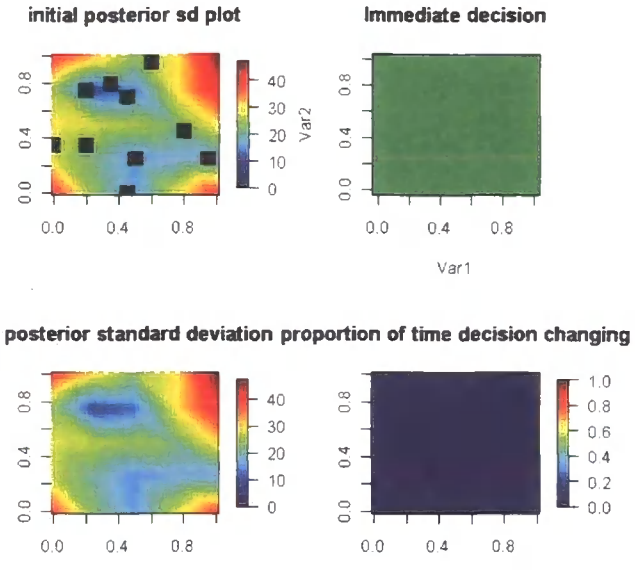


Figure B.8: Stepwise add results, contaminant B where we see no sampling required and therefore no changes to the immediate decision or posterior standard deviation

B.2 Single contaminant stepwise search examples

In addition to the single contaminant example of Section 6.4.1, we look at the remaining two hypothetical contaminants.

Contaminant B

The analysis of contaminant B uses prior beliefs used in Sections 3.6 and 3.7.3. We give a CV of 150 and a UB of 250. The initial herringbone grid was the 33 point grid introduced in Section 6.4.1. We used the initial search to discard $\frac{2}{3}$ of these points, and left 11 for the careful search. In this case it was deemed that no further sampling was necessary, and the decision is to take no action throughout the site, as shown in Figure B.8. This is because of the high upper bound of 250. There are only three values observed which exceed the critical value of 150, and none above the UB. The expected loss associated with an immediate decision was 1.480.

	Imm dec	Cand design
Add	2.802	2.707
Delete	2.802	2.702
Combined	2.802	2.718

Table B.4: Results for the single contaminant stepwise search routines, add, delete and combined, contaminant C

Contaminant C

For contaminant C we use the prior beliefs of Sections 3.6 and 3.7.3. We use a CV of 3 and a UB of 16.

Figures B.9-B.11 show the results of implementing the add, delete and combined search algorithms respectively, for the same prior and loss structures. The results we obtain are shown in Table B.4. We see different results for each case due to the fact that we are simulating, but also because we look at joint relationships in the delete case.

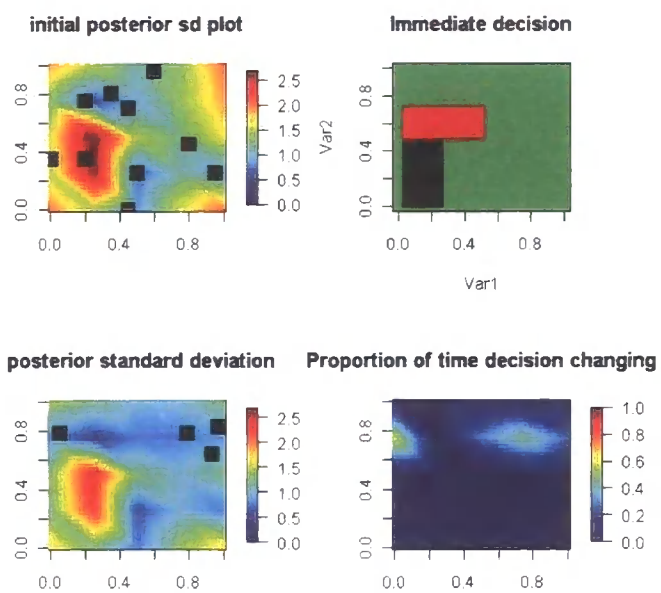


Figure B.9: Stepwise add results, contaminant C

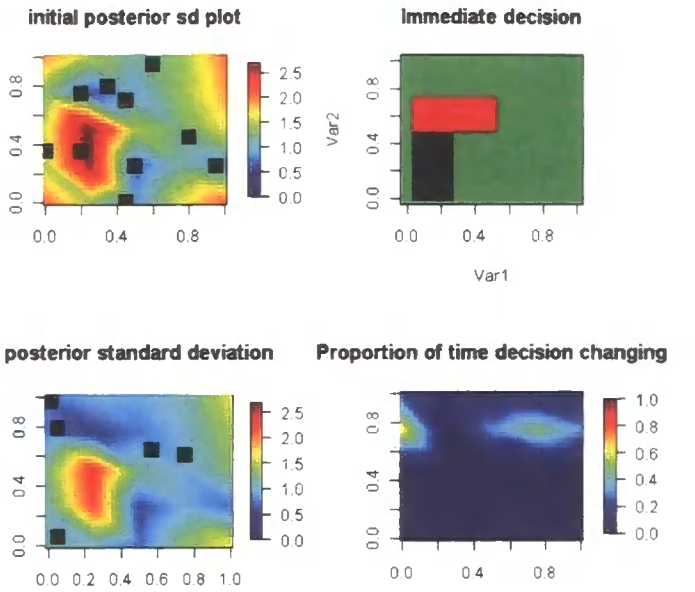


Figure B.10: Stepwise delete results, contaminant C

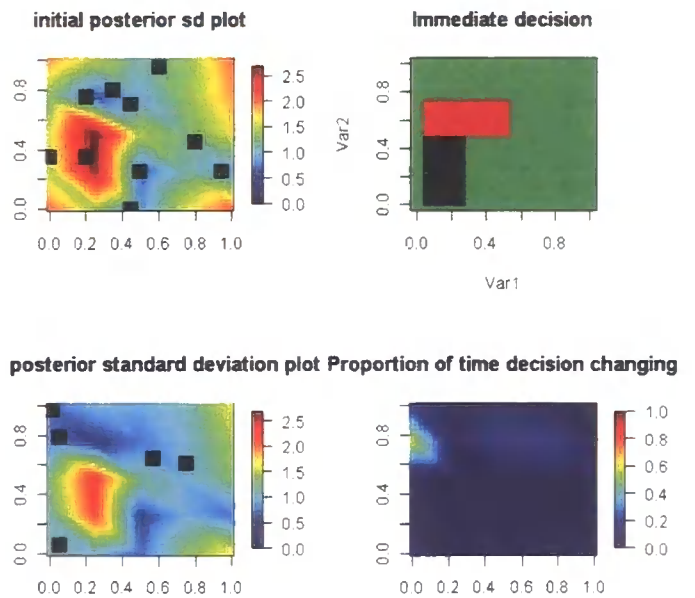


Figure B.11: Stepwise combined results, contaminant C