Sampling Designs for the Spatiotemporal Modelling of Groundwater Quality Monitoring Data

Peter Radvanyi

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Mathematics and Statistics
College of Science and Engineering
University of Glasgow

May 2025

Abstract

Applying appropriate sampling designs in the long-term monitoring of groundwater quality networks is crucial in ensuring that accurate inferences are made about the spatio-temporal distribution of the concentrations of constituents of potential concern (CoPC). Furthermore, the sampling of groundwater monitoring wells and the subsequent analysis of samples induces costs, safety hazards and unintended environmental consequences. Therefore, an optimal sampling design should aim to minimise sample sizes, whilst maximising the value of the information obtained. The problem of finding optimal locations for wells to extend or establish a network has received a lot of attention in the literature. In contrast, fewer approaches have been proposed for optimal sample selection within existing networks, especially in a spatio-temporal context, and the application of these approaches in practice is limited. Current sampling practices often rely on expert judgment and prescriptions by regulatory bodies, which results in datasets that are not well-suited for statistical analysis. Despite its statistical advantages such as generalisability and reducing bias, probability sampling is seldom applied in long-term groundwater quality monitoring. The primary aims of this thesis were to assess common characteristics of long-term groundwater quality data and the use of spatio-temporal models, explore the optimisation of sampling designs through reducing network size, and propose approaches based on probability sampling to support the spatio-temporal modelling of CoPC concentrations.

To compare different approaches to the spatio-temporal modelling of CoPC concentrations via generalised additive models (GAMs) and to assess common characteristics of long-term groundwater quality monitoring data, a comparative study is presented. The study uses synthetic and case study data to evaluate differences in estimating spatio-temporal CoPC concentration surfaces via GAMs with separate and joint smooth terms for space and time. The results highlight the importance of model specification and sampling patterns for obtaining reliable estimates of CoPC concentrations.

In practice, the identification of wells that provide redundant information with respect to the estimation of CoPC concentrations can often be omitted from sampling designs to reduce strain on resources, whilst ensuring that conclusions about spatial and temporal trends are not affected. The demand for a tool to facilitate well redundancy analysis was identified through feedback

shared by the users of the groundwater quality modelling software GWSDAT[1]. In this thesis, a computationally efficient, data-driven approach is proposed for ranking monitoring wells based on their influence on spatio-temporal CoPC concentration models. The approach is based on influential observation diagnostics and is shown to provide rankings similar to a computationally more demanding, cross-validation (CV) based method, through a case study and a simulation study using synthetic groundwater quality data. The approach has also been implemented in GWSDAT.

Thus, omitting redundant wells can be an effective approach for the optimisation of groundwater monitoring networks, but it does not make suggestions on specific sample selection. The generation of spatio-temporal sampling designs, optimised to support the estimation of CoPC concentrations, can help further improve the sustainability of long-term monitoring. In this thesis, it is shown through a literature review that probability sampling designs that aim to draw a spatially and temporally balanced, i.e. evenly spread samples, result in a more precise estimation of CoPC concentration surfaces than simple random designs. Furthermore, it is proposed that by tuning the inclusion probabilities of sample units in balanced designs based on historic data to track the spatial evolution of CoPC concentrations through time, a more precise characterisation of the CoPC plume can be achieved. An original, data-driven methodology is developed for tuning sample inclusion probabilities to be proportional to future predicted distances between the wells and the boundaries of the CoPC plume. Higher probability for selection is given to wells that are predicted to be closer to the plume at a given time. The proposed methodology is shown to provide advantages in characterising plumes given a sufficiently large sample size, using a case study and a simulation study of synthetic CoPC concentration data.

An approach is also proposed for the application of spatio-temporally balanced sampling designs in evaluating the sufficiency of historic sampling intensity. By comparing the CoPC concentration estimates of models using all available historic data and increasingly smaller subsamples selected via balanced designs, it can be assessed whether the monitoring network has been over- or undersampled. This information can then be used to adjust future sampling intensity accordingly.

Finally, consideration is given to the trade-off between high spatial and temporal sampling intensity, given a fixed sample size and monitoring period. In practice, it can be logistically advantageous to perform sampling less frequently, but obtain more samples during each campaign. This results in a spatially high, but temporally low resolution data set. Through a simulation study of synthetic CoPC plume data, it is shown that high spatial resolution data is advantageous when estimating the overall concentration surface, but high temporal resolution data can provide benefits in estimating plume characteristics.

[1] https://gwsdat.net/

Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to thank my supervisors, Prof. Claire Miller, Dr. Craig Alexander and Dr. Marnie Low for their unwavering support and guidance throughout the duration of my doctoral journey. Your encouragement, patience and insightful feedback have been invaluable for the completion of this thesis and my development as a researcher. I am truly grateful to have had the opportunity to learn from you and work with you.

I would also like to thank Dr. Wayne Jones and Dr. Luc Rock for their support and guidance, and for giving me the opportunity to undertake this work. You have always been ready to share your expertise with me and provide help and advice, which I truly appreciate. I would also like to thank you for providing some of the data used in this thesis. Also thank you to Shell Research Ltd. for funding my work.

Thank you to the staff and students of the Statistics Department for your support and solidarity.

Thank you to my family, for always believing in me, even when I did not believe in myself. To my mum, thank you for always being there to talk through difficult times and to my dad, thank you for your encouragement and taking an interest in my work. Nori and Bence, thank you for the fun times we shared when I was at home, they really helped keep my spirits high. To all of you, thank you for the love and support you have given me. I truly would not be here today without them.

Finally, thank you to all my friends who have supported me throughout this journey. Special thanks to Dr. Daniel Kovacs, who was on his own doctoral journey at the same time. Our co-working sessions have helped me push to the end. I will always be grateful for your support.

Declaration

I, PETER RADVANYI, declare that this thesis titled, 'Sampling Designs for the Spatio-Temporal Modelling of Groundwater Quality Monitoring Data' and the work presented in it are my own. I declare that this thesis has been produced in accordance with the University's Code of Good Practice in Research. I confirm that this work was done wholly or mainly while in candidature for a research degree at this University. Where I have consulted the published work of others, this is always clearly attributed. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work. I have acknowledged all main sources of help. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself. The thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice. I acknowledge that if any issues are raised regarding good research practice based on review of the thesis, the examination may be postponed pending the outcome of any investigation of the issues.

Part of the work in Chapter 3 has been presented as a poster in 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, July 2023, with the title 'Computationally Efficient Ranking of Groundwater Monitoring Locations'.

The views and opinions expressed in this body of work are those of the authors and do not necessarily reflect the views, policy or positions of any associated companies including, but not limited to, Shell plc.

Chapter 1

Introduction

This thesis will investigate sampling designs for long-term groundwater quality monitoring networks from a statistical perspective. The presented research will make use of a spatiotemporal statistical modelling framework and probability sampling techniques to propose innovative, data-driven approaches for optimal, economic sample selection to support the characterisation of groundwater quality in terms of solute concentrations. Although the proposed methods were developed with a focus on groundwater quality, their data-driven nature also allows for their broader application in environmental monitoring. This chapter will provide an introduction to environmental sampling, groundwater quality, the optimisation of monitoring designs, groundwater sampling and subsequent data analysis objectives. The spatiotemporal modelling framework and groundwater quality data sets used throughout the thesis will also be introduced. Finally, the outline and main aims of the thesis will be presented.

1.1 Environmental Sampling

Sampling theory is frequently applied in the field of environmental science to provide rigorous protocols for obtaining scientifically reliable information on environmental variables. Stehman and Overton [1994] provide an overview of the application of sampling theory and methods in environmental monitoring surveys. As the authors highlight, environmental objects such as lakes, streams, trees, as well as atmospheric and water pollutant plumes represent real populations with well defined spatial features and characteristics, which are often of interest for statistical inference. In contrast to this, populations in other fields such as epidemiology or economics are often spatially not well defined, and samples are used as starting points for estimating their spatial features (Stehman and Overton [1994]). Depending on the environmental monitoring project, the objectives of sampling and analysis include determining the amount of pollutants

entering an environment, measuring ambient environmental concentrations of analytes to identify short- and long-term trends, detecting accidental releases and studying the fate and transport of contaminants to evaluate the efficiency of remediation systems (C. Zhang [2024]). The analysis of data collected in environmental surveys can yield a description of the population in terms of spatial patterns, trends over time and associations between observed variables. Populations of environmental variables can be considered discrete, consisting of finite individual spatial points such as individuals of a plant or animal species, or continuous, covering a given area such as temperature or air and water quality. Sampling these analytes involves the selection of sampling locations either from the given discrete spatial points or from the continuous area using sampling designs (Stehman and Overton [1994]). Two main approaches can be distinguished in designing sampling strategies to monitor environmental variables. These are probability sampling designs and non-probability sampling designs.

Non-probability sampling designs can rely on expert judgement, process-based analyses or in statistical approaches, mathematical optimisation to determine sample locations and times (Wolf et al. [2016]). To obtain reliable samples using a non-probabilistic approach, extensive knowledge of the underlying processes driving population characteristics is crucial. Additionally, traditional inference principles designed for probability sampling should not be used directly on non-probability samples, as they can result in misguided conclusions (Wolf et al. [2016]). However, non-probability sampling can be justified if strong spatial heterogeneity is known to be present in the population or if other practical considerations have to be accounted for (Speak et al. [2018]). Such considerations can be limited resources for sampling, difficulty in accessing sampling locations, long travel times or having very few potential sampling locations (Stehman and Overton [1994]). Non-probability sampling can lead to high bias in the population estimates if the selection mechanism is not considered (Tutz [2023]), given heterogenous and dependent populations (the selection mechanism should not affect independent, homogeneous populations). In these cases rigorous statistical inference and uncertainty estimation can become difficult. Thus, in many application cases in environmental monitoring, probability sampling methods can aid formal statistical inference on the populations of interest (Kermorvant et al. [2019b]).

Probability sampling relies on stochastic processes to obtain samples from the set of finite or infinite potential locations, where each location is assigned a certain probability of being selected (Wolf et al. [2016]). The advantages of probability sampling designs are that they ensure a representative sample, reduce bias, ensure the generalisability of the results and allow for formal statistical inference (Stehman and Overton [1994]). The probability of a sampling location being included in the design can depend on different factors related to the studied population, as well as the applied selection algorithm. Probabilistic selection algorithms include simple random sampling, systematic sampling and cluster sampling. The first-order inclusion probability

defines the probability of a certain sample being selected, the second-order probability defines the joint inclusion probability of two samples and so forth (Stehman and Overton [1994]). Simple random sampling designs are equal probability designs, because they assign equal first-order inclusion probabilities to each potential sampling location. Unequal probability samples are also frequently applied, with stratified sampling being a common example of such designs, because they allow the disproportionate sampling of different strata (Singh and Mangat [1996]). Unequal probability sampling can be beneficial in cases where a positive correlation exists between inclusion probabilities and response values (Kermorvant et al. [2019b]).

Spatially distributed variables are commonly targets of environmental surveys. Such data often exhibit spatial dependencies that traditional probability sampling designs do not account for (Benedetti et al. [2017]). Spatial dependence is positive, if the observed value of an environmental variable at a given location is more similar to sampling locations in close proximity than to ones farther away (Andersson and Gråsjö [2009]). Therefore, to maximise the amount of useful information obtained using the smallest possible sample size, it is crucial to account for this dependence and avoid the selection of redundant, neighbouring samples. Traditional probability sampling designs such as simple random sampling do not consider this distance-effect and are prone to selecting samples from clusters of neighbouring locations. Sample selection algorithms that attempt to account for spatial dependence between potential sampling locations by aiming to ensure an even spread of samples over the target population are referred to as spatially balanced sampling designs (Benedetti et al. [2017]). These designs allow for drawing both equal and unequal probability samples. Determining inclusion probabilities provides an opportunity to fine-tune sample selection based on additional information about the target population. A more detailed discussion on spatially balanced sampling designs will be provided in Chapter 4 of this thesis.

While probability sampling methods, including spatially balanced sampling are commonly used in other fields of environmental science, their application in long-term groundwater quality monitoring problems is limited (Meray et al. [2022]) due to various practical reasons that will be discussed later in this thesis (see Section 1.2.4 and Chapters 4 and 5). Promoting a more widespread application of probability sampling designs for groundwater quality monitoring would contribute to more effective statistical inferences and thus better protection of groundwater resources. Therefore, the work presented in this thesis will be in the context of concentrations of constituents of potential concern (CoPC) in groundwater as the environmental analytes of interest. It should be noted however, that the proposed novel methodologies are generalisable to other similar environmental analytes as well. The following sections of this chapter will provide an introduction to groundwater, contamination, remediation, long-term monitoring and data analysis through spatiotemporal modelling. At the end of the chapter, the original contributions of this thesis will be highlighted, along with an overview and summary of the primary aims.

1.2 Groundwater

Based on the definition from Hornberger and Perrone [2019], groundwater refers to water found beneath the earth's surface in the saturated zones below the water table, where it fills the spaces and pores between soil particles or fractured rock. The geological formations saturated with water are called aquifers. Aquifers can be made up of different materials such as clay, sand, gravel and fractured stone such as sandstone or limestone. The properties of the aquifer medium largely determine the speed at which groundwater can flow.

Groundwater is an essential resource used worldwide for drinking, irrigation and industrial purposes. Many countries rely on groundwater as the primary source of fresh water. For example, in the European Union 75% of the population depend on groundwater for their water supply and 65% of drinking water is supplied by groundwater (*Groundwater - European Commission* [2024]). As Hornberger and Perrone [2019] points out, most of the fresh water on Earth, around 69.6% is locked in icecaps and glaciers. Of all the remaining liquid fresh water, 30.1% is found beneath the earth's surface. This makes groundwater the largest reservoir of liquid fresh water. Groundwater also plays a crucial role in the balance of the water cycle, by maintaining wetlands and supplying rivers during dry periods. Thus, groundwater is vital to sustaining biodiversity, which highlights the global importance of its protection from pollution and depletion.

1.2.1 Groundwater Contamination

Groundwater quality can be affected by the release of undesirable substances that can cause adverse effects in human populations and the environment (P. Li et al. [2021]). Various naturally occurring and anthropogenic substances of concern can come into contact and be transported by groundwater to natural or man-made outlets (Bear and Cheng [2010]). Due to the relatively slow movement and hidden nature of water in the subsurface, CoPC can remain undetected and be released over long periods of time (Singhal and Gupta [2010]). Naturally occurring solutes in the ground that can cause adverse effects in humans when ingested include minerals and metals such as manganese, arsenic and lead, organic pollutants such as bacteria and products of decaying organic matter, and seawater and brackish water which can also seep into fresh water aquifers, especially in coastal regions (P. Li et al. [2021]).

Human activities can often result in the release of harmful substances into the environment, which can eventually make their way into groundwater as well. There are many sources of anthropogenic groundwater contamination. The main categories are described by Singhal and Gupta [2010] as industrial, agricultural, mining related, waste disposal related and miscellaneous chemical pollutants. The authors also distinguish CoPC based on the characteristics of the

source, which can be point-like, such as hydrocarbon storage tanks, leaky pipes and landfills, or diffuse sources such as agricultural fields or natural mineral and metal deposits. The most common agricultural CoPC are pesticides, herbicides and nitrates from fertilizers (P. Li et al. [2021]) which are of concern because they can affect large areas. Groundwater contamination can also often occur around damaged or improperly installed storage tanks and sewage systems as well as unprotected landfills and municipal and hazardous waste disposal sites (Singhal and Gupta [2010]). These can release organic CoPC and various household and industrial chemicals. Water soluble CoPC can travel in groundwater via diffusion or with the bulk motion of the water i.e. advection, thus forming CoPC plumes that can spread within the aquifer. According to Singhal and Gupta [2010], the most common organic groundwater CoPC are petroleum hydrocarbons often released from compromised underground or above-ground storage tanks for refineries, central heating or petrol stations. Non-oxygenated hydrocarbons have relatively low solubilities but potentially have negative health effects and can be very persistent (Shih et al. [2004]). Benzene, toluene, ethylbenzene and xylenes (BTEX) are substances generally present in all gasoline products and are the most water soluble and toxic compounds associated with non-oxygenated gasoline. There is limited data on the effects of these substances in humans but their permissible limit in drinking water is less than one part per billion (Singhal and Gupta [2010]). Halogenated hydrocarbons (e.g. carbon tetrachloride) are of greater concern since they are stable in the environment and can be accumulated and enriched in living organisms (P. Li et al. [2021]), causing brain disorders and liver damage (Singhal and Gupta [2010]). In urban environments, common petroleum hydrocarbon CoPC are the BTEX compounds and methyltert-butyl-ether (MTBE), which is a fuel additive classed by the U.S Environmental Protection Agency (USEPA) as a possible human carcinogen (Squillace et al. [1997]). According to Singhal and Gupta [2010], most hydrocarbons have limited dissolution between aqueous and organic phases, therefore they are referred to as non-aqueous phase liquids or NAPLs. NAPLs with densities lower than water are called light NAPLs (LNAPLs). Gasoline, kerozene and benzene can form LNAPLs. NAPLs with densities higher than water are called dense NAPLs (DNAPLs), of which common examples include coaltar, chloroform and chlorinated hydrocarbons. There are numerous other potential groundwater polluting substances and sources, which are not in the scope of this thesis to cover. A detailed but not exhaustive list can be found at *Contamination of Groundwater | U.S. Geological Survey* [2024] or in the paper by P. Li et al. [2021]. This thesis will mainly focus on releases and spills of petroleum hydrocarbon compounds such as BTEX.

1.2.2 Long-Term Monitoring and Remediation

Due to increasing human activity around the world, groundwater resources are becoming more and more vulnerable to contamination. Due to groundwater being out of sight and difficult to

access, contamination events are not easy to identify and its consequences are hard to assess (Singhal and Gupta [2010]). Therefore, it is important to perform constant and regular monitoring and implement early warning and control systems at vulnerable groundwater sites (Lyons et al. [2023]). Most countries have legislation in place to protect groundwater resources from contamination. In the European Union this legislation is the Water Framework Directive (*Water Framework Directive - European Commission* [2024]). Depending on the use case of the groundwater resource, the concentrations of potentially harmful substances need to be kept under safe thresholds, determined based on their effects on human health and the environment.

The removal of CoPC from groundwater, or the removal of the contaminated soil and water itself is often required by law (S. Zhang et al. [2017]). Alternatively, the contamination can be monitored over long periods of time, possibly decades, until its concentration is naturally diluted below safety limits (Rügner et al. [2006]). The act of neutralising or removing CoPC from groundwater is referred to as remediation. Active remediation can happen in a variety of ways. S. Zhang et al. [2017] provide an overview of groundwater remediation technologies proposed in the literature. Contaminated groundwater can be pumped out of the soil and subsequently treated and disposed of. For certain CoPC a less intrusive, in-situ treatment approach can be used in which reagents can be pumped into the groundwater to initiate chemical and biological reactions such as oxidation, reduction or biodegradation that can neutralise the harmful compounds. Active remediation technologies are typically more intrusive, costly and have higher environmental footprints due to the amount of required materials, machinery, transportation and personnel (Meray et al. [2022]). The concept of sustainable remediation (U.S. Sustainable Remediation Forum [2009]) has emerged to address some of the problems with traditional remediation approaches by focusing on less intense, passive remediation practices. Sustainable remediation considers the net impact of the remediation process itself on the surrounding environment and promotes the use of monitored natural attenuation or MNA (Rügner et al. [2006]) where possible. MNA relies on containment and the long-term monitoring of natural processes to achieve remediation objectives rather than direct intervention. In the United Kingdom, the Sustainable Remediation Forum (SuRF) is an initiative set up by the independent, non-profit organisation Contaminated Land: Applications in Real Environments (CL:AIRE), to progress the understanding of sustainable remediation and land management [1].

As mentioned above, groundwater is by nature difficult to access. Places where groundwater reaches the surface without mixing with surface water are rare, especially in urban settings. Therefore, monitoring the quality of groundwater requires drilling boreholes or wells, which provide constant access to the aquifers for the use of sampling equipment (Bear and Cheng [2010]). A collection of monitoring wells established within the area of interest is referred to as the groundwater monitoring network. Such monitoring networks are used to obtain infor-

[1] https://claire.co.uk/projects-and-initiatives/surf-uk

mation about the state of the groundwater system. There are certain aspects of groundwater quality such as temperature, pH and electrical conductivity that can be monitored remotely via sensors placed in wells. There are only a few CoPC whose concentrations can be inferred from these in-situ measurements (Schmidt et al. [2018]). Most commonly, the direct observation of CoPC concentrations requires sampling and subsequent laboratory analysis. Thus, long-term monitoring involves frequent sampling campaigns to collect information on the concentrations of undesirable substances within aquifers over time (Lyons et al. [2023]). Ultimately, the main aim of long-term monitoring in the context of MNA, is to confirm the stability of the spatial distribution of CoPC, detect changes and anomalies in their mobility and observe the continued reduction of their concentrations in the groundwater (Meray et al. [2022]).

1.2.3 Sampling Design Optimisation

Due to the relatively high cost and environmental footprint of drilling, the number of monitoring wells should be kept minimal but adequate. Therefore, their positioning requires thorough consideration that should account for aquifer characteristics, hydrological conditions and site accessibility. There is extensive literature on the optimal design of groundwater quality monitoring networks, with the problem of finding optimal locations for installing new monitoring wells receiving most of the attention (Farlin et al. [2019]). Methods for solving such optimisation problems range from purely data-driven to purely process based approaches. Farlin et al. [2019] provide a summary of the mathematical and statistical tools that have been applied to this problem, which include geostatistical methods (Loaiciga [1989]), principal components and cluster analysis (Daughney et al. [2012]), Bayesian optimisation (Nowak et al. [2012]) and numerical groundwater flow modelling (Wöhling et al. [2016]). In practice, however, the application of these methods is still rather uncommon. Well locations are often determined based on input from regulators and consultants (Meray et al. [2022]). This is partly due to the low availability of tools that do not require special training in statistics on the side of the user.

As Farlin et al. [2019] also highlighted, developing optimal sampling strategies for already established monitoring networks has received much less attention. Groundwater sampling campaigns and subsequent laboratory analyses also induce a strain on resources, since they require transportation, equipment and personnel (Wu et al. [2005]). Thus, selecting when and which wells to sample is essential for increasing the sustainability of long-term monitoring operations. An optimal sampling design should aim to minimise the amount of sampling events, whilst maximising the amount of useful information that can be obtained from those samples with respect to the monitoring objectives. Most statistical methods proposed to achieve this goal rely on estimating the spatiotemporal distribution of CoPC concentrations (Meray et al. [2022]). However, many of the available optimisation approaches such as the one proposed by Farlin et al. [2019] fo-

cus only on the spatial aspect of sampling design, and assume an adequate sampling frequency. McLean [2018] proposed two objective functions that could be used to optimise sample selection in existing monitoring networks in a spatiotemporal framework, but these methods result in non-probabilistic designs and have not been evaluated in terms of how precisely the resulting samples could predict CoPC concentrations. Ultimately, a methodology that could provide a complete, optimal sampling strategy for monitoring periods of any given length would be highly valuable in practice. The work presented in this thesis, will look to develop such a methodology using spatially balanced sampling designs.

1.2.4 Current Groundwater Sampling Design Practices

The implementation of sampling design optimisation methods in practice is very limited. This is because sampling strategies at groundwater monitoring sites often rely on narrowly focused periodic sampling based on subjective expert judgements and directives, resulting in sporadic sampling designs (Lyons et al. [2023]). Sampling is commonly carried out for the purposes of compliance with regulations. The sampling frequency is often prescribed by these regulations, while sampling locations are commonly selected by expert judgments and regulator opinions (Meray et al. [2022]). Furthermore, the collected data are often archived without meaningful statistical assessment. Thus, groundwater sampling strategies are not typically designed with subsequent statistical analysis in mind. This can hinder long-term monitoring operations in achieving their goals, which were described in Section 1.2.2. As mentioned in Section 1.2.3, one of the challenges for the more widespread adoption of sampling optimisation techniques is the availability of easy to use tools, that provide extensive data analysis and optimisation features (Meray et al. [2022]). Therefore, developing methodologies that can be integrated into such tools could contribute to the adoption of better practices in the sampling of long-term groundwater monitoring networks.

1.2.5 Groundwater Monitoring Data Analysis

The collected data on CoPC concentrations along with other hydrogeological parameters can ultimately be used to achieve the groundwater monitoring objectives through the use of statistical modelling tools. Common characteristics of interest are the spatial distribution of CoPC concentrations, plume mass, area, average concentration and time series of CoPC concentrations within monitoring wells (Jones et al. [2014]). The spatial distributions can be estimated by simply interpolating concentrations between monitoring well locations or fitting a concentration surface using geostatistical approaches such as Kriging, inverse distance weighting or smooth-

ing splines (McLean et al. [2019]). The spatial distribution of concentrations can help delineate the CoPC plume and assess the extent of the effected area. Monitoring the spatial distribution over time can identify changes in the stability of the plume and confirm the continued reduction of concentrations (Meray et al. [2022]).

1.2.6 Spatial and Spatiotemporal Modelling

As McLean et al. [2019] highlights, the spatial characteristics of groundwater CoPC plumes are most commonly assessed separately from temporal features. In most cases, spatial analysis is done by fitting statistical models to separate monitoring events, or to data sets consisting of multiple events within a given time-period. Temporal analyses are often only concerned with changes and trends in solute concentrations at individual sampling locations or monitoring wells. McLean et al. [2019] also point out that in contrast with other environmental disciplines such as forecasting precipitation or stream flow, combined spatiotemporal modelling of groundwater quality data is rarely done in practice. McLean et al. [2019] showed that a joint, spatiotemporal modelling framework can provide better estimation accuracy of groundwater quality characteristics than repeated spatial analyses while requiring smaller sample sizes. Thus, using this approach can contribute to increasing the sustainability of long-term monitoring operations. The application of spatiotemporal modelling tools for groundwater quality monitoring is a fairly recent development. One of the most prominent software tools integrating this approach is Groundwater Spatiotemporal Data Analysis Tool or GWSDAT (Jones et al. [2014]).

1.3 GWSDAT

GWSDAT is an open-source software tool designed to model, interpret, and visualize trends in groundwater quality monitoring data. A complete description of the software's capabilities and structure is available in the publication by Jones et al. [2014] and on the GWSDAT website[2], while its source code can be found on GitHub[3]. Created during a collaboration between the University of Glasgow and Shell Global Solutions International BV, it employs a unique, efficient spatiotemporal modelling framework (Evers et al. [2015]) based on penalized splines (P-splines) (Eilers and Marx [1996]). This efficient approach contributed to GWSDAT's success and popularity worldwide (Jones et al. [2022]), especially in long-term groundwater contamination monitoring and remediation.

[2]https://gwsdat.net/
[3]https://github.com/WayneGitShell/GWSDAT

GWSDAT analyses time-series data of solute concentrations and groundwater elevations observed in monitoring wells. It can generate a series of spatial plots showing estimated concentrations and groundwater elevations on the monitored site. It can also outline the solute plume, compute plume diagnostics such as mass and area, and report trends in the time-series of each monitoring well. It also allows for modelling the thickness of non-aqueous phase liquid (NAPL) layers. The coordinates of the sampling wells is the only required input as spatial information, but GWSDAT also supports the import of site maps in the form of GIS shape-files. GWSDAT is built on the open-source statistical programming language R (R Core Team [2020]).

Part of the work undertaken in this thesis builds on and addresses issues and requests identified directly by the user base of GWSDAT, thus contributing to the development of methods in real-life applications. These include analysing the influence of monitoring wells on model predictions and generating optimal sampling designs using statistical methods.

1.3.1 Spatiotemporal Modelling Framework

The solute concentration surface in GWSDAT is estimated using a unique P-splines based spatiotemporal modelling framework from concentration measurements collected from monitoring wells along with corresponding coordinates and sampling dates. Jones et al. [2014] provide an overview of this framework as it is implemented in the software and Evers et al. [2015] provide a more in depth discussion, particularly on the problem of selecting an optimal smoothing parameter. As this modelling approach will be the primary tool used throughout this thesis to evaluate sampling designs, this section will introduce its structure and how it is applied to groundwater monitoring data.

Splines are an approach for non-parametric regression that describe the relationship between covariates and responses in a flexible manner (McLean et al. [2019]). For example, let y_i be a set of responses and x_i be corresponding covariates with $i = 1, 2, ..., n$ where n is the number of observations. A model for the relationship between y_i and x_i can be described as:

$$y_i = f(x_i) + \varepsilon_i, \tag{1.1}$$

where $f(x)$ represents a non-parametric regression function of the covariates and $\varepsilon_i \sim N(0, \sigma^2)$ represents random variation. As Fahrmeir et al. [2013] discuss, in the case of splines, the non-parametric function $f(x)$ can be represented by polynomials of a given degree defined on specific intervals of the covariate domain with smoothness constraints to ensure smooth fusion at the knots (κ) between the polynomial pieces. This class of basis functions are called polynomial

splines. A set of polynomial splines for a given degree and knots configuration can be represented by different approaches, with B-splines being a common choice. B-spline basis functions are constructed from $(l+1)$ piecewise polynomials of degree l, which are joined in an $(l-1)$-times continuously differentiable way (Fahrmeir et al. [2013]). The function $f(x)$ can thus be represented through a linear combination of B-spline basis functions $B_j(x)$ and corresponding basis coefficients α_j such that

$$f(x) = \sum_{j=1}^{m} \alpha_j B_j(x), \tag{1.2}$$

where the basis coefficients α_j can be estimated using a least-squares approach (McLean et al. [2019]). The spatiotemporal B-splines regression function is an extension of the one-dimensional function described in equation 1.2 to three covariate dimensions. For spatiotemporal data, which are indexed over space (coordinates x_1 and x_2) and time (t), the regression function $f(x_1,x_2,t)$ can be expressed in the form:

$$f(x_1,x_2,t) = \sum_j \sum_k \sum_l \alpha_{jkl} B_j(x_1) B_k(x_2) B_l(t), \tag{1.3}$$

where x_1 and x_2 represent spatial coordinates for easting and northing, t represents time and B_J, B_k and B_l are the corresponding B-spline basis functions with $j = 1,2,...,n$, $k = 1,2,...,m$ and $l = 1,2,...,p$. This corresponds to the tensor product of the marginal B-spline bases, which can be computed efficiently through row-wise Kronecker products (McLean et al. [2019]). The number of basis functions used to construct the regression function is crucial as it controls the level of smoothness, but its selection is subjective and involves a bias-variance trade off (McLean et al. [2019]). To overcome the difficulty in the selection of an appropriate number of basis functions, P-splines were proposed by Eilers and Marx [1996]. As described by Fahrmeir et al. [2013], in the P-splines approach, a high number of basis functions are used (20-40 per dimension) to ensure a high level of flexibility for representing even highly complex functions, and an additional penalty term (λ) is introduced to penalize overfitting. These penalties are based on the derivatives of the regression function since these characterise its properties. The first derivative of a function constructed of B-spline basis functions can be expressed as a function of the first-order differences of the corresponding basis coefficients as

$$\frac{\partial}{\partial x} \sum_j \alpha_j B_j^l(x) = l \cdot \sum_j \frac{\alpha_j - \alpha_{j-1}}{\kappa_j - \kappa_{j-1}} B_{j-1}^{l-1}(x). \tag{1.4}$$

The penalties can then be based on these differences to avoid large values of the first derivative

and obtain a smooth function. Using equidistantly placed knots provides a simple way to express these difference penalties (Fahrmeir et al. [2013]). Similarly, higher order (d) differences can also be used to induce a smooth function in terms of d-th order derivatives. A common choice for this penalty is the use of the second derivative, such as

$$\lambda \int (f''(x))^2 dx, \tag{1.5}$$

since the second derivative measures the curvature of the function. In the context of spatiotemporal models, Evers et al. [2015] highlights that using a second-order penalty can result in abnormally high peaks or low valleys in the model especially in large gaps between data points, and that using a first-order penalty can to some extent mitigate this effect. The smoothing parameter λ can be determined through the application of various optimality criteria such as the penalised least squares criterion (PLS), Akaike's information criterion (AIC), Bayesian information criterion (BIC), cross-validation (CV) or generalised cross-validation (GCV). In GWSDAT's modelling framework, the smoothing parameter λ is estimated using a Bayesian approach introduced by Evers et al. [2015], which is described in this section.

When modelling the spatiotemporal distribution of concentrations of CoPC in a groundwater aquifer (representing the population), the aim is to describe the relationship between the measured concentration of the CoPC in the water samples (representing the response) and the coordinates of the monitoring wells from which the samples were obtained, as well as the dates on which the samples were taken (representing the covariates). Thus, let y_i be the natural logarithm of solute concentrations associated with covariates easting x_1, northing x_2 and time t. The solute concentrations are commonly log-transformed to normalise their distribution. Chapter 2 provides a more in-depth discussion on this issue. Let $i = 1,...,N$, where N is the total number of observations. The solute concentrations can then be represented by the model:

$$y_i = f(x_1, x_2, t) + \varepsilon_i, \tag{1.6}$$

where $f(x_1, x_2, t)$ is the spatiotemporal regression function made up of B-spline basis functions as described in equation 1.3 and ε_i represents random variability associated with sampling and analytical techniques that determine the concentration of the solute in the sample, with distribution $N(\mu, \sigma^2)$. The spatiotemporal smooth $f(x_1, x_2, t)$ aims to capture the spatiotemporal dependence of y_i in the systematic part of the model. The model in equation 1.6 can be expressed in vector matrix form, which gives:

$$\boldsymbol{y} = \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{1.7}$$

where $\boldsymbol{y}$ is the vector of solute concentration observations, $\boldsymbol{B}(\boldsymbol{x})$ is the matrix of basis functions (columns) evaluated at each data point (rows), $\boldsymbol{x}$ is the matrix of covariates (coordinates and time points), $\boldsymbol{\alpha}$ is the vector of basis coefficients and $\boldsymbol{\varepsilon}$ is the vector of random errors. Using the P-splines implementation with equidistantly placed knots over the spatiotemporal domain, the values of the coefficients $\boldsymbol{\alpha}$ are chosen to minimise the objective function

$$S(\boldsymbol{\alpha}) = \|\boldsymbol{y} - \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha}\|^2 + \lambda \|\boldsymbol{D}_d \boldsymbol{\alpha}\|^2, \tag{1.8}$$

where the matrix $\boldsymbol{D}_d$ computes the successive d-th order differences, in the case of GWSDAT the first-order differences, across the sequence of α-s in each of the covariate dimensions (easting and northing coordinates and time). λ is a non-negative smoothing parameter that controls the degree of smoothness of the function. By minimising the objective function for a given value of λ we get the estimator of the parameters:

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{B}^T\boldsymbol{B} + \lambda \boldsymbol{D}_d^T \boldsymbol{D}_d)^{-1}\boldsymbol{B}^T\boldsymbol{y}. \tag{1.9}$$

The fitted solute concentration values are then given by:

$$\hat{\boldsymbol{y}} = \boldsymbol{B}\hat{\boldsymbol{\alpha}} = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B} + \lambda \boldsymbol{D}_d^T \boldsymbol{D}_d)^{-1}\boldsymbol{B}^T\boldsymbol{y}, \tag{1.10}$$

where the trace of the hat matrix $(\boldsymbol{B}^T\boldsymbol{B} + \lambda \boldsymbol{D}_d^T \boldsymbol{D}_d)^{-1}\boldsymbol{B}^T$ which creates the fitted values from $\boldsymbol{y}$ by analogy with standard linear models corresponds to the effective degrees of freedom (Evers et al. [2015]).

Increasing the λ value induces a higher degree of smoothness, making the model less flexible. In contrast, a lower λ value increases the flexibility of the model. If $\lambda = 0$, than the expression in 1.9 gives the ordinary least squares estimate for $\hat{\boldsymbol{\alpha}}$ (Jones et al. [2014]). Thus, the selection of an appropriate smoothing parameter λ is crucial to avoid capturing too much of the noise in the data (overfitting) or generating a model that is too smooth (underfitting).

In GWSDAT, the smoothing parameter λ is estimated using a Bayesian approach introduced by Evers et al. [2015]. Let M_λ be the model obtained by a particular value of λ. The likelihood function can then be derived from

$$\boldsymbol{Y}|\boldsymbol{\alpha}, \sigma^2, M_\lambda \sim \mathcal{N}(\boldsymbol{B}\boldsymbol{\alpha}, \sigma^2 \boldsymbol{I}_n), \tag{1.11}$$

where $\boldsymbol{Y} \in \mathbb{R}^n$, $\boldsymbol{B} \in \mathbb{R}^{n \times m}$ and $\boldsymbol{\alpha} \in \mathbb{R}^m$. For a fixed value of λ, the normal-inverse gamma is a conjugate prior for the parameters $\boldsymbol{\alpha}, \sigma^2$:

$$\boldsymbol{\alpha}, \sigma^2 | M_\lambda \sim \mathcal{NIG}(\mathbf{0}, (\lambda \boldsymbol{D}^T \boldsymbol{D})^{-1}, a, b), \tag{1.12}$$

where Evers et al. [2015] suggest using $a = b = 0.0001$ to acknowledge the uncertain prior information on the parameter σ^2. These values of a and b were used throughout the thesis when employing this modelling framework. The posterior distribution of λ can be shown to be

$$f_{M_\lambda|\boldsymbol{Y}} \propto \lambda^{rank(\boldsymbol{D}^T\boldsymbol{D})/2} \frac{|\boldsymbol{B}^T\boldsymbol{B} + \lambda \boldsymbol{D}^T\boldsymbol{D}|^{-1/2}}{\{2b + \boldsymbol{y}^T[\boldsymbol{I_n} - \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B} + \lambda \boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{B}^T]\boldsymbol{y}\}^{a+n/2}} f_{M_\lambda}^{prior}, \tag{1.13}$$

which is a special case of comparison of Bayesian linear models, using λ as the model index (Evers et al. [2015]). The value of λ which maximises the posterior density known as the maximum a posteriori (MAP) value is selected as the smoothing parameter in the P-splines model. The computational efficiency of the approach is improved by exploiting the sparsity of the design and penalty matrices. A more detailed discussion is given by Evers et al. [2015].

Model smoothness in P-splines models can still be controlled by adjusting the number of basis functions, referred to as the number of segments (nseg) in GWSDAT, in the three covariate dimensions. However, as discussed in relation to equation 1.2, the selection of the number of basis functions is subjective and involves a trade off between bias and variance (McLean et al. [2019]). Increasing the number of basis functions gives the model more flexibility and results in lower bias but higher variance. Moreover, it can also significantly increase computation times. On the other hand, reducing it reduces variance but increases bias. The P-splines framework proposed by Eilers and Marx [1996] aimed to overcome this dilemma by using a high number of basis functions and introducing a penalty on the basis coefficients. Claeskens et al. [2009] provide an analysis comparing regression splines without penalty using fewer basis functions and smoothing splines with penalty using more basis functions. The authors concluded that using fewer basis functions resulted in better asymptotic rates. GWSDAT allows for limited manual control of the number of basis functions nonetheless, in case a more flexible model is warranted.

Throughout this thesis, the above described framework will be used to model groundwater quality monitoring data and subsequently evaluate sampling designs. Relevant chapters will provide a description of specifications that were used to fit the models.

1.4 Groundwater Monitoring Data Sets

Various long-term groundwater quality monitoring data sets (simulated and case study data sets) will be used throughout this thesis to test the proposed methods. This section aims to introduce these data sets and give an overview of their characteristics. A common feature of groundwater monitoring data is that they consist of at least three main variables. These are the coordinates of the sampling well (easting and northing), the time the samples were taken and the concentration of the investigated solute or solutes measured within the samples. The simulated data sets used in this thesis contain only these main variables for a single hypothetical solute. The case study data sets may or may not contain additional variables measured in-situ at the wells, such as the sampling depth, groundwater level, temperature, electric conductivity or pH. The following sections will provide an overview of each data set used in this thesis.

1.4.1 Simulated data set 1

The first simulated data set was generated by Evers et al. [2015] from a partial differential equation (PDE) representing a highly idealised model of the spread of solutes in groundwater. The PDE is described by the following expression:

$$\frac{\partial y}{\partial t} = D \times \left(\frac{\partial^2 y}{\partial x_1^2} + \frac{\partial^2 y}{\partial x_2^2} \right) + \omega_1(x_1, x_2) \frac{\partial y}{\partial x_1} + \omega_2(x_1, x_2) \frac{\partial y}{\partial x_2} \tag{1.14}$$

where y denotes the CoPC concentration, x_1 and x_2 represent the spatial coordinates and $t \in [0, 1]$ is time. D is a constant that controls the speed with which the CoPC spreads. D is multiplied by a term that describes the diffusion of the CoPC in groundwater. The last two terms represent advection and describe how the spread of the CoPC is affected by the groundwater flow in each direction. ω_1 and ω_2 denote the direction and velocity of groundwater flow in two dimensions.

The solute concentration surface was generated by computing the numerical solution to the PDE over a 100×100 meter grid with 1×1 meter cells and 167 time points. The concentration values representing the observations were obtained by sampling the concentration surface at 29 sampling locations. The arrangement of the sampling locations is shown on Figure 2.13a. All the above design parameters were selected to mimic groundwater quality monitoring sites. The resulting data set consisted of the monitoring well coordinates, the times of sampling events and the simulated concentrations of the hypothetical CoPC. As described by Evers et al. [2015], well-specific and measurement noise were added to the observations to mimic the sampling and analytical errors of groundwater monitoring data. The multiplicative noise was generated

using Gaussian errors with standard deviation chosen to correspond to a signal-to-noise ratio of 10:1. A weak correlation between errors of coefficient 0.05 was introduced for observations originating from the same well to represent well-specific errors. For modelling purposes in this thesis, the above described noisy data were used. The details of this will be discussed in Chapter 2. Figure 2.20a shows the spatial distribution of solute concentrations at the final time point.

1.4.2 Simulated data sets 2, 3 and 4

Data sets 2, 3 and 4 were created by (McLean et al. [2019]). The data sets represented three hypothetical CoPC plumes and were generated using a groundwater flow model based on the modular, finite difference groundwater flow model developed by the U.S. Geological Survey, MODFLOW (Harbaugh et al. [2000]) and a solute transport model based on MT3D (Zheng [1990]). Each of the three plumes represented a different level of complexity with regards to the spatial distribution of CoPC concentrations. Hence, they will be referred to as simple, medium and complex plumes. Each data set contained 450000 data points of spatial point coordinates, times and concentrations for a hypothetical solute. There were a total of 20 time points representing sampling events at 6 month intervals for a total period of 10 years. There were 22500 spatial point coordinates on a 1×1 km^2 grid. Figure 3.1 shows the spatial geometry of the CoPC plumes at the final time point. Table 1.1 shows the detailed specifications of the models, including hydrological and geological parameters McLean et al. [2019].

The partial differential equation used in MODFLOW to represent groundwater flow is described in the following equation:

$$\frac{\partial}{\partial x}\left(K_{xx}\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}\frac{\partial h}{\partial y}\right)+\frac{\partial}{\partial z}\left(K_{zz}\frac{\partial h}{\partial z}\right)+W=S_s\frac{\partial h}{\partial t}, \tag{1.15}$$

where K_{xx}, K_{yy} and K_{zz} are values of hydraulic conductivity along the x, y and z coordinate axes, h is the potentiometric head, W is a volumetric flux per unit volume representing sources and sinks of water with $W < 0$ for outflow and $W > 0$ for influx, S_s is the specific storage of the porous material and t is time. The details of what hydrogeological process or aquifer characteristic each term represents can be found in Harbaugh et al. [2000].

Equation 1.15 combined with initial and boundary conditions, describes transient three dimensional groundwater flow in a heterogeneous and anisotropic medium, provided that the coordinate directions are aligned with the principal axes of hydraulic conductivity. The partial-differential equation is solved using the finite-difference method.

Table 1.1: Summary of parameters and specifications used in the MODFLOW/MT3D models to simulate the three hypothetical contamination plumes (McLean et al. [2019]).

Model aspect	**Model value**
Model discretisation	
Model domain	$1x1$ km^2
Vertical discretisation	1 layer, 10 m thick, confined
Flow parameters (MODLFOW)	
Horizontal conductivity (K)	**Simple plume**: fixed $K = 15md^{-1}$ (sand) **Medium plume**: mean $K = 15md^{-1}$, standard deviation $log_{10}(sd(K)) = 0.4$ **Complex plume**: mean $K = 15md^{-1}$, standard deviation $log_{10}(sd(K)) = 0.9$
Porosity (n)	0.25
Boundary conditions	Constant head cells at east (0 m) and west (-1 m) boundaries resulting in a hydraulic gradient of 0.001 m/m. Active cells elsewhere. Groundwater recharge rate of 1 mm d^{-1}
Time discretisation	Total simulation time: 10 years
Transport parameters (MT3D)	
Longitudinal dispersivity	6 m
Horizontal and vertical transverse dispersivity	0.1 m
Diffusion	Ignored
Advection	Method of characteristics (MOC) scheme (time discretisation based on a Courant number of 0.75)
Boundary conditions	Fixed concentration cells on the west boundary (0 mg l^{-1}) representing clean groundwater inflow, active concentration cells elsewhere. The contamination source area ($100m^2$) is represented by assigning a concentration groundwater recharge of 100 mg l^{-1} to 36 model cells. Outside the contaminated area, the concentration of recharge is 0 mg l^{-1}.
Initial conditions	Assigned concentration of 0 mg l^{-1}

The governing partial-differential equation underlying the solute transport model in MT3D is:

$$R\frac{\partial C}{\partial t} = \frac{\partial}{\partial x_i}\left(D_{ij}\frac{\partial C}{\partial x_j}\right) - \frac{\partial}{\partial x_i}(v_i C) + \frac{q_s}{\theta}C_s - \lambda\left(C + \frac{\rho_b}{\theta}\bar{C}\right), \tag{1.16}$$

where C is the concentration of CoPC dissolved in groundwater, t is time, x is the distance along the respective Cartesian coordinate axis, D_{ij} is the hydrodynamic dispersion coefficient, v is the seepage or linear pore water velocity, q_s is the volumetric flux of water per unit volume of aquifer representing sources (positive) and sinks (negative), C_s is the concentration of the sources and sinks, ρ_b is the bulk density of the porous medium, $\bar{C}$ is the concentration of CoPC adsorbed to the porous medium, λ is the rate constant of the first-order rate reactions, θ is the porosity of the porous medium (dimensionless) and R is called the retardation factor. R is defined as

$$R = 1 + \frac{\rho_b}{\theta}\frac{\partial \bar{C}}{\partial C}, \tag{1.17}$$

where the first term on the right-hand side of the equation represents hydrodynamic dispersion, while the second term represents advection. The details of what hydrogeochemical process or aquifer characteristic each term represents can be found in Zheng [1990]. The effects of molecular diffusion are generally negligible compared to mechanical dispersion. The former only becomes important at very low groundwater velocities. Thus, the effects of molecular diffusion were ignored for the simulation of the hypothetical plumes. The hypothetical medium and complex plumes were simulated using a heterogeneous hydraulic conductivity field, while the simple plume assumed homogeneous conductivity with a fixed $K = 15md^{-1}$ (see Table 1.1), which is considered representative of a sand aquifer.

1.4.3 Case study data set 1

The first groundwater quality monitoring case study data set used in this thesis is available publicly in GWSDAT[4] by the name GWSDAT Basic Example. It contains a total of 520 observations in the form of solute concentration measurements from 11 monitoring wells. The data were collected over a monitoring period between 2002 to 2006, with irregular sampling frequency and a sporadic spatial sampling design. The information recorded in the data set are the well identifiers, the well coordinates (easting and northing), the dates of the sampling events, the groundwater levels, the analysed CoPC and their measured concentrations in μg/l. The analysed CoPC were the petroleum hydrocarbons benzene, toluene and xylene (BTEX). A more detailed discussion on the analysis of the data set will be presented in Chapter 2. Figure 2.1 shows the

[4] https://stats-glasgow.shinyapps.io/GWSDAT/

coordinates of the monitoring wells and the observations collected from these wells over the monitoring period. Figure 2.11 shows the benzene concentration surface at the final time point estimated from this data using GWSDAT.

1.4.4 Case study data set 2

The other case study used in this thesis is the more comprehensive of the two example data sets available for GWSDAT (see footnote 4) by the name GWSDAT Example. The anonymised groundwater quality monitoring data come from the long-term monitoring of a decommissioned petrol station. It consists of groundwater levels and concentration measurements for five different CoPC in groundwater samples obtained from 32 existing monitoring wells. The relative spatial coordinates of these wells are also part of the data set. The observations were collected over a 4-year monitoring period. The five solutes of concern that were investigated were ethylbenzene, toluene, total petroleum hydrocarbons (TPH), nitrate and sulphate. The data also include measurements of NAPL thickness. Figure 2.1 shows the coordinates of the monitoring wells and the observations collected from these wells over the monitoring period. Figure 2.12 shows the ethylbenzene concentration surface at the final time point estimated from the data using GWSDAT. Figure 3.3 shows these estimates obtained using the P-splines modelling approach in R along with estimated groundwater level contour lines, as well as the layout of the monitoring well network.

1.4.5 On handling non-detects

Groundwater monitoring data sets often contain so called non-detects (NDs). These are a class of observations, where the concentration of the CoPC was below the detection limit of the laboratory analysis. In the case study data set these observations are noted by $ND < x$, where x is the detection limit. There are different approaches to treating NDs in practice, such as substituting the value with x or $\frac{1}{2}x$, or using various statistical methods such as regression on order statistics (ROS) or maximum likelihood estimation (MLE) (Gibbons et al. [2009]). Substituting is one of the most commonly used approaches, but it can introduce significant bias, especially in cases with a high proportion of NDs. According to Gibbons et al. [2009], substitution methods can be adequate in practice if the detection frequency is above 80%. MLE and ROS methods are commonly used for handling NDs without introducing substantial bias. In the case study data set, the proportion of NDs is $\sim 33\%$, considering all measured solutes. This could justify the application of MLE or ROS based methods in practice. However, in this thesis a substitution based solution was deemed sufficient, since the model estimates are only used to compare the

efficiency of different sampling designs which are applied to the same data set. Thus, in this thesis, NDs were substituted with $\frac{1}{2}x$, i.e. half of the detection limit value.

1.5 Thesis Aims & Overview

In summary, the primary focus of this thesis is the development of novel statistical approaches and methodologies for optimising the sampling designs of existing, long-term groundwater monitoring well networks. These novel approaches aim to reduce sample sizes by selecting optimal wells and times to sample based on the analysis of historic data. At the same time, they aim to ensure that sufficient amounts of information is collected for the subsequent spatiotemporal modelling of groundwater CoPC concentrations. Consequently, the work presented in this thesis could help reduce the costs and environmental footprint associated with groundwater sampling campaigns and improve sustainability and estimation quality. Whilst this thesis focuses on long-term groundwater monitoring, the methods presented here can be generalised to other areas of environmental science as well. The following paragraphs summarise the novel methodological contributions of this thesis by chapter.

Building on the work by McLean et al. [2019] on spatial vs spatiotemporal modelling in groundwater quality monitoring, Chapter 2 will evaluate different approaches to spatiotemporal modelling using generalised additive models (GAMs). Through an original comparative study using various synthetic and real groundwater contamination data, the chapter will look to find optimal spatiotemporal GAMs, and compare them to the modelling approach employed in GWSDAT (Evers et al. [2015]). It will also explore the characteristics of long-term groundwater monitoring data and elucidate corresponding design challenges related to spatiotemporal sample distribution. Later chapters will continue to explore the optimisation of sampling designs specifically in the context of supporting spatiotemporal modelling.

In Chapter 3, a novel optimisation approach is developed that can be used to analyse the influence of monitoring wells on the predictions from spatiotemporal groundwater contamination models. The approach is based on the statistics of influential observations and aims to establish a ranking of wells within monitoring networks, where a lower rank corresponds to less influence on model predictions. The approach was evaluated in a simulation study by comparison to cross-validation (CV). The ranking supports decision makers in omitting less influential monitoring wells from future sampling campaigns, thus optimising sample size. The proposed approach is currently implemented in GWSDAT.

While well influence analysis can be used to omit wells, it does not generate complete, spatiotemporal sampling designs, therefore, the following chapters will focus on this problem.

Chapter 4 introduces stochastic sampling techniques in the context of environmental monitoring and the concept of spatial balance. It provides an overview and explores the current literature on spatially balanced sampling approaches and their potential use in generating optimal designs for groundwater contamination monitoring. These approaches ensure that the selected samples are evenly spread over the target area, by avoiding the sampling of neighbouring locations and thus, allowing for better statistical analyses. The chapter also compares different spatially balanced sampling techniques, to select the most appropriate one to use in the following chapter for developing data-driven sampling designs. An important aspect considered in the comparison is the extendability of the technique to three dimensions (space and time). Based on the comparisons, the local pivotal methods (LPM) developed by Grafström et al. [2012], were found to be the most appropriate.

Chapter 5 then builds on Chapter 4 to develop a novel data-driven method to tune sample inclusion weights in the LPM spatially balanced sampling technique. The presented method is based on analysing the spatial evolution of the CoPC plume over time through historic observations and assigning a higher weight to monitoring wells that are predicted to be closer to the plume at the upcoming sampling campaign. The novel method was evaluated in a simulation study using synthetic contamination data by comparison with simple random sampling (SRS) and LPM with equal inclusion weights. The sampling designs with varying sample sizes drawn by these different methods were used to fit spatiotemporal models and estimate CoPC concentrations on the monitored site over time. Thus, the sampling designs were evaluated on how well they could be used to estimate these concentrations given decreasing sample sizes.

In Chapter 6, a novel application of spatiotemporally balanced sampling designs is proposed for evaluating the sufficiency of past sampling intensity based on historic data. By comparing the concentration estimates of models using the historic data and increasingly smaller subsamples selected via balanced designs, it can be assessed whether the monitoring network has been over- or undersampled. This information can then potentially be used to adjust future sampling intensity. The integration of the proposed approach in GWSDAT is also discussed. The chapter will also discuss the distribution of samples within a spatiotemporal design. In practice, it can be logistically advantageous to perform sampling less frequently, but obtain more samples during each campaign. This results in a spatially high, but temporally low resolution data set. The performed simulation study aims to identify the impacts of varying high spatial and temporal sampling intensity on estimates, given a fixed sample size and monitoring period.

Finally, Chapter 7 provides a summary of the results obtained throughout this thesis, with a discussion about their significance and potential for future developments.

Chapter 2

Spatiotemporal Groundwater Contamination Modelling Using GAMs

The benefits of using a joint spatiotemporal modelling framework instead of repeated spatial models in the context of analysing groundwater monitoring data have been demonstrated by McLean et al. [2019]. Using spatiotemporal P-splines to estimate solute concentration surfaces leads to more precise model predictions and is substantially less sensitive to the number of observations available when compared to repeated spatial models such as spatial P-splines. In spatiotemporal P-splines models, the covariates (spatial coordinates and time) are expressed in a non-parametric regression function, which consists of the linear combination of B-spline basis functions and corresponding basis coefficients (see equation 1.3 in Chapter 1).

The P-splines model introduced in Section 1.3.1 of Chapter 1 is a particular case of the broader generalised additive model (GAM) framework (Hastie and Tibshirani [1986]), which enables a variety of ways to approach spatiotemporal modelling. Due to their high flexibility, GAMs have been used frequently in environmental sciences, ecology, epidemiology (Ravindra et al. [2019]) and more recently in water resource management to estimate the spatial and temporal distribution of analytes. Despite their widespread use in groundwater level and vulnerability assessment (Mosavi et al. [2021]; Naghibi et al. [2017]; Zamanirad et al. [2020]; Motevalli et al. [2019]) and their ability to estimate solute concentration surfaces in a spatiotemporal framework, there are very few examples of their application in groundwater contamination monitoring (Sorichetta et al. [2013]).

This chapter aims to provide an introduction to the spatiotemporal analysis of groundwater contamination monitoring data, which will be used throughout the thesis to evaluate spatiotemporal sampling designs, through an outline comparison of various GAM approaches and the P-spline models also used by GWSDAT. The work presented in this chapter also provides an overview

on the characteristics of groundwater contamination data and various challenges with regards to creating sampling designs. The aim of the comparison was not a rigorous analysis, but an evaluation of different approaches to the spatiotemporal modelling of groundwater contamination using GAMs on case study and synthetic data sets. The most suitable GAMs were identified through a comparative study and their performances were assessed on the basis of different information criteria. The predicted concentration surfaces were also compared to the spatiotemporal P-splines approach via the output from GWSDAT. A simulation study using synthetic groundwater monitoring data was also performed using the most effective GAMs and they were evaluated based on their precision in estimating solute concentration surfaces.

2.1 Generalized Additive Models

Generalized additive models (GAMs) were first proposed by Hastie and Tibshirani [1986]. They are akin to generalized linear models (GLMs), which are a special case of GAMs, where the linear predictor is replaced by an additive predictor, which is a sum of unspecified smooth functions of the covariates. A wide variety of linear or non-linear smooth functions can be used to describe the relationships between the responses and the corresponding predictor variables, including splines and P-splines. The expected value of the response variable is then linked to the additive predictor through a link function such as the identity or log functions. While the predictor functions are generally non-parametric, GAMs may also contain parametric terms. Multi-dimensional smoothers are also possible within the GAM framework. Another definition of GAMs is provided by S. Wood et al. [2016] from the perspective of likelihood estimation. In statistical models with smooth regular likelihood, where the likelihood decomposes into a sum of independent terms, each contributed by a response variable from a single parameter exponential family distribution, then such a model is a GAM. Thus, for response variable Y an exponential family distribution is specified (such as normal, binomial or Poisson) along with a link function that relates the expected values of Y to the predictor variables x_i (such as the identity, log or inverse functions). The assumptions of exponential family distribution and link function for the response variables can be determined through an analysis of the distribution of the response data conditioned on the covariates, and by model selection criteria such as AIC and BIC. Similarly to linear regression, another assumption that should be investigated when modelling data using GAMs is the independence of errors.

In general, GAMs take on the following structure as described by S. Wood [2017], for example including parametric and two-dimensional terms as well:

$$g(\mathbf{E}(Y_i)) = X_i^* \theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + ... + f_m(x_{mi}), \tag{2.1}$$

where g is a link function, $\mathbf{E}(Y_i)$ is the expected value of the response variable Y_i for $i = 1, 2, ..., n$, X_i^* is a row of the model matrix for strictly parametric components, θ is the corresponding parameter vector and f_j; $j = 1, 2, ..., m$ are smooth functions of the covariates x_{ki}.

GAMs have been widely used to solve statistical problems due to their high flexibility in model specification (Fang and Chan [2015]). They have the same interpretability advantages as GLMs, where the contribution of each explanatory variable to the prediction is clearly defined. However, GAMs allow for more flexibility, because the relationships between dependent and independent variables are not assumed to be linear. The predictive functions that will be used do not need to be specified a priori (Larsen [2015]). Therefore, GAMs are especially useful when dealing with highly non-linear and non-monotonic relationships between response and explanatory variables (Fang and Chan [2015]). However, GAMs raise two theoretical problems. Namely, how to represent the smooth functions and specifying how smooth they should be (S. Wood [2017]).

In general, smoothers can be grouped into three classes: local regression (loess), smoothing splines and regression splines such as B-splines, P-splines and thin plate splines (Larsen [2015]). The P-splines modelling approach described in Section 1.3.1, which is also used in GWSDAT, is a particular case of the GAM framework. In the GWSDAT implementation, the three covariates (easting, northing and time) are included in a single smooth term, where the smoothing function is P-splines. In all cases, the smoothness of the functions is controlled by some form of regularization. This usually involves applying a smoothing parameter to the function, which can be specified or estimated from the data. The higher the value of the smoothing parameter, the smoother the function will become. As the smoothing parameter approaches its highest possible value, the function becomes linear. Estimating the smoothing parameter from the data is computationally expensive but can produce more accurate models (Larsen [2015]). The two most common approaches to estimating the smoothing parameter are the generalised cross-validation criteria (GCV; Golub et al. [1979]) and the mixed model approach via restricted maximum likelihood (REML; Corbeil and Searle [1976]). More novel approaches can be used as well such as that proposed by S. Wood [2004], based on QR decomposition and singular value decomposition. The implementation in GWSDAT uses a unique Bayesian framework to estimate the smoothing parameter. This framework was proposed by Evers et al. [2015] and is summarised in Section 1.3.1 in Chapter 1.

Variable selection is important when dealing with GAMs that contain a large number of variables. Measuring the strength of individual predictors can be done via different methods such as stepwise selection (forward and backward), ridge regression (Guisan et al. [2002]) and different shrinkage-based approaches (Marra and S. Wood [2011]). Term selection can also be automatized by modifying the smoothness selection methods to turn more penalized functions into the zero function, thus essentially removing them from the model (S. Wood [2017]). Interactive

term selection can be performed by comparing GCV or AIC scores of models with and without the term in question. Backwards stepwise selection involves looking at the individual p-values of the fitted model terms, dropping the least significant one and re-fitting until all terms are significant (S. Wood [2021]).

GAMs are ultimately estimated by maximizing the penalized likelihood function. This can be achieved by using a local scoring algorithm (extension of a backfitting algorithm) or in the case of regression splines, casting the GAM as a GLM and solving it using penalized iteratively reweighted least squares (PIRLS) (Larsen [2015]). As Fahrmeir et al. [2013] describes, if the functions in a GAM such as Equation 2.1 are estimated with basis functions, a large linear model is obtained. Using the vectors $\mathbf{f}_1, ..., \mathbf{f}_m$ of functions evaluated at the covariates $x_1, ..., x_m$, the model

$$\mathbf{y} = \mathbf{f}_1 + ... + \mathbf{f}_m + \mathbf{X}\beta + \varepsilon$$

can be expressed as

$$\mathbf{y} = \mathbf{V}_1\gamma_1 + ... + \mathbf{V}_m\gamma_m + \mathbf{X}\beta + \varepsilon,$$

with appropriately defined design matrices $\mathbf{V}_i$ and coefficient vectors γ_i. For every model component $\mathbf{f}_i = \mathbf{V}_i\gamma_i$, the vector γ_i can be decomposed as

$$\gamma_i = \tilde{\mathbf{X}}_i\beta_i + \tilde{\mathbf{U}}_i\tilde{\gamma}_i,$$

where $\tilde{\mathbf{X}}_i$ and $\tilde{\mathbf{U}}_i$ are design matrices. Using appropriate design matrices, vector β_i of fixed effects and vector $\tilde{\gamma}_i \sim N(0, \tau_j^2\mathbf{I})$ of independent and identically distributed random effects can be obtained. The vector of function evaluations can then be expressed as

$$\mathbf{f}_i = \mathbf{V}_i(\tilde{\mathbf{X}}_i\beta_i + \tilde{\mathbf{U}}_i\tilde{\gamma}_i) = \mathbf{X}_i\beta_i + \mathbf{U}_i\tilde{\gamma}_i,$$

with $\mathbf{X}_i = \mathbf{V}_i\tilde{\mathbf{X}}_i$ and $\mathbf{U}_i = \mathbf{V}_i\tilde{\mathbf{U}}_i$. By applying this decomposition to all model components, a large generalised linear mixed model (GLMM) is obtained (which is an extension of GLMs that also includes random effects in the linear predictor), where the fixed effects consist of β and $\beta_i, i = 1, ..., n$, and the random effects consist of $\tilde{\gamma}_i, i = 1, ..., n$.

GAMs can also be interpreted from a Bayesian perspective, because the penalty terms are equivalent to treating the coefficients of GLMs as random variables with normally distributed priors

(Larsen [2015]). The connection to Bayesian methods is clear, given that the estimation of smooth functions in a GAM imposes a prior belief that the predictive functions will follow smooth patterns. As mentioned above, the automatic smoothing parameter selection in the GWSDAT modelling framework also follows a Bayesian approach where the aim is to maximise the maximum a posteriori probability (MAP) estimate. This approach is discussed in in Section 1.3.1 in Chapter 1.

2.1.1 The Use of GAMs in Water Resource Management

Due to their high flexibility, GAMs have frequently been used to model environmental, ecological and epidemiological data (Ravindra et al. [2019]). The following are examples of studies where researchers used GAMs in the fields of hydrology and hydrogeology. These studies represent just a small subset of the literature available that demonstrates the potential benefits of using this modelling approach for water quality management. For example, Morton and Henderson [2008] advocate for the use of GAMs for the estimation of non-linear trends in water quality in the presence of correlated errors. The authors conclude that GAMs have many advantages that make them an appropriate tool for modelling water quality monitoring data. In a recent paper by Liu et al. [2020], a GAM was used to analyse water quality data to identify correlations between chlorophyll-a concentrations and other water quality parameters. GAMs have also been used in groundwater potential mapping. Falah et al. [2017] compared the prediction accuracy of GAMs with other popular GIS-based statistical methods such as Frequency Ratio (FR), Statistical Index (SI) and Weight-of-Evidence (WOE), to investigate groundwater spring potential. Their results showed that GAM was on par with the other established methods. Its prediction accuracy was only 8.4% lower than that of SI, which produced the best outcome with 85.4%.

With the increasing demand for groundwater as a freshwater resource, in recent years it has become more important to explore and develop accurate groundwater level prediction methods. The application of ensemble-based machine learning and data mining techniques for this purpose is becoming more widespread. GAMs can also be used in this context as a handful of papers have demonstrated. For example, Mosavi et al. [2021] evaluated the performance of four ensemble models (two bagging and two boosted models) including a GAMBoost model (Generalized additive model by likelihood based boosting) and reported satisfactory results with all four (albeit with the bagging models outperforming boosted models). Naghibi et al. [2017] performed a comparative assessment of five different ensemble models including a GAM and reported similar performance for all approaches in terms of groundwater well potential mapping. In another example, Zamanirad et al. [2020] used different machine learning algorithms including a GAM to model the influence of groundwater exploitation on land subsidence susceptibility. The authors concluded that in their case the GAM produced the best susceptibility

model. A practical application of GAMs in a data mining model was performed by Motevalli et al. [2019], who used different data mining techniques including GAMs to successfully evaluate the vulnerability of groundwater to salinization in a coastal aquifer. In another study, Sorichetta et al. [2013] evaluated the importance of different explanatory variables associated with nitrate contamination in groundwater, using different groundwater vulnerability assessment methods such as WOE and Logistic Regression (LR). The latter was extended using a GAM.

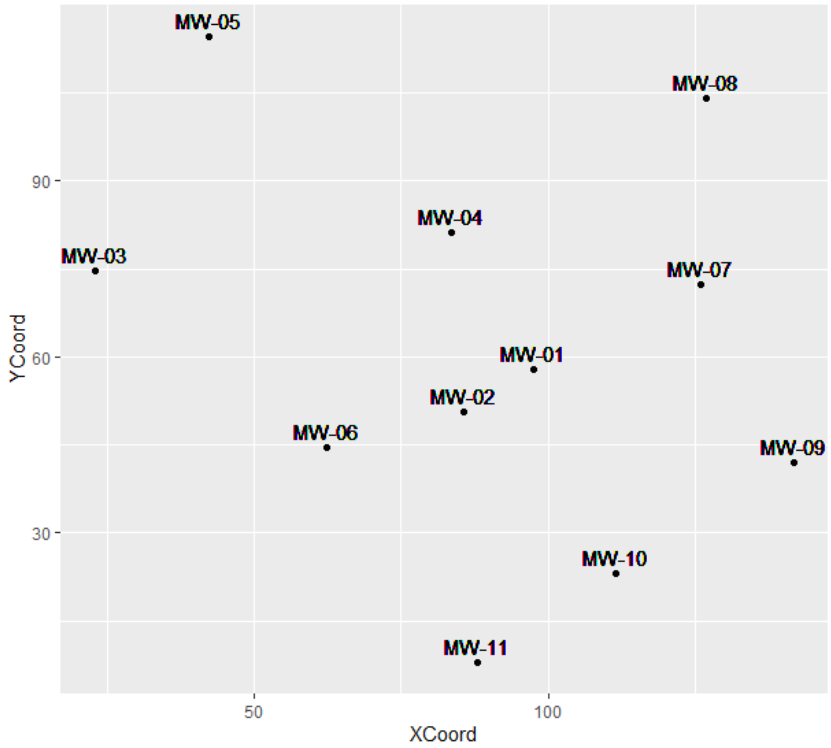
There is very little in the literature about GAMs being applied to groundwater quality monitoring data sets in particular. However, based on the above, there are potential benefits of using this modelling approach in this specific context. One of the major advantages of using GAMs on groundwater monitoring data is the ability to construct a spatiotemporal model. It is possible to represent the spatial and temporal effects of a model in a single smooth function within the GAM framework (see Section 2.2). The use of GAMs to identify and model spatiotemporal patterns has been well established. As McLean et al. [2019] showed, a spatiotemporal approach to modelling groundwater monitoring data might be more effective than the traditional, separated approaches. In conclusion GAMs appear to be appropriate tools for the explanatory and predictive modelling of spatiotemporal groundwater monitoring data.

2.2 Fitting GAMs to Groundwater Monitoring Data Using *mgcv* in R

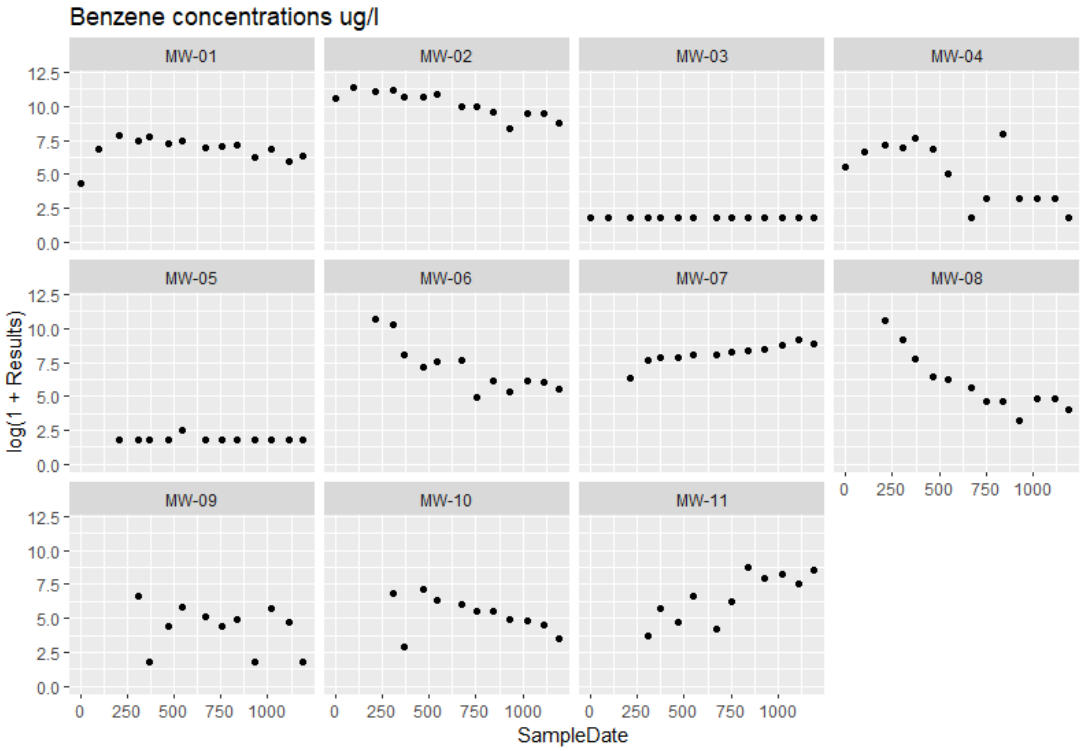
GAMs were fitted to the case study groundwater monitoring data sets 1 and 2 described in Sections 1.4.3 and 1.4.4 respectively, using the *mgcv* package (S. Wood [2021]) in R (R Core Team [2020]). In the fitted models, the measured concentrations of benzene and ethylbenzene were considered as the response variables for case study 1 and case study 2 respectively. For both data sets, the monitoring well coordinates (easting and northing) and the sampling dates (time) were the explanatory variables, representing the spatial and temporal dimensions respectively. Overall, data set 1 contained 137 observations of benzene concentrations collected from 11 wells over a 4-year period (see Figure 2.1), while the data set 2 contained 382 ethylbenzene observations collected from 32 monitoring wells, also over a 4-year period (see Figure 2.2). Six different model structures or approaches were investigated to find the most optimal one i.e. the one that produced the best fit for the data. The aim of the models was to produce a smooth spatiotemporal estimate of the CoPC concentration surfaces for the monitoring sites. The investigated approaches differed in which covariates were estimated jointly in smooth terms and what types of smooth functions were applied. The following sections will detail how the six model structures were constructed.

For all modelling approaches, the number of basis functions k was chosen based on recommendations from S. Wood [2021]. According to the author, the choice of k is generally not critical. It should be large enough to maintain enough degrees of freedom to represent the underlying truth reasonably well, but small enough to maintain reasonable computational efficiency. In this study, k was ultimately chosen to minimise AIC through trial and error, while satisfying the above conditions. The *gam.check* function contains an assessment of how adequate the chosen value of k was in the models. This check was also performed to verify that the selected k value was reasonable. The assessment relies on a metric named the k-index, for which values < 1 indicate a too small k value (S. Wood [2017]). The smoothing parameter was estimated automatically using the GCV criterion. The standard errors of the fitted values are estimated in *mgcv* based on the Bayesian posterior covariance matrix of the parameters, V_p in the fitted GAM object. The significance of smooth terms in the GAMs was evaluated using the *summary.gam* function in *mgcv*. The p-values indicating the significance of smooth terms are based on a test statistic that analyses the frequentist properties of Bayesian confidence intervals for smooths and it was introduced byMARRA and WOOD [2012]. The details of this test are further discussed in S. Wood [2021].

A $log(y+1)$ transformation was applied to the observed CoPC concentrations in both data sets (see Sections 1.4.3 and 1.4.4). In both groundwater monitoring data sets the concentration distributions were skewed towards low concentrations and non-detects, as shown on Figure 2.3. The log transformation was applied to normalise the distribution of concentration values to minimise the impact of this low tail on the model estimates. This skewness is also a common feature of groundwater contamination monitoring data, since often not all wells are effected by the contamination in the monitoring network. Hence, most of the data collected over time will include non-detects (concentrations close to or equal to zero). The one was added to the observed concentration values to mitigate the presence of these non-detect, or zero values. Figure 2.4 shows the distribution of log-transformed concentration values in case study data set 1 (see Section 1.4.3) to illustrate the effect of log-transformation on the data. In this case, zero inflation can be observed in the data distribution even after transformation. A Box-Cox transformation (Box and Cox [1964]) could also be applied to normalise the distribution of concentrations, however the log-transformation is used in the GWSDAT modelling framework, therefore this transformation will be utilised throughout this analysis. The evaluation of the models was performed on this transformed scale. The measurement errors associated with the analytical methods that determine the solute concentrations in samples and applied to the simulated concentration values were assumed to be normally distributed, therefore, the identity function was used as link function in the GAMs.

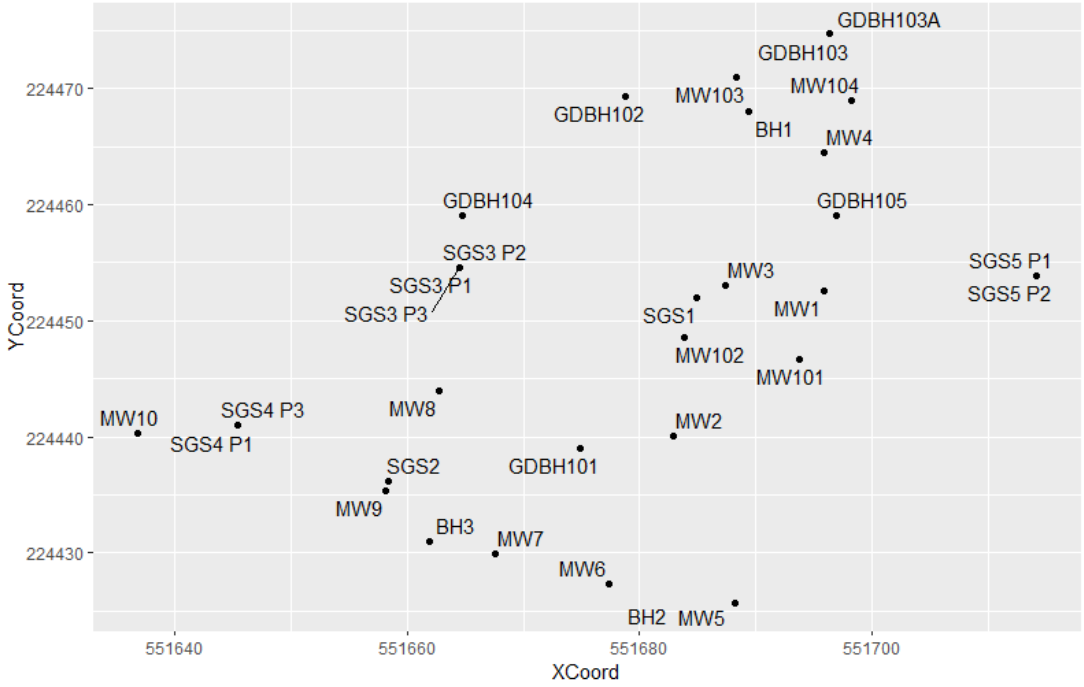


(a) Monitoring well coordinates in case study 1

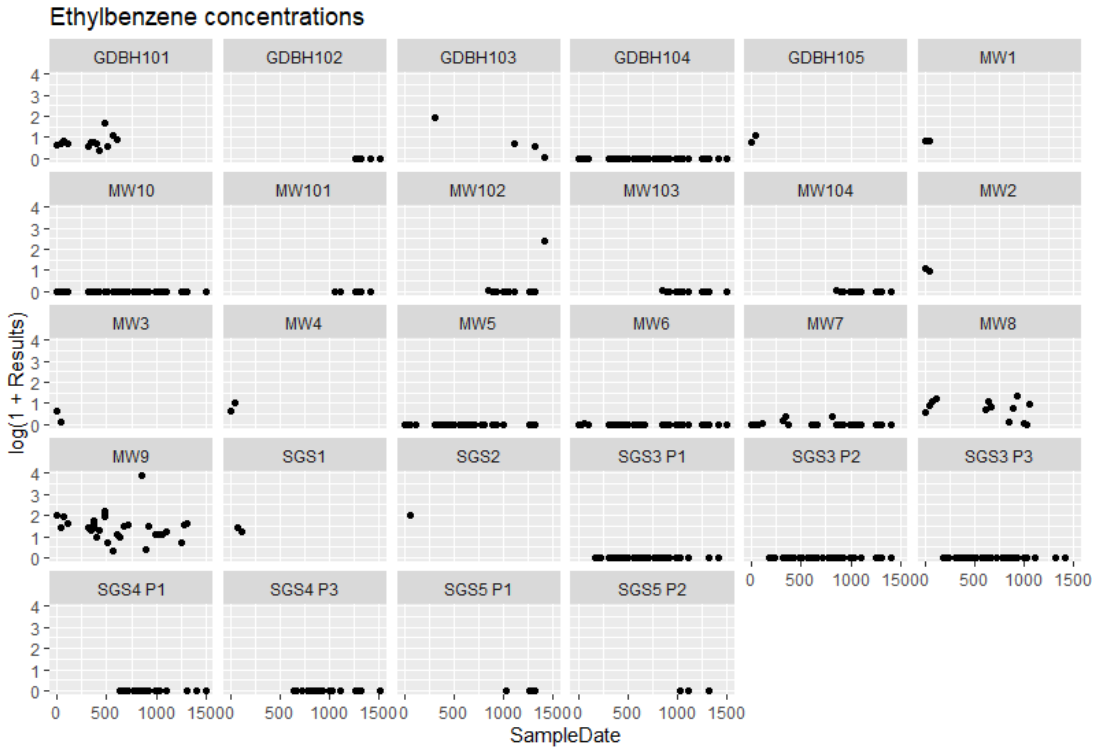


(b) Log benzene concentrations in monitoring wells observed over time

Figure 2.1: Observations of the log-transformed concentration of benzene (μg/l) in groundwater samples obtained from the 11 monitoring wells over the 4-year monitoring period in case study 1.



(a) Monitoring well coordinates in case study 1



(b) Log benzene concentrations in monitoring wells observed over time

Figure 2.2: Observations of the log-transformed concentration of benzene (μg/l) in groundwater samples obtained from the 32 monitoring wells over the 4-year monitoring period in case study 2.

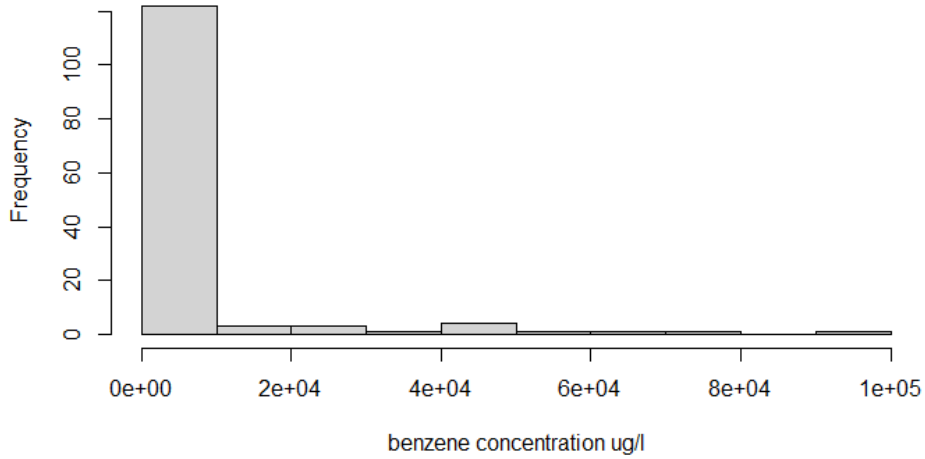


Figure 2.3: Distribution of observed benzene concentration values in case study data set 1 (see Chapter 1.4.3), in $\mu g l^{-1}$.

Figure 2.4: Distribution of log-transformed benzene concentration values in case study data set 1 (see Chapter 1.4.3), $\mu g l^{-1}$.

2.2.1 Separate thin plate splines approach

In this approach, the monitoring well coordinates (spatial effects) and sampling dates (temporal effects) were represented using separate smooth functions (bivariate and univariate respectively). These smooth functions of the spatial and temporal components were constructed using penal-

ized thin plate regression splines (S. N. Wood [2003]). The model took the following form:

$$E(\mu_i) = \beta_0 + f_{12}(x_{i1}, x_{i2}) + f_3(x_{i3}) \qquad i = 1, 2, ..., n, \tag{2.2}$$

where $E(\mu_i)$ is the expected value of the log transformed response $\mu_i = log(1 + y_i)$, $f_{12}(x_{i1}, x_{i2})$ is a bivariate smooth function (thin plate spline) of the covariates representing the spatial coordinates of the monitoring wells, $f_3(x_{i3})$ is a smooth function (thin plate spline) of the covariates representing the sampling time points and β_0 is the intercept term while n denotes the number of observations.

Thin plate regression splines rely on the penalization of the integral of the square of the second derivative as its smoothness measure. The example model 1.1 provided in Chapter 1, can be estimated using thin plate splines by finding the function that minimises

$$||\mathbf{y} - \mathbf{f}||^2 + \lambda \int f''(x)^2 dx, \tag{2.3}$$

where $\mathbf{y}$ is a vector of y_is, $\mathbf{f}$ is the corresponding vector of $f(x_i)$ values and $||\cdot||$ is the Euclidean norm. λ is the smoothing parameter which balances the badness of fit measured by the first term and model wiggliness measured by the second term. Thin plate splines are obtained as the solution of the generalisation of this minimisation to problems in which f is a function of any finite number $d \geq 1$ of covariates and the order m of differentiation in the wiggliness penalty can be any integer satisfying $2m > d$ (S. N. Wood [2003]).

Case study data 1

Based on the *summary.gam* function of the model in *mgcv*, both smooth terms were significant with p-values of $< 2e^{-16}$ and 0.000157 for spatial and temporal terms respectively. The values of k were 11 for the spatial smooth term and 3 for the temporal smooth term, selected to minimise the AIC values. These values of k were shown to be adequate by the k-index values in the *gam.check* model diagnostics. The log-transformed observations of case study data 1 ranged from 1.792 to 11.430. The fitted values of the model ranged from 0.995 to 10.881 and had a median estimated standard deviation of 0.405 with a range of 0.361 to 0.440. Figure 2.5 shows the resulting model with 95% confidence intervals compared to the log-transformed observations.

Case study data 2

Based on the *summary.gam* function of the model in *mgcv*, only the spatial smooth term was significant with a p-value of $< 2e^{-16}$ while the temporal smooth term had a p-value of 0.156. The values of $k = 24$ for the spatial smooth and $k = 5$ for the temporal smooth were shown to be adequate by the k-index values in the model diagnostics. The log-transformed observations of case study data 2 ranged from 0.0005 to 3.892. The fitted values of the model ranged from -0.044 to 1.832 and had a median estimated standard deviation of 0.048 with a range of 0.029 to 0.217. Figure 2.6 shows the resulting model with 95% confidence intervals compared to the log-transformed observations.

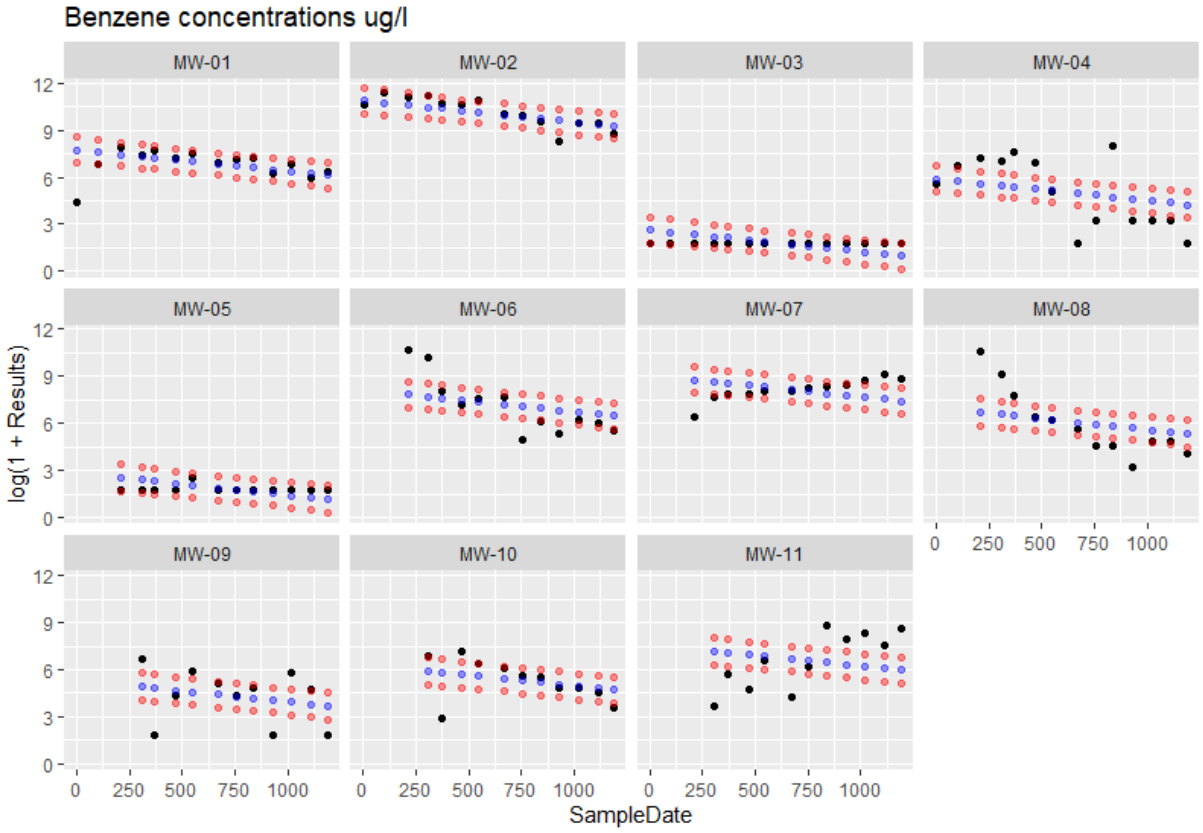


Figure 2.5: Observed concentrations over time (black), fitted values (blue) from the separate thin plate splines model and corresponding 95% confidence intervals (red) for each monitoring well in case study 1.

2.2.2 Separate tensor product splines approach

The second approach was only different from the first approach in that it used penalized tensor product splines (see Equation 1.3 in Chapter 1 for the formal definition) to represent the smooth functions instead of penalized regression splines. In contrast to thin plate regression splines,

tensor product splines allow different degrees of smoothness in the dimensions represented by the function.

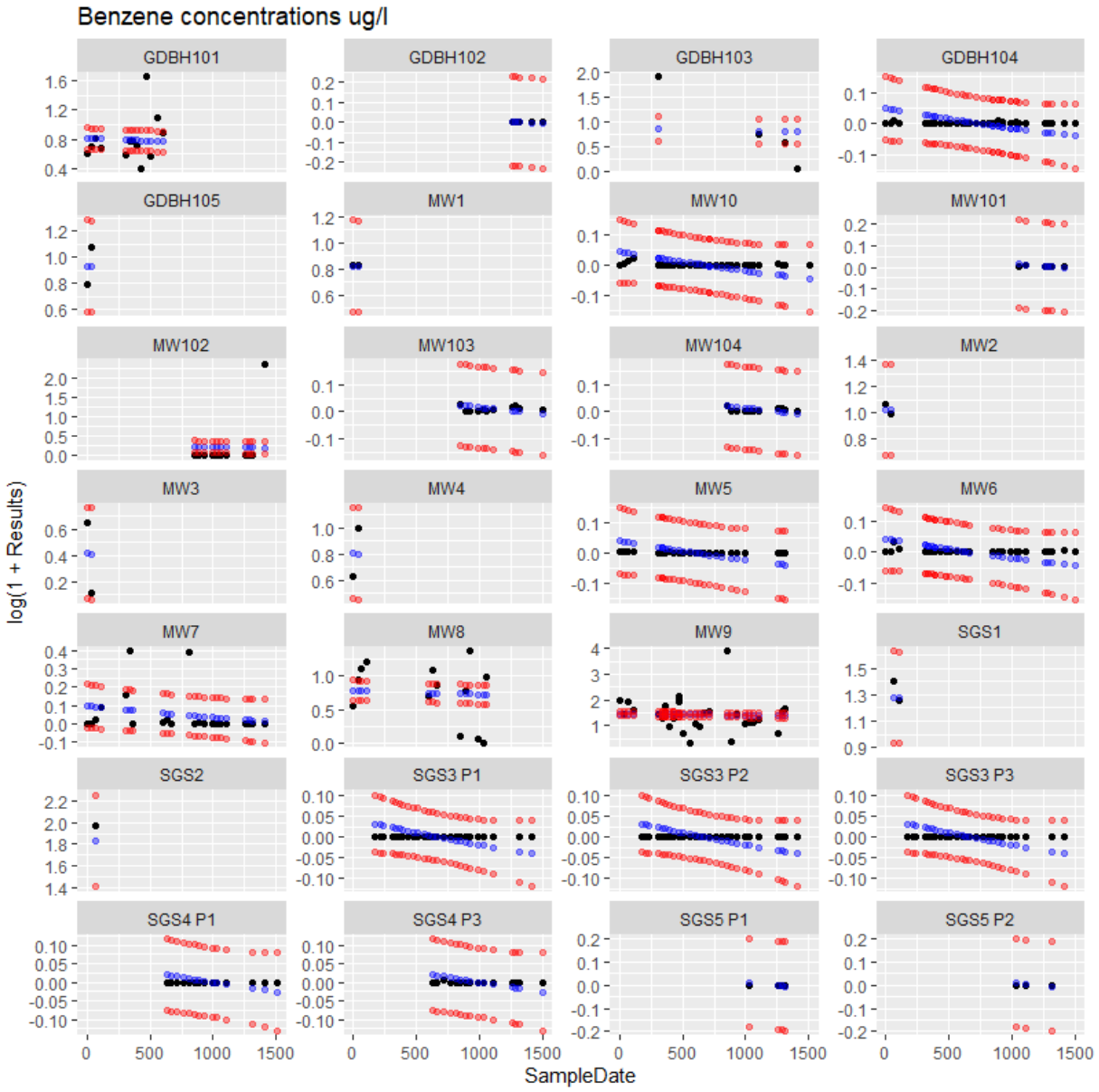


Figure 2.6: Observed concentrations over time (black), fitted values (blue) from the separate thin plate splines model and corresponding 95% confidence intervals (red) for each monitoring well in case study 2.

The model took the following form:

$$E(\mu_i) = \beta_0 + f_{12}(x_{i1}, x_{i2}) + f_3(x_{i3}) \qquad i = 1, 2, ..., n, \tag{2.4}$$

where $E(\mu_i)$ is the expected value of the log transformed response $\mu_i = log(1 + y_i)$. The spatial

explanatory variables were x_{i1} (x coordinates) and x_{i2} (y coordinates). The temporal explanatory variable was x_{i3} (sampling dates). f_{12} and f_3 are smooth functions (tensor product splines) and β_0 is the intercept term. n denotes the number of observations.

Case study data 1

Based on the *summary.gam* function of the model in *mgcv*, both smooth terms were significant with p-values of $< 2e^{-16}$ and 0.000157 for spatial and temporal terms respectively. Similarly to the separate thin plate splines approach, the values of k were 11 for the spatial smooth term and 3 for the temporal smooth term, selected to minimise the AIC values. These values of k were shown to be adequate by the k-index values in the *gam.check* model diagnostics. The fitted values of the model ranged from 0.991 to 10.901 and had a median estimated standard deviation of 0.406 with a range of 0.362 to 0.441.

Case study data 2

Based on the *summary.gam* function of the model in *mgcv*, only the spatial smooth term was significant with a p-value of $< 2e^{-16}$ while the temporal smooth term had a p-value of 0.159. The values of $k = 10$ for the spatial smooth term and $k = 5$ in the temporal smooth term were shown to be adequate by the k-index values in the model diagnostics. The fitted values of the model ranged from -0.044 to 1.867 and had a median estimated standard deviation of 0.048 with a range of 0.029 to 0.230.

2.2.3 Trivariate thin plate splines approach

In the third approach, a single three-dimensional penalized thin plate regression spline is used to represent the smooth function. The resulting trivariate smooth term contained both spatial and the temporal covariates. The model took the following form:

$$E(\mu_i) = \beta_0 + f_{123}(x_{i1}, x_{i2}, x_{i3}) \qquad i = 1, 2, ..., n, \tag{2.5}$$

where $E(\mu_i)$ is the expected value of the log transformed response $\mu_i = log(1 + y_i)$. The spatial explanatory variables were x_{i1} (x coordinates) and x_{i2} (y coordinates). The temporal explanatory variable was x_{i3} (sampling dates). f_{123} is the penalized thin plate regression spline smoother and β_0 is the intercept term. n denotes the number of observations.

Case study data 1

Based on the *summary.gam* function of the model in *mgcv*, the smooth term was significant with a p-value of $< 2e^{-16}$. The value of $k = 12$, selected to minimise the corresponding AIC, was also shown to be adequate by the k-index values in the *gam.check* model diagnostics function. The fitted values of the model ranged from -0.052 to 8.323 and had a median estimated standard deviation of 0.479 with a range of 0.295 to 0.834.

Case study data 2

Based on the *summary.gam* function of the model in *mgcv*, the spatiotemporal smooth term was significant with a p-value of $< 2e^{-16}$. The value of $k = 20$ in the trivariate smooth term was shown to be adequate by the model diagnostics produced by the *gam.check* function, with a k-index value of 1.15. The fitted values of the model ranged from -0.504 to 0.928 and had a median estimated standard deviation of 0.058 with a range of 0.041 to 0.159.

2.2.4 Trivariate tensor product splines approach

The fourth approach modified approach 3 in that the three-dimensional smoother was a penalized tensor product spline instead of the penalized thin plate regression spline. According to S. Wood [2021] this approach is much less costly computationally in multi-dimensional cases. This model structure is the most similar to the GWSDAT framework (see Section 1.3.1 in Chapter 1). The main difference between this and the GWSDAT implementation is the method used for the estimation of the smoothing parameters. Whilst GWSDAT uses the Bayesian approach proposed by Evers et al. [2015], this implementation used generalised cross-validation (Golub et al. [1979]). The model took the following form:

$$E(\mu_i) = \beta_0 + f_{123}(x_{i1}, x_{i2}, x_{i3}) \qquad i = 1, 2, ..., n, \tag{2.6}$$

where $E(\mu_i)$ is the expected value of the log transformed response $\mu_i = log(1 + y_i)$. The spatial explanatory variables were x_{i1} (x coordinates) and x_{i2} (y coordinates). The temporal explanatory variable was x_{i3} (sampling dates). f_{123} is the tensor product spline smoother and β_0 is the intercept term. n denotes the number of observations.

Case study data 1

Based on the *summary.gam* function of the model in *mgcv*, the smooth term was significant with a p-value of $< 2e^{-16}$. The value of $k = 4$ was also shown to be adequate by the k-index values in the model diagnostics. The fitted values of the model ranged from 1.658 to 11.231 and had a median estimated standard deviation of 0.396 with a range of 0.326 to 0.680. Figure 2.7 shows the resulting model with 95% confidence intervals compared to the log-transformed observations.

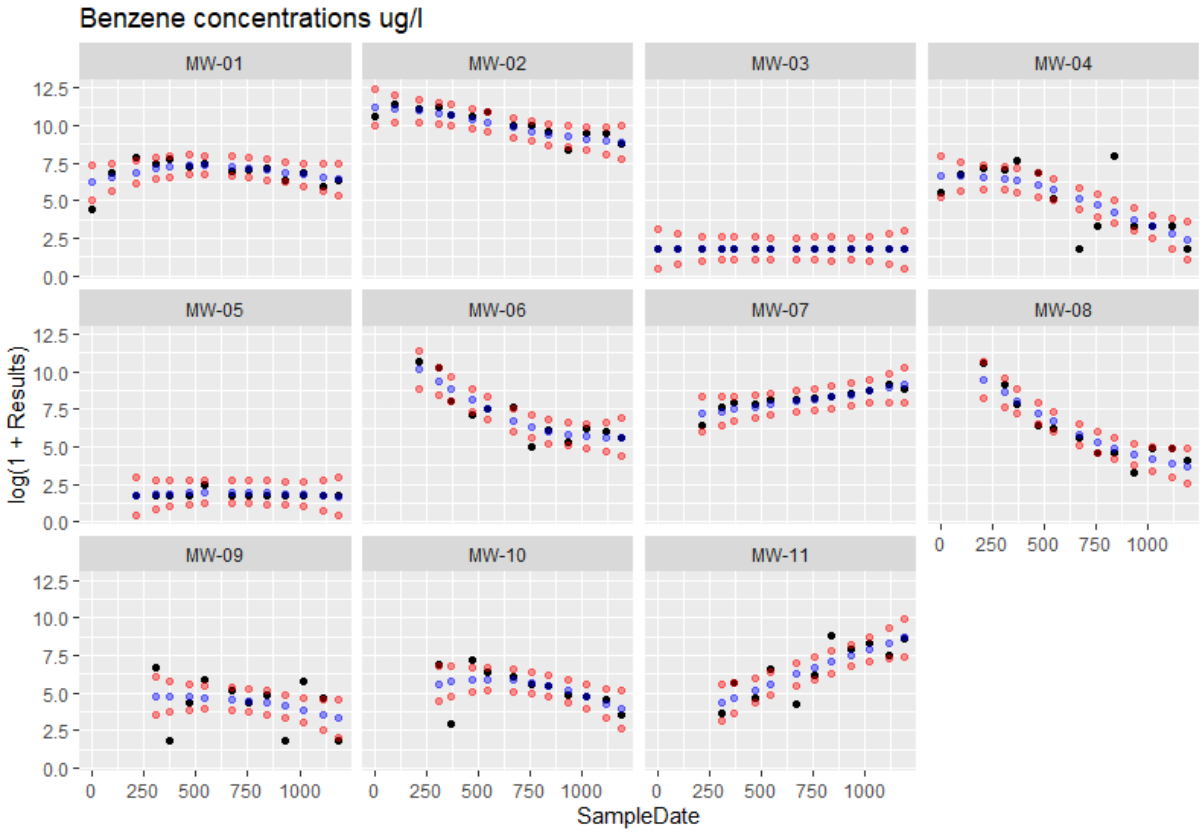


Figure 2.7: Observed concentrations over time (black), fitted values (blue) from the trivariate tensor product splines model and corresponding 95% confidence intervals (red) for each monitoring well in case study 1.

Case study data 2

Based on the *summary.gam* function of the model in *mgcv*, the spatiotemporal smooth term was significant with a p-value of $< 2e^{-16}$. The value of $k = 5$ for the trivariate smooth term was shown to be adequate by the model diagnostics produced by the *gam.check* function, with a k-index value of 1.08. The fitted values of the model ranged from -0.075 to 1.877 and had a median estimated standard deviation of 0.077 with a range of 0.040 to 0.232. Figure

2.8 shows the resulting model with 95% confidence intervals compared to the log-transformed observations.

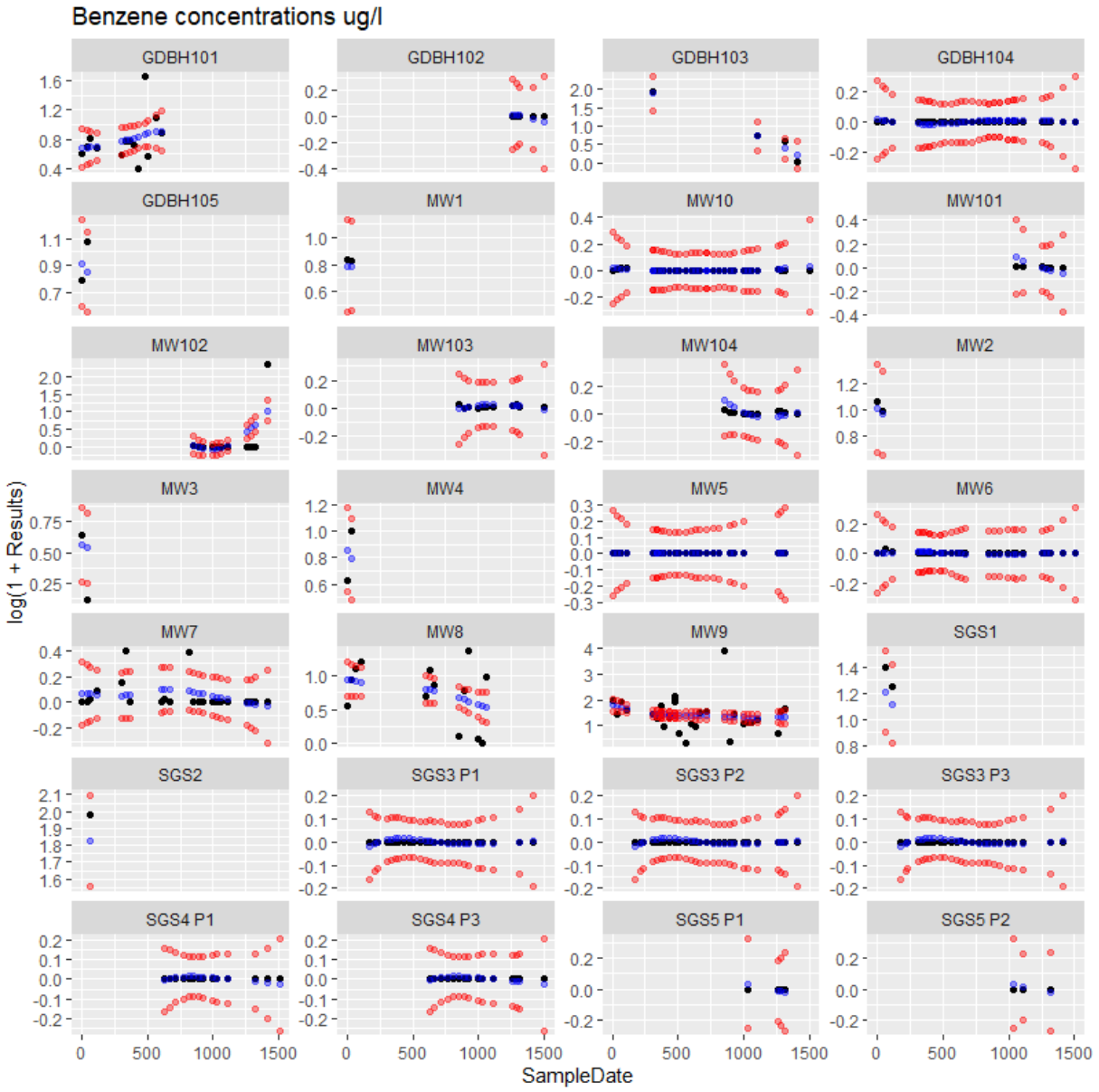


Figure 2.8: Observed concentrations over time (black), fitted values (blue) from the trivariate tensor product splines model and corresponding 95% confidence intervals (red) for each monitoring well in case study 2.

2.2.5 Well-based thin plate splines approach

In the fifth approach, a penalized thin plate regression spline smoother contained the sampling dates but a replicate of the smooth was produced for each level of a factor variable representing each monitoring well. This was done by using the *by* function of the *mgcv* package. Due to the

monitoring well name being a factor variable, centering constraints were applied to the smooth by the function. According to S. Wood [2021], because of this, the *by* variable had to be included as a parametric term as well. The model took the following form:

$$E(\mu_i) = \beta_0 + f(x_{i3})\gamma_i \qquad i = 1,2,...,n, \tag{2.7}$$

where $E(\mu_i)$ is the expected value of the log transformed response $\mu_i = log(1+y_i)$ with $i = 1,2,...,n$ representing the number of observations. The sampling dates are represented by the explanatory variable x_{i3}. f is a thin plate regression spline smoother and β_0 is the intercept term. γ_i is a factor variable corresponding to each observation indicating which monitoring well the observation originates from.

Case study data 1

Based on the summary of the model in *mgcv*, the smooth terms for the individual wells were significant except for wells 3 and 5 with p-values of 1 and 0.874 respectively. The values of $k = 11$ for all wells were shown to be adequate by the k-index values in the model diagnostics. The fitted values of the model ranged from 1.522 to 11.428 and had a median estimated standard deviation of 0.315 with a range of 0.171 to 0.630.

Case study data 2

Based on the summary of the model in *mgcv*, the smooth terms for the individual wells were mostly insignificant with some exceptions. The values of $k = 4$ for all factor levels were shown to be adequate by the model diagnostics produced by the *gam.check* function, with a k-index value of 1.04 for all wells. The fitted values of the model ranged from -0.409 to 1.978 and had a median estimated standard deviation of 0.065 with a range of 0.038 to 0.225.

2.2.6 Well-based tensor product splines approach

The sixth approach was different from the fifth in that it used a penalized tensor product spline to represent the smooth rather than a penalized regression spline. The model took the following form:

$$E(\mu_i) = \beta_0 + f(x_{i3})\gamma_i \qquad i = 1,2,...,n, \tag{2.8}$$

where $E(\mu_i)$ is the expected value of the log transformed response $\mu_i = log(1+y_i)$ with $i = 1,2,...,n$ representing the number of observations. The sampling dates are represented by the explanatory variable x_{i3}. f is a tensor product spline smoother and β_0 is the intercept term. γ_i is a factor variable corresponding to each observation indicating which monitoring well the observation originates from.

Case study data 1

Based on the summary of the model in *mgcv*, the smooth terms for the individual wells were significant except for wells 3 and 5 with p-values of 1 and 0.846 respectively. The values of $k = 11$ for all factor levels were shown to be adequate by the k-index values in the model diagnostics. The fitted values of the model ranged from 1.566 to 11.158 and had a median estimated standard deviation of 0.306 with a range of 0.140 to 0.513.

Case study data 2

Based on the summary of the model in *mgcv*, the smooth terms for the individual wells were mostly insignificant with some exceptions. The values of $k = 5$ for all factor levels were shown to be adequate by the model diagnostics produced by the *gam.check* function, with a k-index value of 1.03 for all wells. The fitted values of the model ranged from -0.369 to 1.978 and had a median estimated standard deviation of 0.065 with a range of 0.038 to 0.229.

2.3 Model Selection

The six approaches were evaluated using AIC and BIC criterion, adjusted R^2 values as well as residual diagnostics using the *gam.check* function within the *mgcv* package. This produces residual plots, Q-Q plots, provides information about the convergence of the smoothing parameter estimation and can indicate whether the number of basis functions, k was adequate. In general, BIC is better suited for the selection of explanatory models whilst AIC is better suited for the selection of predictive models (Chakrabarti and Ghosh [2011a]). In this study, both AIC and BIC were evaluated to obtain an indication as to which models might perform better at explanatory and predictive modelling.

2.3.1 Case study data 1

Table 2.1 shows the model performance metrics for data set 1, which contained benzene concentrations from 11 wells over 4 years of monitoring (see Figure 2.1 and Section 1.4.3 in Chapter 1). The two separate smooth term approaches produced very similar model fits, with penalized thin plate regression splines performing slightly better, but not significantly. Both approaches explained about 76.5% of the variation in the data. The result indicates that with separate smooths for spatial and temporal effects, the choice of smoother type (penalized tensor product splines or penalized thin plate regression splines) is not critical. Both types resulted in similarly good fits for the data.

The trivariate smooth case showed a different trend. When spatial and temporal effects were included in the same smooth, the choice of smoothing function significantly impacted the outcome. Model 4 performed better using tensor product splines than model 3 using penalized thin plate regression splines. In fact Model 4 even outperformed Models 1 and 2, while Model 3 did not. The reason for this discrepancy could be the different k value. However, both cases were optimized through trial and error, therefore a more fundamental difference between the two approaches is more likely causing the different performances.

Table 2.1: Comparison of performance metrics for GAMs fitted to case study data set 1 (see Section 1.4.3 in Chapter 1) using *mgcv*.

Model	No. of Smoothers	k	Smoother Type	AIC	BIC	R^2
1	2	11, 3	thin plate regression spline	488.05	525.63	0.765
2	2	11, 3	tensor product spline	488.10	525.83	0.765
3	1	12	thin plate regression spline	563.46	595.58	0.587
4	1	4	tensor product spline	411.20	497.66	0.879
5	1 (per well)	11	thin plate regression spline	303.24	458.49	0.95
6	1 (per well)	11	tensor product spline	251.30	439.80	0.967

The well-based models 5 and 6 seemed to provide the best fit out of the 6 approaches. Between these models, Model 6 performed better with the tensor product spline smoother. The high R^2 values (95% and 96.7% for models 5 and 6 respectively) indicate that the variance in the CoPC concentration data is almost perfectly explained by the these models. In the well-based modelling approaches, each smooth function only fits the observations of a single well, whilst having a high flexibility with $k = 11$. In contrast models 1-4 use 1 or 2 smooth functions for all observations. This difference in the number of data points to be fitted could explain the high R^2 values observed for models 5 and 6. However, these high R^2 values could also indicate that the models are overfitting, which would reduce their utility in interpolating concentration

values over time. Plotting the well-based models did in fact reveal overfitting in wells that provided more variable observations. Moreover, the more fundamental problem with this model design is that it only considers the time-series of CoPC concentrations at the sampling locations separately. It does not aim to produce an estimate of the CoPC concentration surface, making it unsuited for deriving spatial patterns. To do so, would require additional interpolation.

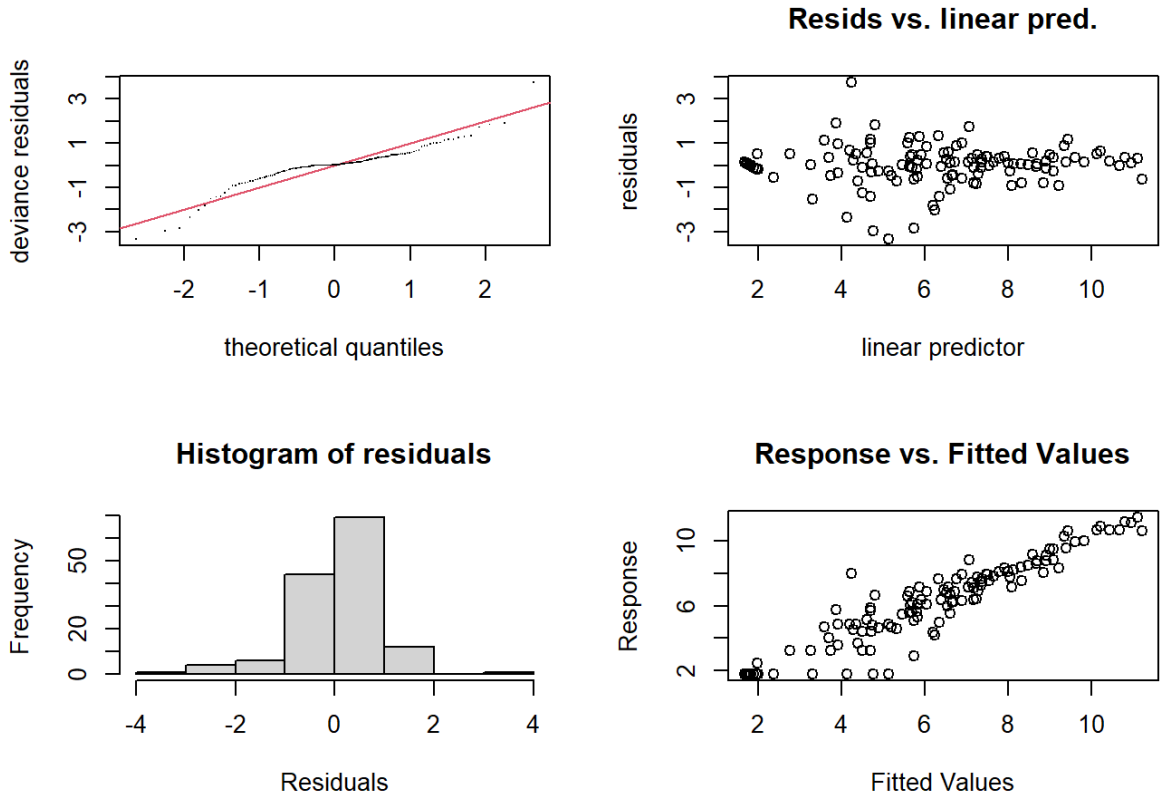


Figure 2.9: Residual diagnostics for the trivariate tensor product splines smooth model, applied to the observations of benzene concentrations from case study data 1 (see Section 1.4.3 in Chapter 1).

Another aspect to consider is that the trivariate smooth term models 3 and 4 were computationally the most efficient and least costly, since the smoothing parameter estimation only had to be performed once. Whereas, the separate smooth term models 1 and 2 need to estimate the smoothing parameters for the spatial and temporal terms separately. Therefore, overall, the trivariate spatiotemporal smooth term model using tensor product splines (model 4) seems to be the most adequate choice for modelling CoPC concentrations in data set 1. Figure 2.9 shows the residual diagnostics for this approach. The Q-Q plot of the residuals indicates that their distribution has heavier tails than would be expected of a normal distribution. The histogram and the response vs. fitted values plots suggest higher variance and more overestimation of low concentration values by the model. The concentration data is dominated by low concentrations

(zero-inflation) and the ballooning of fitted values at certain spatial locations could be causing this trend. The aim here is the description of the spatial and temporal trends, but for rigorous statistical inference, the assumption of normally distributed residuals would warrant more examination.

The autocorrelation of the residuals was examined for each monitoring well to verify the assumption that residuals are independent. No indication of residual covariance was observed. The variograms of the residuals at each time slice also showed no indication of spatial correlation. This suggested that the models captured the spatial and temporal patterns in the contamination data well.

Overall the diagnostics indicate that this approach results in a reasonable fit for the contamination data to allow for the analysis of spatial and temporal trends, but the fit could be further improved by addressing the trends observed in the residual diagnostics. Model 4 is also the most similar to the P-splines framework used in GWSDAT, as it includes all three covariates in a single smooth, spline-based function, and allows for varying degrees of smoothness across the covariate dimensions. It should be noted that the trend exhibited in the bottom left corner of the response versus fitted values and residuals versus linear predictor plots is the result of the lower limit imposed on the concentration values by the applied transformation (see Section 2.2), in which adding 1 to the observations before the log transformation fixed the lowest possible value.

2.3.2 Case study data 2

Table 2.2 shows the model performance metrics for data set 2, which contained ethylbenzene concentrations from 32 wells over 4 years of monitoring (see Figure 2.2 and Section 1.4.4 in Chapter 1). Just as in the previous case, Approaches 1 and 2 showed a very similar performance, with the former achieving a slightly better fit. The R^2 values indicate that both models explain about 76.2% of the variation in the data. The results show that the choice of smoother is once again, not critical when spatial and temporal effects are estimated by separate smooths, i.e. the two models produced equally adequate fits. Based on the BIC scores alone, these two approaches produce the best model fit.

The single smooth models exhibited the same behaviour with respect to each other as with the data set 1. Model 4 performed better than Model 3, indicating that using tensor product splines as smooth functions rather than thin plate regression splines is preferred when using a single spatiotemporal smooth. Based on the model scores, Model 3 produced a significantly worse fit that Approaches 1, 2 or 4. Model 4 outperformed Models 1 and 2 when looking at the AIC

scores and the R^2 values, but not when looking at the BIC scores. It might be the case that Model 4 fits the data better than Models 1 and 2 but is farther from the truth, since AIC tends to select more complex models than BIC due to the lower degree of penalization of complexity in AIC (Fahrmeir et al. [2013]). Figure 2.10 shows the residual diagnostic plots for approach 4, to allow for comparison with the diagnostics of the same approach applied to case study data 1. It should be noted that the concentration estimates are shown on the log-scale, hence the appearance of negative concentration values.

Table 2.2: Comparison of performance metrics for GAMs fitted to case study data set 2 (see Section 1.4.4 in Chapter 1) using *mgcv*.

Model	No. of Smoothers	k	Smoother Type	AIC	BIC	R^2
1	2	24, 5	thin plate regression spline	66.89	168.20	0.762
2	2	10, 5	tensor product spline	66.94	168.47	0.762
3	1	20	thin plate regression spline	424.12	469.17	0.371
4	1	5	tensor product spline	39.25	261.99	0.794
5	1 (per well)	4	thin plate regression spline	0.36	235.49	0.815
6	1 (per well)	5	tensor product spline	15.01	249.63	0.808

The autocorrelation of the residuals was examined for each monitoring well to verify the assumption that residuals are independent. No indication of residual correlation was observed. The variograms of the residuals at each time slice also showed no indication of spatial correlation. This suggested that the models captured the spatial and temporal patterns in the contamination data well. The diagnostics indicate that the model fit is not as adequate as in the case of data set 1. The residual distribution is slightly right skewed. The Q-Q plot also indicates deviance from normal distribution at the higher quantiles, due to the zero-inflated distribution of the data. The fitted values generally match the observations, but outliers are present.

Similarly, Models 5 and 6 seemed to produce the best model fit when looking at the AIC scores and R^2 values, but not if the BIC scores were also taken into account. However, next to the problem of overfitting, as discussed above in Section 2.3.1, Models 5 and 6 are conceptually not adequate for producing spatiotemporal estimates of the CoPC concentration surface.

Overall, Models 1,2 and 4 seemed to be the best models for the data set 2 dataset. Given the computational efficiency of the trivariate smooth approach (see Section 2.3.1) and the lower AIC score, Model 4 appears to be the most adequate choice in this case.

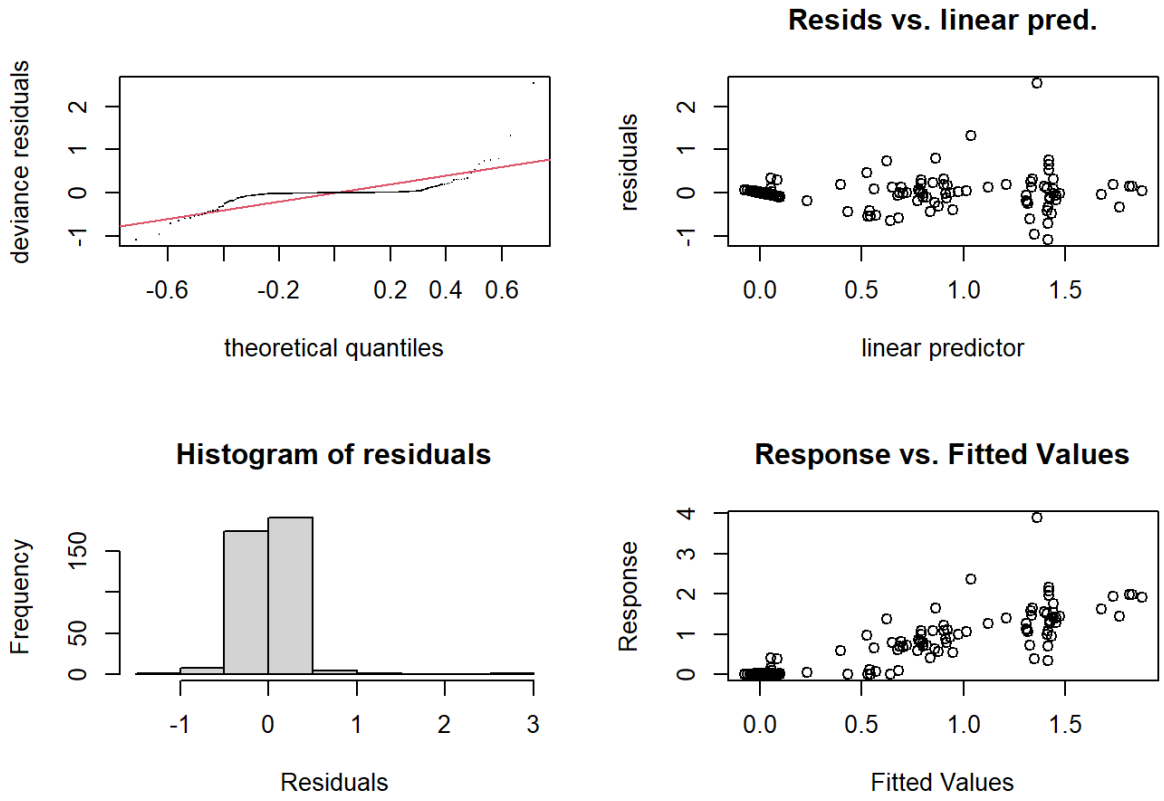
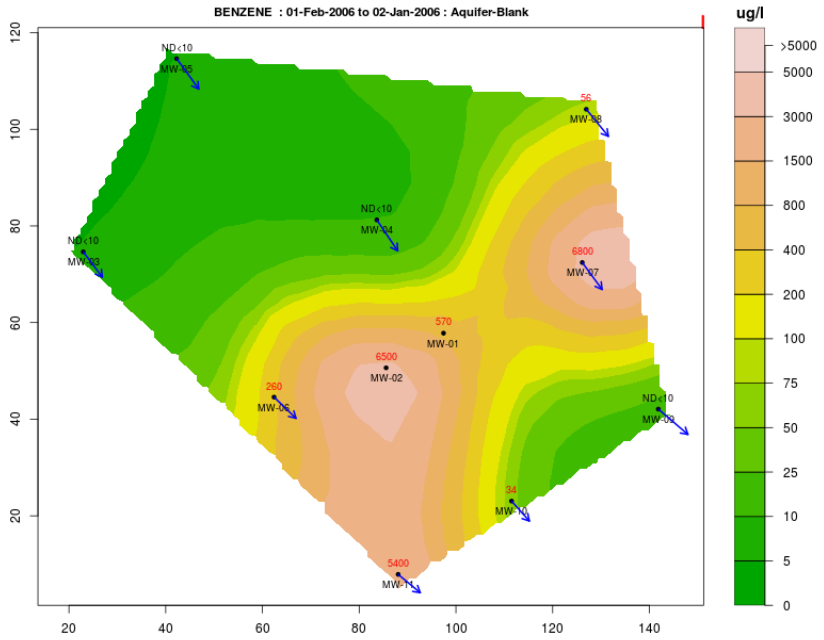


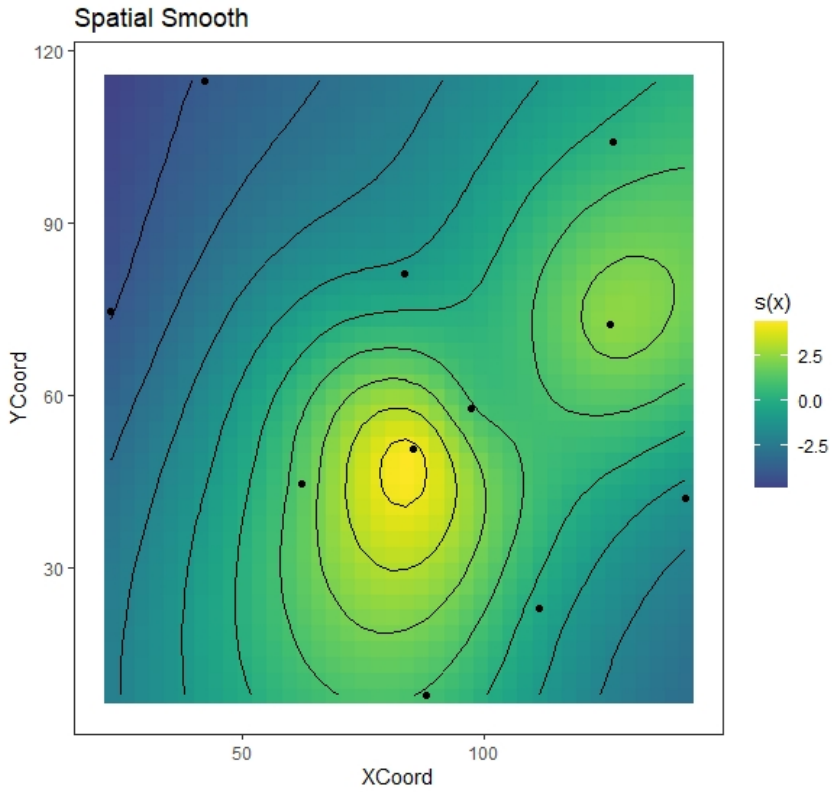
Figure 2.10: Residual diagnostics for modelling approach 4, applied to the observations of benzene concentrations from case study data 2.

2.4 Comparing GAMs to GWSDAT Models

As shown in Section 1.3.1, GWSDAT fits a trivariate (spatiotemporal) tensor product spline to the groundwater sampling data, with a penalty term based on the first derivative to control its smoothness. The value of the penalty term is optimised using a Bayesian approach also described in Section 1.3.1. To evaluate and illustrate the similarities between the CoPC concentration surfaces estimated using GWSDAT and the GAMs using separate spatial and temporal thin plate spline smooths, a qualitative comparison was made for the two case study data sets. The comparison was made using spatial estimates of concentrations at the final time point in the data sets, and using GAM approach 1. In this qualitative comparison, only the spatial trends were considered, not the actual concentration values themselves, therefore the presented spatial plots obtained from GAMs show concentration estimates in the log-transformed state. A quantitative comparison between the two approaches will be shown in Section 2.5 using synthetic groundwater contamination data (see Section 1.4.1).

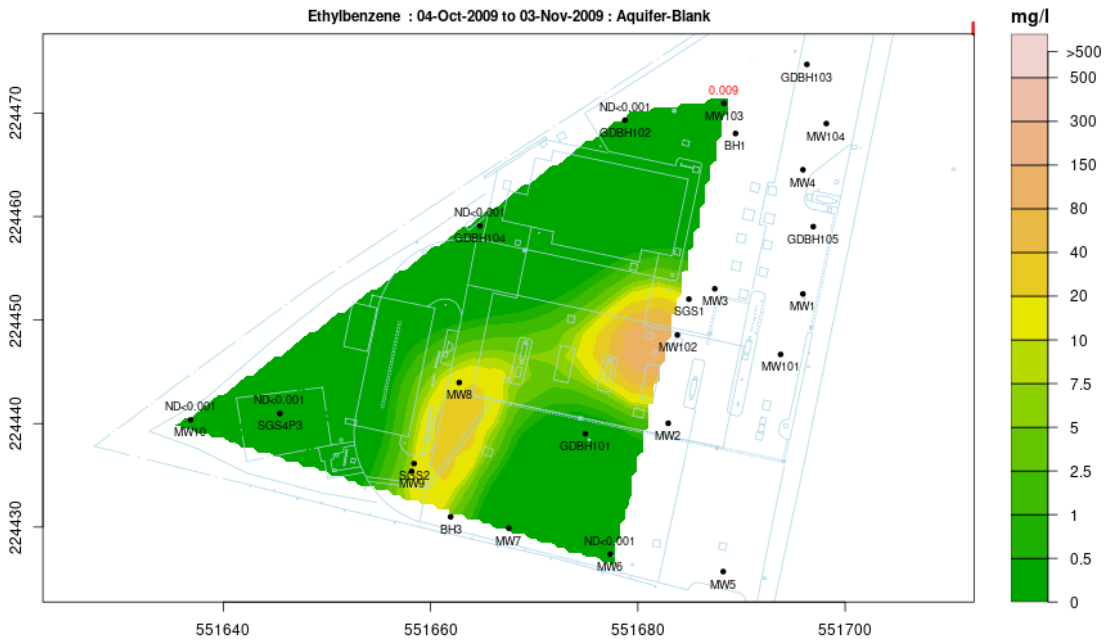


(a) Estimated benzene concentration surface via GWSDAT

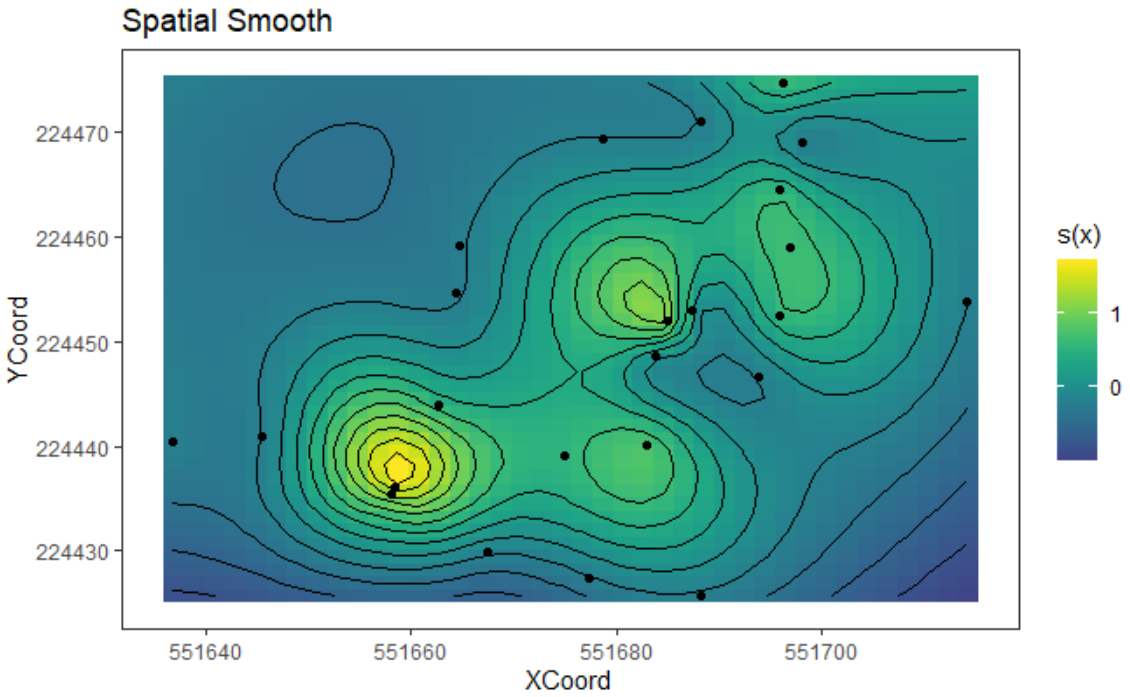


(b) Estimated benzene concentration surface via GAM approach 1 on the log-transformed scale

Figure 2.11: Comparison of spatial trends in the estimated benzene concentration surfaces at the final time point in data set 1 using GWSDAT and the separate thin plate splines smooth approach fitted in *mgcv*.



(a) Estimated ethylbenzene concentration surface via GWSDAT



(b) Estimated ethylbenzene concentration surface via GAM approach 1

Figure 2.12: Comparison of estimated ethylbenzene concentration surfaces at the final time point in data set 2 using GWSDAT and the separate thin plate splines smooth approach fitted in *mgcv*.

2.4.1 Case study data 1

Figure 2.11 shows that the GAM predicted a similar spatial trend as the modelling approach used in GWSDAT. Higher concentrations of benzene were expected in the groundwater around well MW-02 at the centre of the monitoring site and MW-07 at the north-eastern boundary. The GAM approach was able to estimate these spatial features effectively. It should be noted that GWSDAT only shows estimates of concentrations within the area enclosed by the monitoring wells (the convex hull), whereas the GAM approach shows the estimates over the entire grid cell.

2.4.2 Case study data 2

Figure 2.12 shows the comparison of the spatial trends of estimated concentrations at the final time point in case study data set 2. Once again, similar trends can be observed on the two images. Based on the GWSDAT output, peaks in ethylbenzene concentration levels were expected towards the southern boundary, between wells MW9 and MW8, and towards the centre of the monitoring network around well MW102. The GAM model estimated higher concentration levels at the former location, whereas GWSDAT at the latter. It should also be noted that GWSDAT did not produce estimates at the eastern edge of the network, where the GAMs estimated additional contamination hotspots, thus making adequate comparisons more difficult.

Overall, both the GAM and GWSDAT modelling approaches were able to detect the main spatial trends in the benzene and ethylbenzene concentrations in data sets 1 and 2. In the following sections, a simulation study using the synthetic groundwater contamination data set 1 (see Section 1.4.1) will be discussed, in which GAM approaches 1 and 4 were further compared using model diagnostics, information criteria and prediction errors.

2.5 Comparison Using Synthetic Groundwater Contamination Data

In the investigation described in this section, GAMs were fitted to the synthetic groundwater contamination monitoring data set 1 (see Section 1.4.1 in Chapter 1) using the *mgcv* package in R (S. Wood [2021]). The models were evaluated and compared using AIC (Akaike [1998]) and BIC (Schwarz [1978]) criteria, R^2 and root mean square prediction errors (RMSPEs) as well as qualitative analyses of estimated spatial trends. Additional residual diagnostics were also computed to assess model fit. The predictions of the selected models were also compared

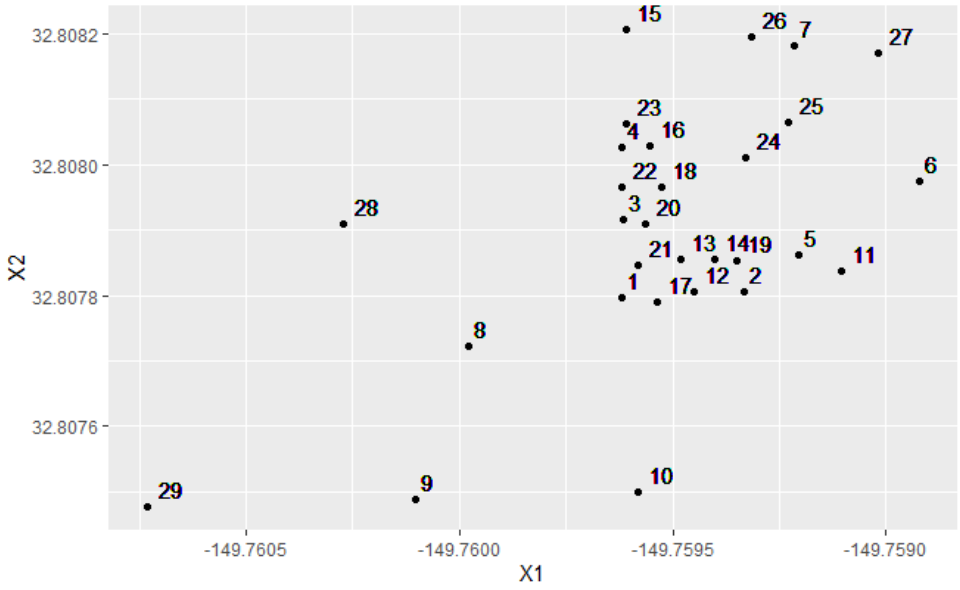
to predictions from models fitted to the same data set using GWSDAT (see Section 1.3). The GAM approaches used for the analysis were the separate thin plate splines smoothing approach (see Section 2.2.1) and the trivariate tensor product splines smoothing approach (see Section 2.2.4), since these were the most adequate models from Section 2.2, where they were fitted to the two case study data sets. One of the aims of the comparison between the two models using synthetic contamination data was to assess if the model performance metrics suggest similar results as in Section 2.2. Additionally, using synthetic data allowed for calculating prediction errors over the estimated concentration surface. Since estimating the concentration surface is the primary goal of modelling groundwater contamination data, the prediction error metric was important for comparing the investigated models. The simulated contamination data used for the analysis will be described in the following section.

2.5.1 Simulated Contamination Data

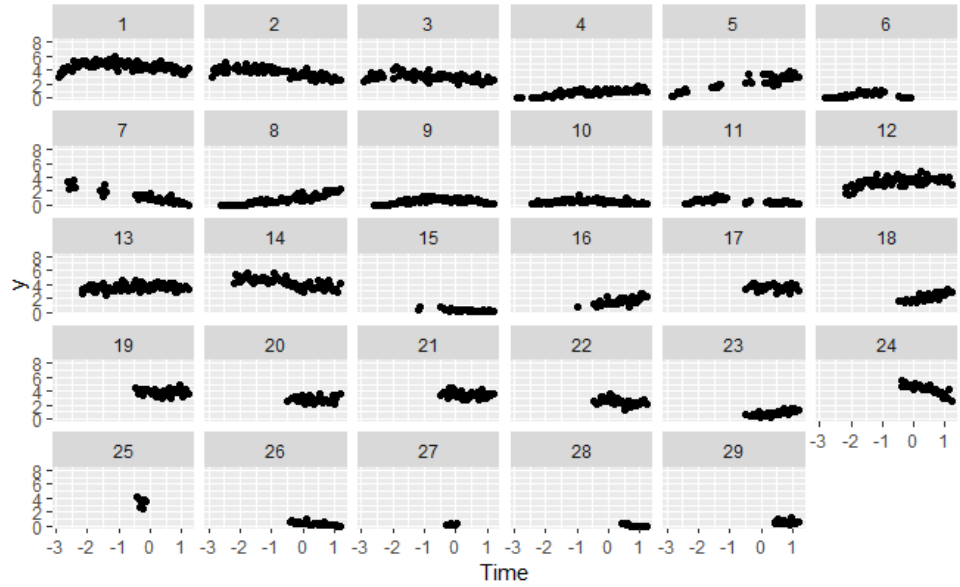
Section 1.4.1 provides a detailed description on how the simulated contamination data were generated and how they are structured. Using the CoPC concentration surface and the 29 monitoring wells located on the hypothetical site (see Figure 2.13a), two sampling scenarios were created. These sampling scenarios provided the observations for consequent modelling. Scenario 1 represented a more realistic sampling scenario in the sense that sampling times and locations were selected more sporadically, which created gaps in the collected data. This can also be seen on Figure 2.13b, which shows the collected observations in each monitoring well. In practice, monitoring wells are often sampled sporadically due to operation costs, as explained in Section 1.2.4. Scenario 2, on the other hand, contained continuous information for all wells over time, as if samples had been taken at each time slice considered in the data. The collected samples in each well over time are shown on Figure 2.13c. The data from sampling scenario 1 will be referred to as realistic data, while data from sampling scenario 2 will be referred to as complete data. As described in Section 1.4.1, multiplicative Gaussian noise was added to the sample data to mimic groundwater solute concentration observations, which commonly have associated measurement and analytical errors. The amount of error added resulted in a signal-to-noise ratio of 10:1 and the well-specific correlation coefficient between the errors was set to a very small value of 0.05. Similarly to the case study data sets in Section 2.2, a $log(y+1)$ transformation was applied to the synthetic observations to normalise the right skewness of the data (see Figure 2.14) and avoid the presence of zeros within the log transformation.

2.5.2 Exploratory Data Analysis

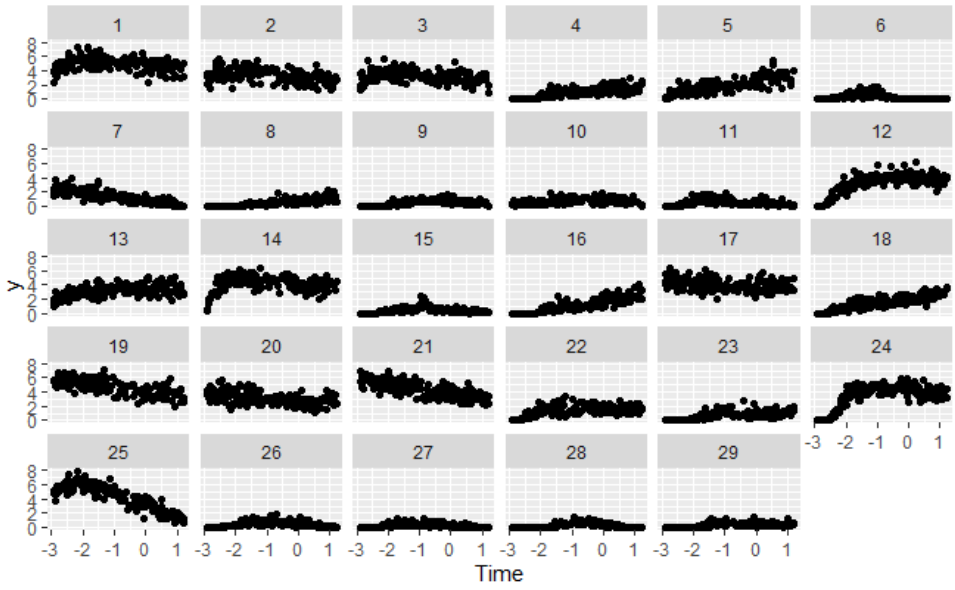
A brief exploratory data analysis was carried out prior to fitting the GAMs to the sample data sets. Firstly, the coordinates of the monitoring wells were plotted on a 2-dimensional grid on Figure 2.13a. The figure shows the arrangement of monitoring wells on the hypothetical contaminated site. The wells are arranged in a way that is roughly diagonal from the south-west to the north-east. The north-east corner contains most of the wells, with higher density towards the centre of the site. Some wells were placed sparsely in the south-west area of the site.



(a) Locations of monitoring wells 1-29 on the hypothetical site



(b) CoPC concentrations in the wells over time in the realistic sampling case



(c) CoPC concentrations in the wells over time in the complete sampling case

Figure 2.13: Exploratory plots showing the locations of monitoring wells and the observed concentration values in them over time in the realistic and complete sampling cases.

Secondly, the observations of CoPC concentrations over time were plotted for every monitoring well. Figure 2.13b shows concentrations under the scenario with sporadic data collection (realistic sampling design) and Figure 2.13c shows concentrations under the scenario with an idealistic data collection (complete sampling design). The y-axis represents the concentration of the CoPC in the groundwater sample and the x-axis represents time.

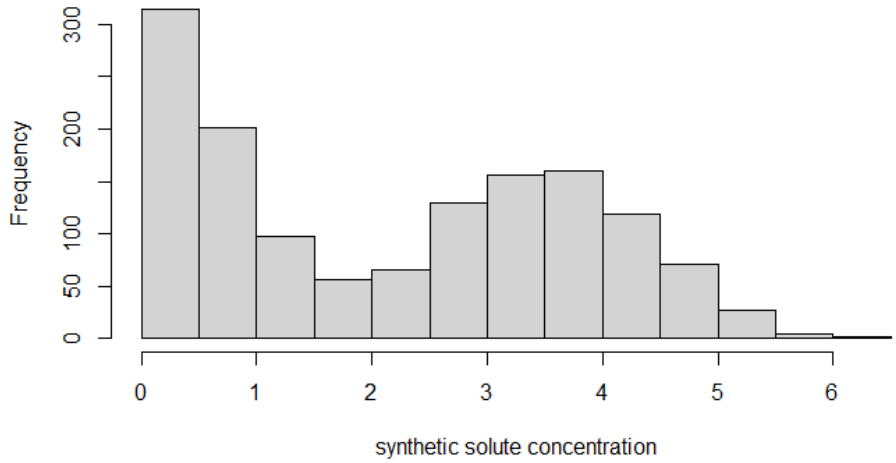


Figure 2.14: Distribution of synthetic observations of solute concentration values in the realistic sampling scenario of simulated contamination data set 1.

The similarities in the two figures are present since the two scenarios share a common data pool, but some of the samples have been removed from the complete scenario to create the realistic scenario to mimic commonly used sampling practices (see Section 1.2.4). The missing data points can be observed by comparing the two figures. In both scenarios, most monitoring wells seem to show a slightly decreasing trend in CoPC concentrations over time, following a sharp increase at the beginning of the monitoring period (such as wells 1, 2, 3, etc.), indicating the release and eventual attenuation of the contamination. On the other hand, a few of the wells show a slightly increasing trend (such as wells 4, 5, 12 etc.), indicating the slow migration of the CoPC plume towards these wells.

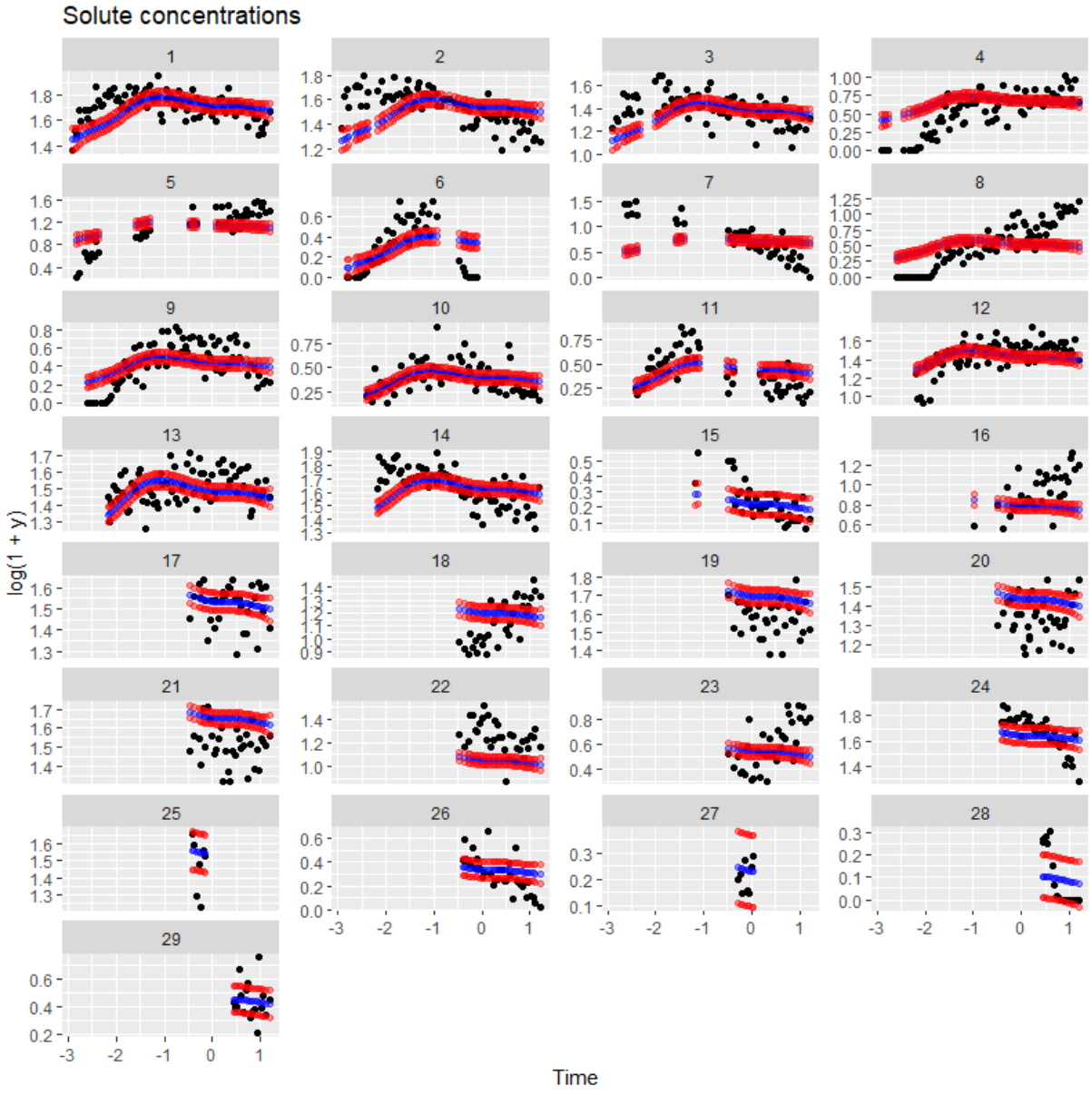


Figure 2.15: Observed synthetic solute concentrations over time (black), fitted values (blue) from the separate thin plate splines smooth term models and corresponding 95% confidence intervals (red) for each monitoring well in simulated contamination data 1 (see 1.4.1).

Figure 2.14 shows the distribution of CoPC concentration levels observed in the realistic sampling scenario. Similarly to the case study data sets (see Figure 2.3), the distribution exhibits skewness towards low concentration values. The distribution also appears to be bimodal. This indicates that most of the observations came from wells that were not substantially affected by the contamination. The wells that were more significantly affected, indicate a mean concentration of ~ 3.5.

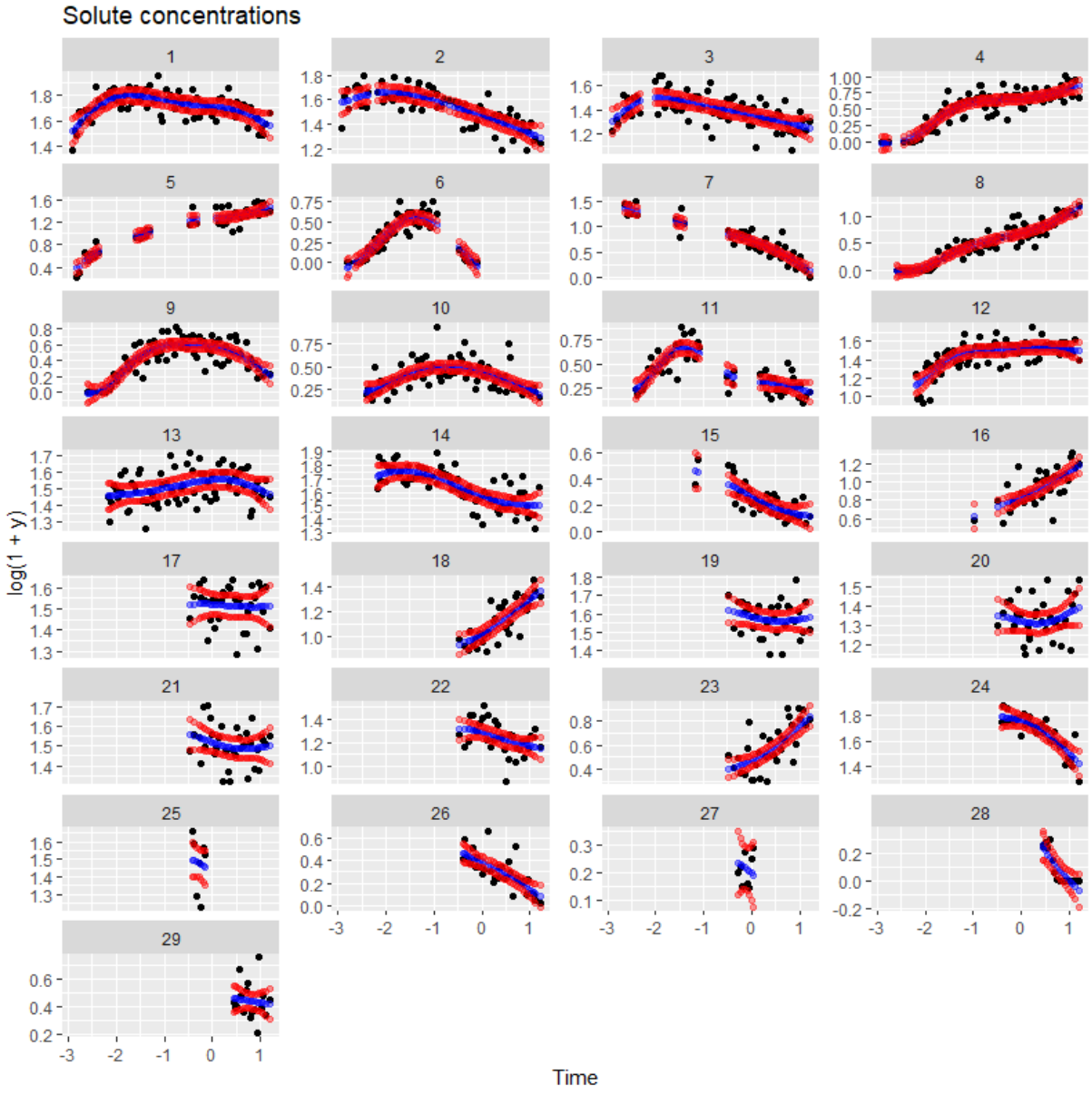
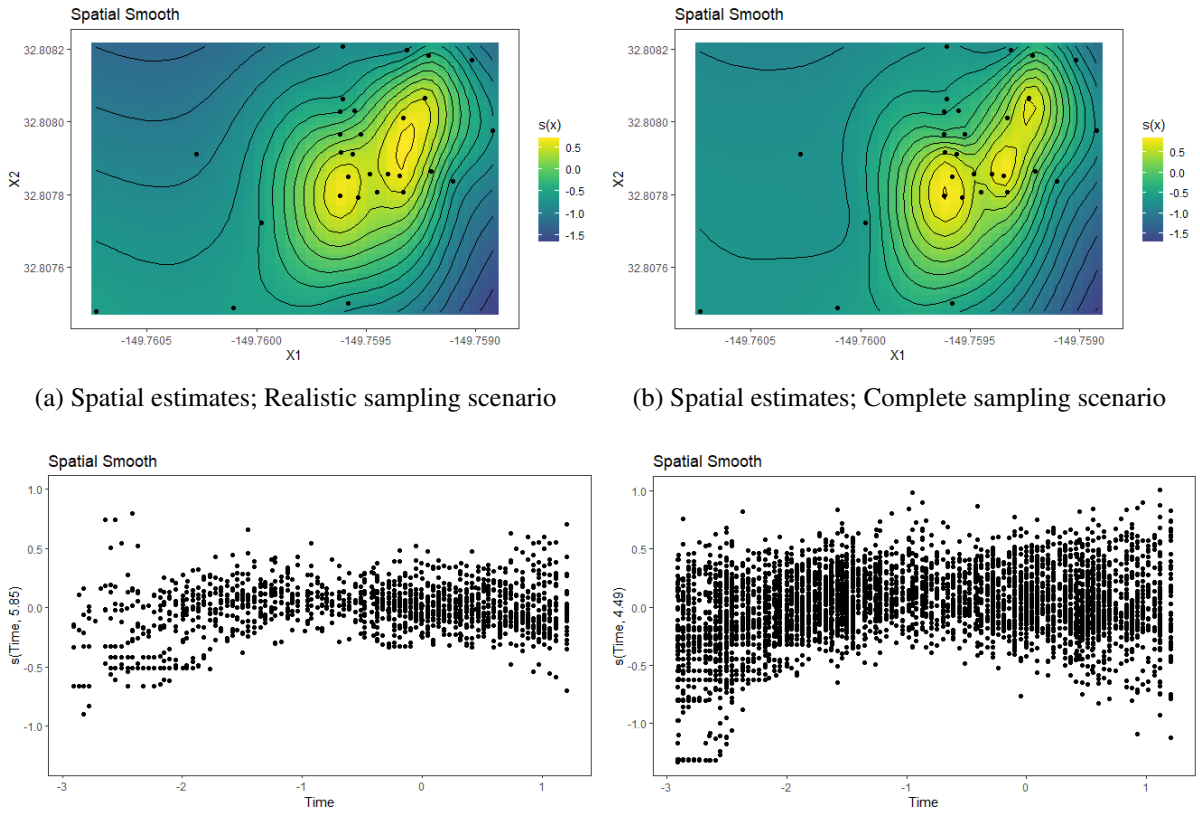


Figure 2.16: Observed synthetic solute concentrations over time (black), fitted values (blue) from the trivariate tensor product spline smooth term models and corresponding 95% confidence intervals (red) for each monitoring well in simulated contamination data 1 (see 1.4.1).

2.5.3 Model Fitting

The two investigated GAM approaches, i.e. the separate thin plate splines smoothers and the trivariate tensor product splines smoother models (see Sections 2.2.1 and 2.2.4) were fitted to each of the two simulated data sets i.e. the realistic sampling and complete sampling scenarios resulting in a total of four statistical models. These GAMs were used since they were the two most adequate GAM approaches from Sections 2.2 and 2.3. To summarise, the first approach defines two separate smooth terms for spatial (easting and northing) and temporal (observation time) covariates, while the second approach defines a single smooth term for all three. In both cases the degree of smoothness can be adjusted for each term by specifying the possible number of basis functions k.



(a) Spatial estimates; Realistic sampling scenario

(b) Spatial estimates; Complete sampling scenario

(c) Temporal estimates; Realistic sampling scenario

(d) Temporal estimates; Complete sampling scenario

Figure 2.17: Comparison of spatial and temporal estimates at the final time point using the separate thin plate spline smooth terms model in the investigated sampling scenarios (realistic and complete sampling). The results display the estimated log-transformed CoPC concentrations.

As mentioned in Section 2.2, S. Wood [2021] suggests that the choice of k is generally not critical as long as it is reasonably large enough to represent the data well, while not having a detrimental effect on computational efficiency. The author also suggests that if the degree of smoothness chosen by the model is at or close to the limits of the selected value, k can

be increased and the model re-fitted to check if the more flexible model results in a better fit. Taking this into consideration, the value of k was selected in this analysis to minimize AIC and BIC criteria to achieve reasonable complexity and avoid overfitting. The smoothing parameter was estimated using the generalized cross-validation (GCV) criterion (Golub et al. [1979]).

2.5.4 Model Assessment

This section will present model diagnostics and the estimated CoPC concentration surfaces to assess the model assumptions and how well the two investigated GAM approaches fit the simulated groundwater contamination data given different sampling scenarios. The estimated concentration surfaces will also be compared to the output from GWSDAT.

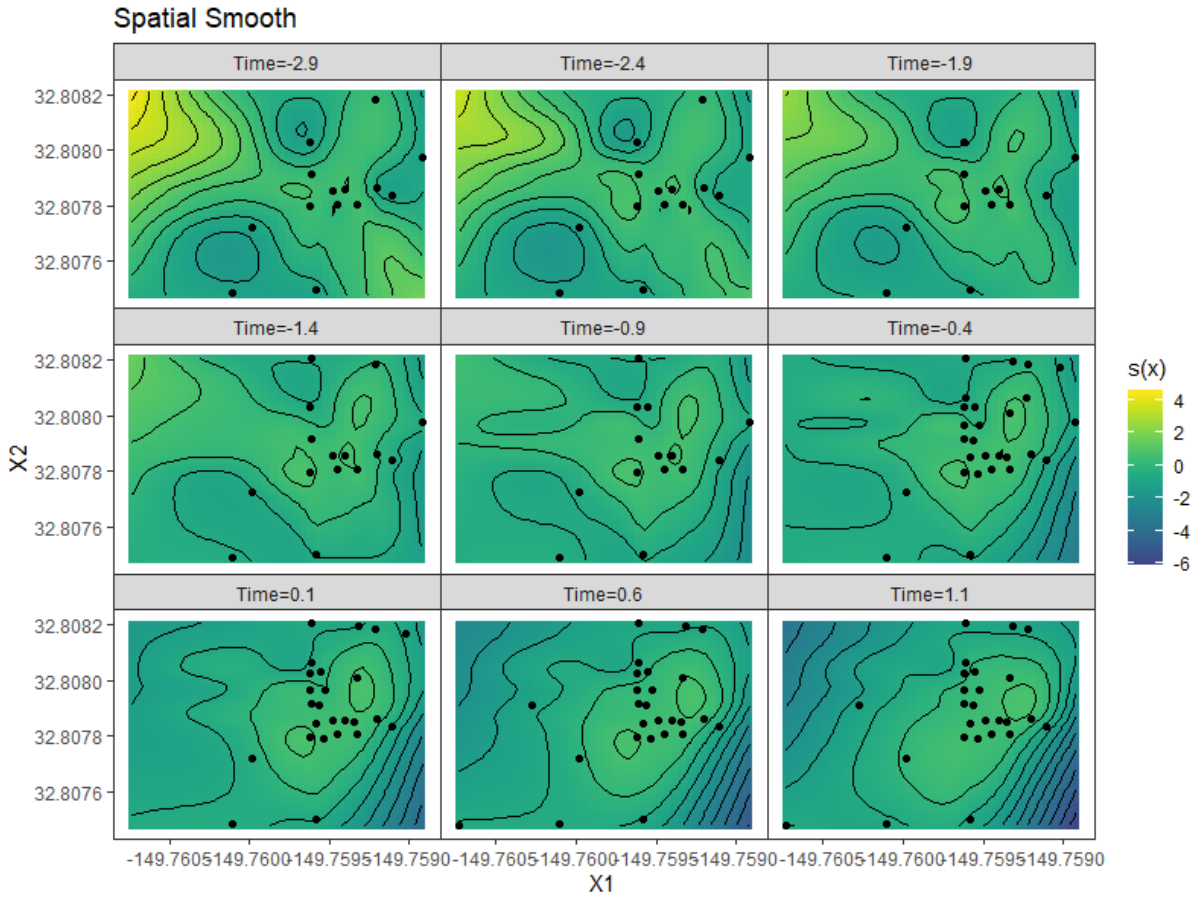


Figure 2.18: Estimated CoPC concentration surfaces using the trivariate tensor product spline smooth term model at different time points in the realistic sampling scenario. The solute concentrations are log-transformed.

The model summary diagnostics in *mgcv* showed that for the separate thin plate regression splines model, both smooth terms were significant with p-values $< 2e^{-16}$ when fitting the real-

istic data set. The chosen k value of 10 for the temporal smooth term was adequate, however, the k value of 20 for the temporal smooth term appeared to be too small based on a k-index value of 0.63 with a p-value of $< 2e^{-16}$ in the *gam.check* diagnostics. The true values of the log-transformed concentration data in this scenario ranged from $4.479e^{-11}$ to 1.956, while the fitted values ranged from 0.069 to 1.783. The corresponding estimated standard errors had a median of 0.024 with a range of 0.016 to 0.070. In the trivariate tensor product splines approach, the selected k values of 10, 10 and 6 for the two spatial and the one temporal dimension respectively, were shown to be adequate by the *gam.check* diagnostics with a k-index of 1.09. The fitted values ranged between -0.070 and 1.798 with corresponding estimated standard errors between 0.021 and 0.070 with a median of 0.025. Figures 2.15 and 2.16 show the true and fitted values for the two modelling approaches with 95% confidence intervals for each individual monitoring well in the simulated contamination data set in the realistic sampling scenario. These figures show that the trivariate smooth term approach is able to capture the observed data more precisely than the separate smooth terms approach (see for example wells 2, 4, 6, 8, etc.), despite the similar range and median standard error of the fitted data.

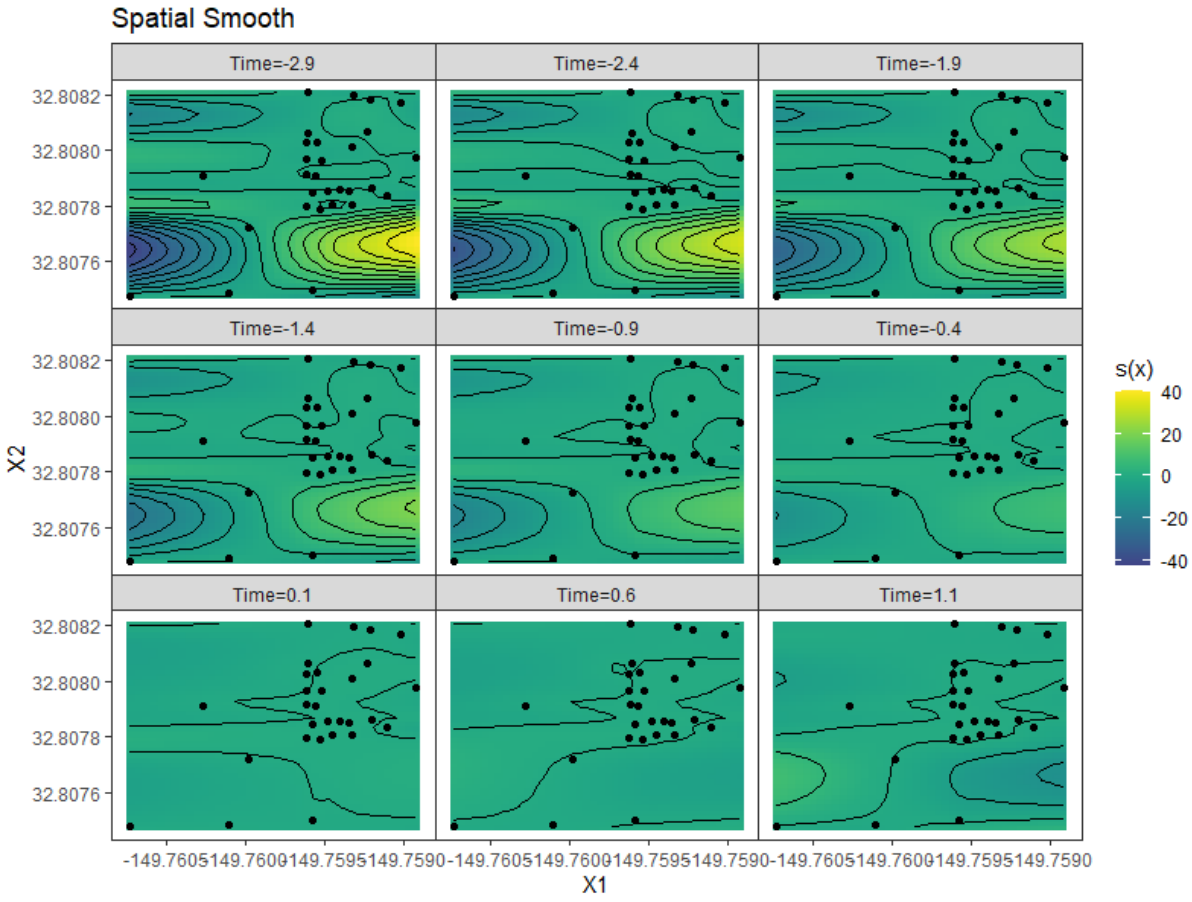
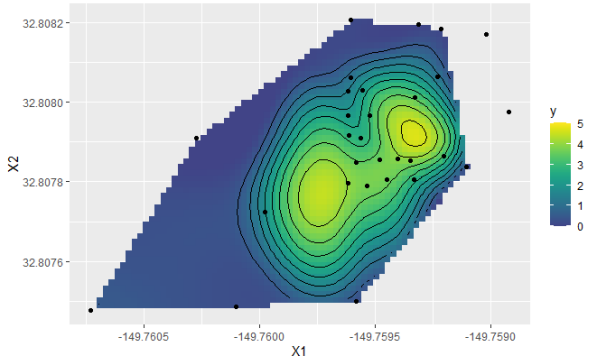


Figure 2.19: Estimated CoPC concentration surfaces using the trivariate tensor product spline smooth term model at different time points in the complete sampling scenario. The solute concentrations are log-transformed.

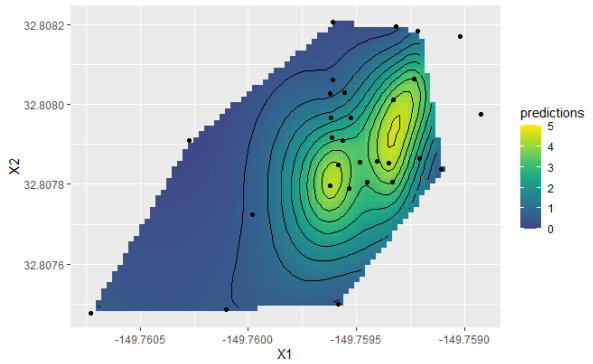
When modelling the idealistic monitoring data set, both smooth terms in the separate thin plate regression splines approach were significant with p-values of $< 2e^{-16}$. The chosen k value of 10 for the temporal smooth term was adequate, however, the k value of 29 for the spatial smooth term appeared to be too small based on a k-index value of 0.58 with a p-value of $< 2e^{-16}$ in the *gam.check* diagnostics. The fitted values had a range of -0.174 to 1.895 with corresponding estimated standard errors between 0.023 and 0.028 with a median of 0.023. Using the trivariate tensor product splines approach in this scenario, the *gam.check* diagnostics indicated that the selected values of k (10, 10 and 6 for easting, northing and time respectively) were adequate. In this case, log-transformed concentration data ranged from $6.972e^{-12}$ to 2.181. The fitted values ranged from -0.261 and 1.972 with corresponding estimated standard errors between 0.024 and 0.061 with a median of 0.028.

The fitted models were used to predict the CoPC concentration surface over the investigated hypothetical site. Figure 2.17 shows a comparison of the smooth effects plots from the separate smooth term models in the realistic and complete sampling scenarios. For these models, spatial and temporal smooth effects are plotted separately on Figures 2.17a, 2.17c, 2.17b and 2.17d. The spatial smooth effects plots display the log transformed estimated CoPC concentration surface at the final time point. Since the trivariate smoothing approach involved a single spatiotemporal smooth term, the spatial smooth effects were also plotted at several fixed time points on Figures 2.18 and 2.19.

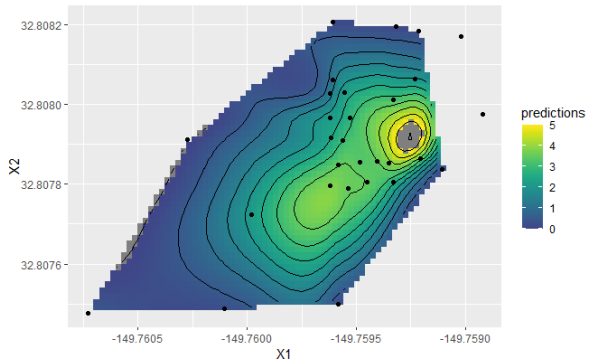
The temporal smooth effects of the separate smooth terms approach exhibit a similar concave down parabolic shape in both sampling scenarios even with the realistic data containing substantially fewer data points especially early on in the monitoring period. In general the trend indicates an initial increase and eventual decrease in CoPC concentrations overall on the hypothetical site. The trends of the spatial smooth effects are also generally similar between the realistic and the complete data scenarios, with the complete data exhibiting more spatial variation overall, especially in the north-east region of the site. The spatiotemporal smooth effect plots for the models using the trivariate smooth term approach show different trends based on which sampling scenario they were applied to. In the realistic sampling scenario, the trivariate smooth approach estimated similar spatial trends as the separate smooth terms approach, identifying the areas where CoPC concentrations were highest (see Figure 2.20a). In contrast, in the complete sampling scenario, the trivariate smooth approach produced substantial ballooning outside the convex hull at the south-western and south-eastern boundaries of the hypothetical site. Ballooning refers to the phenomenon where in sub-domains of the data set (spatial regions or time periods) that lack sufficient amounts of information on solute concentrations, the model estimates can become significantly inflated or deflated (McLean et al. [2019]). This effect can be mitigated by adjusting the number of smoothing basis functions and the type of smoothing parameter used to prevent overfitting and propagate a smoother model (Evers et al. [2015]).



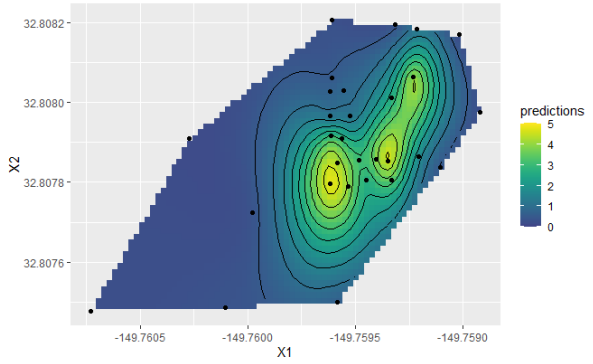
(a) True simulated CoPC concentrations



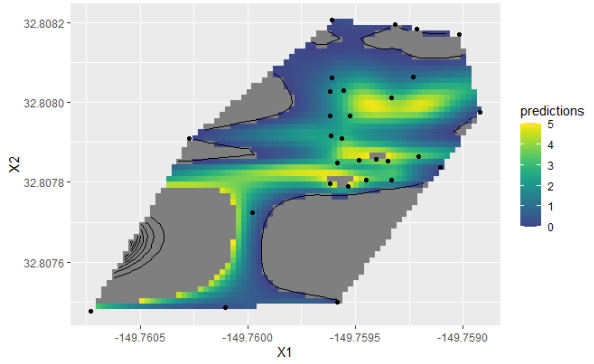
(b) Realistic data, separate smooth terms GAM estimates



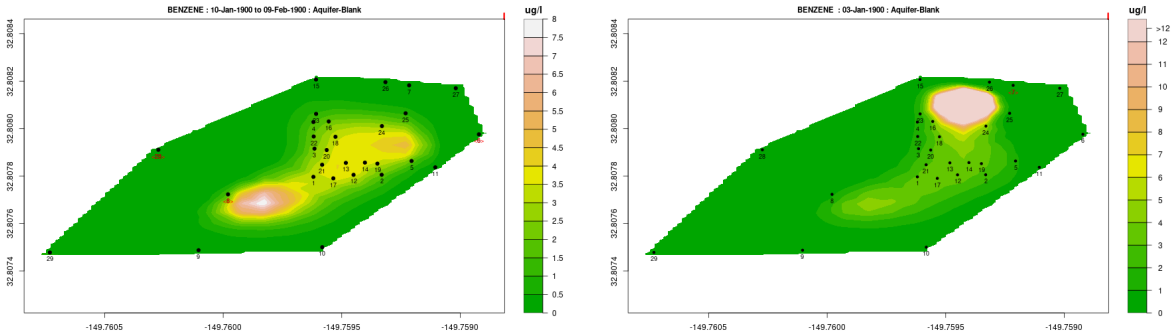
(c) Realistic data, trivariate smooth term GAM estimates



(d) Complete data, separate smooth terms GAM estimates



(e) Complete data, trivariate smooth term GAM estimates



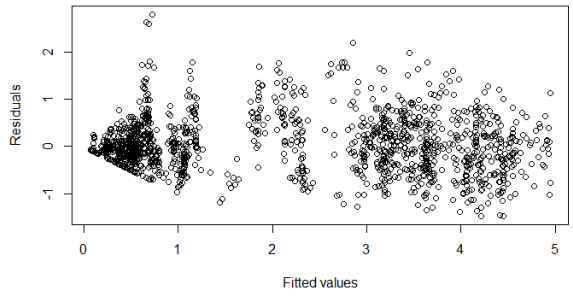
(f) Realistic data, GWSDAT estimates

(g) Complete data, GWSDAT estimates

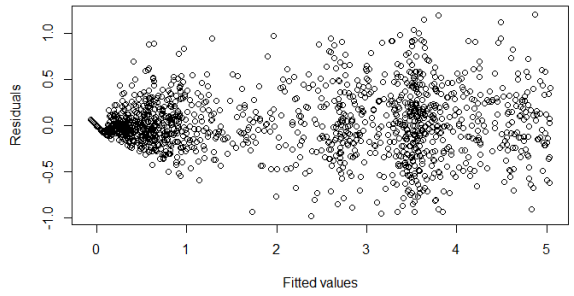
Figure 2.20: Comparison of estimated CoPC concentration surfaces at the final time point using separate spatial and temporal and trivariate smooth term GAMs in the realistic and complete sampling scenarios. The true CoPC concentration surface of the simulated contamination data set (see Section 1.4.1) is also displayed, as well as the estimates produced by GWSDAT. The results are displayed on the original scale, without the log-transformation. The grey areas represent regions where the estimated values exceed the bounds of the shown concentration range.

Figure 2.20 shows the estimated CoPC concentration surfaces over the convex hull at the final time point after model fitting, and the true simulated concentration surface on 2.20a. The log transformation of the estimated concentration values have been reversed to represent the results in the original scale on these figures. The model estimates support the observations of the smooth effect plots showed. Both the separate and trivariate smooth GAM approaches capture the spatial trends in the data well in the realistic sampling scenario (see Figures 2.20b and 2.20c). However, the trivariate smooth approach results in substantial ballooning when modelling the observations from the complete sampling scenario as shown by grey areas in Figure 2.20e), which suggests that the model is too flexible and results in overfitting. The test data shows two peaks in the CoPC concentrations, one in the centre of the site and one at the north-east boundary (see Figure 2.20a). The GAMs fitted to the realistic sampling data capture the north-east peak well but predict the one in the centre of the site to be localised around one of the monitoring wells. This is expected as there is a lack of information in the samples within the central region due to a lack of monitoring wells. Thus, the models estimated the plume to be around the monitoring well that produced the highest CoPC concentrations. The trivariate smooth model seems to capture the location of the central peak better but estimates much lower concentrations here than on the north-east. The complete data model using approach 1 produces similar results in terms of spatial trends as the realistic data models, including the location and extent of the contamination plume (see Figure 2.20d). This suggests that despite the additional information, no substantial improvement was achieved in terms of how well spatial trends were estimated. GWSDAT was able to estimate the location of the central contamination peak more precisely, but with higher CoPC concentrations than the true data under the realistic sampling scenario (see Figure 2.20f). However, using the complete sampling scenario data set caused the GWSDAT

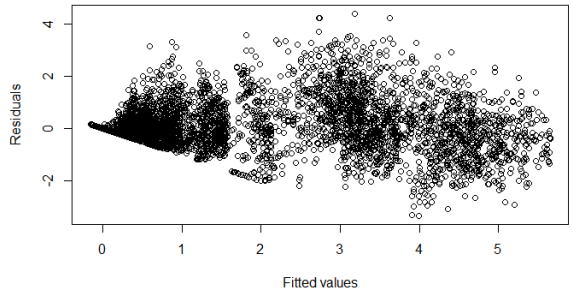
model to produce ballooning in concentration estimates on the northern boundary of the convex hull (see Figure 2.20g). Residual diagnostics were also considered in order to assess the model fits.



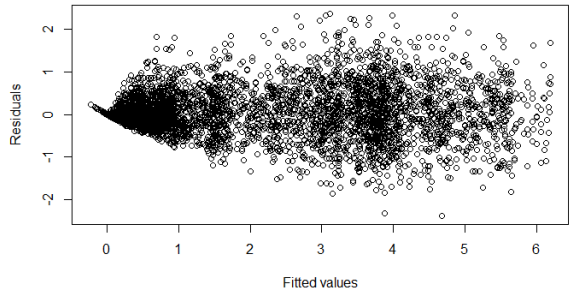
(a) Realistic data, separate smooth terms approach



(b) Realistic data, trivariate smooth term approach



(c) Complete data, separate smooth terms approach

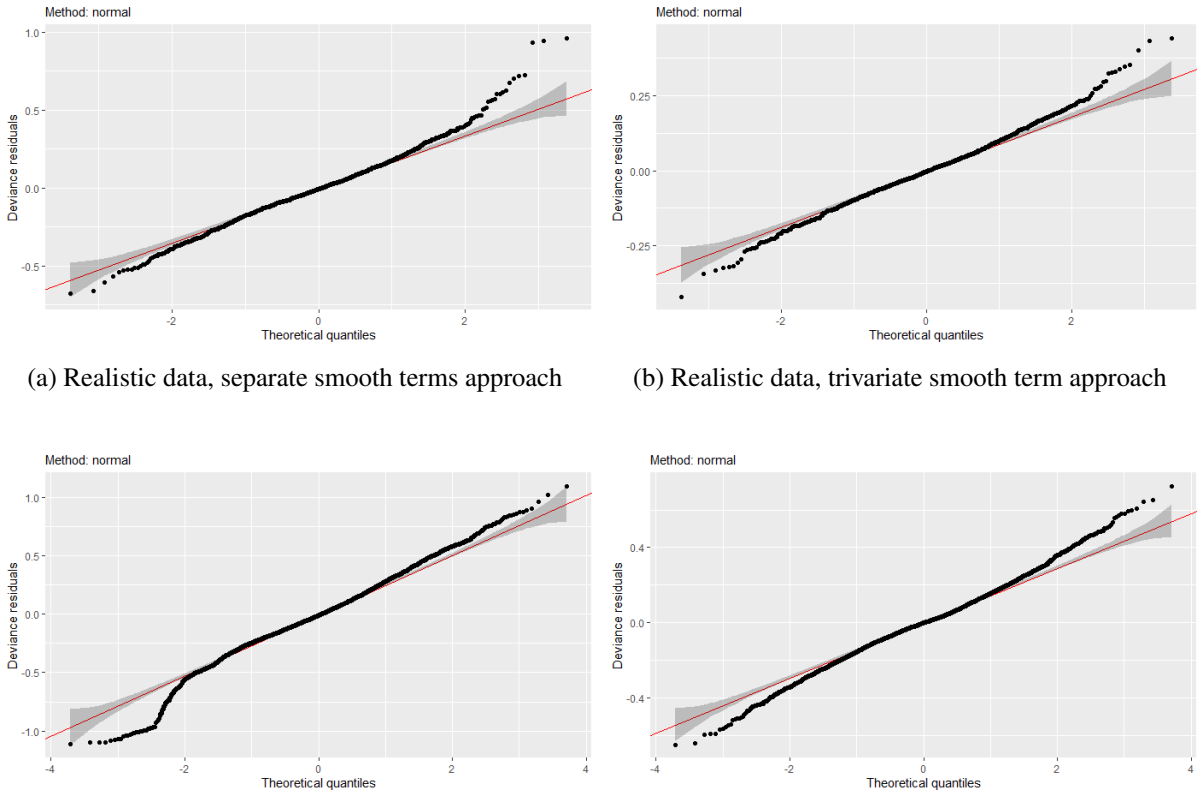


(d) Complete data, trivariate smooth term approach

Figure 2.21: Residuals versus fitted values for the investigated GAM approaches fitted using *mgcv* to the data obtained using realistic and complete sampling scenarios.

Figure 2.21 shows the residuals against the fitted concentration values for the two investigated modelling approaches in the realistic and complete monitoring scenarios. The residuals generally exhibit random deviations from zero, indicating that the models approximate the relationship between CoPC concentrations and covariates reasonably. Furthermore, there is no strong indication for the presence of outliers. They do however indicate the presence of some trend. At higher CoPC concentration values, the residuals appear to be spread evenly around zero suggesting that the variances of the errors are equal beyond a certain concentration value. The trend exhibited by the residuals of observations with low concentrations could be the result of the 10% multiplicative measurement noise that was added to the simulated data on the log scale (see Section 1.4.1). The added measurement noise structure results in smaller values having smaller associated errors than larger values. A notable difference between the separate and trivariate smooth modelling approaches is that the range of fitted values appears to be more smooth for the trivariate approach. The separate smooth terms approach appears to estimate that most so-

lute concentrations fall within certain smaller ranges, thus the transition between these ranges appears less smooth. This could be the result of the separate smoothing of spatial and temporal covariates in this approach. The bands could correspond to average concentration levels detected in the monitoring network at different sampling events in time.



(a) Realistic data, separate smooth terms approach

(b) Realistic data, trivariate smooth term approach

(c) Complete data, separate smooth terms approach

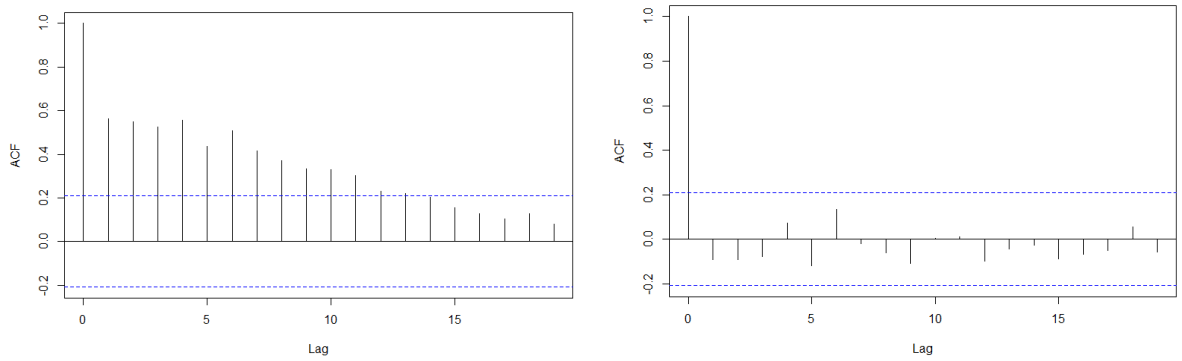
(d) Complete data, trivariate smooth term approach

Figure 2.22: Q-Q plots for the investigated modelling approaches in the two data scenarios.

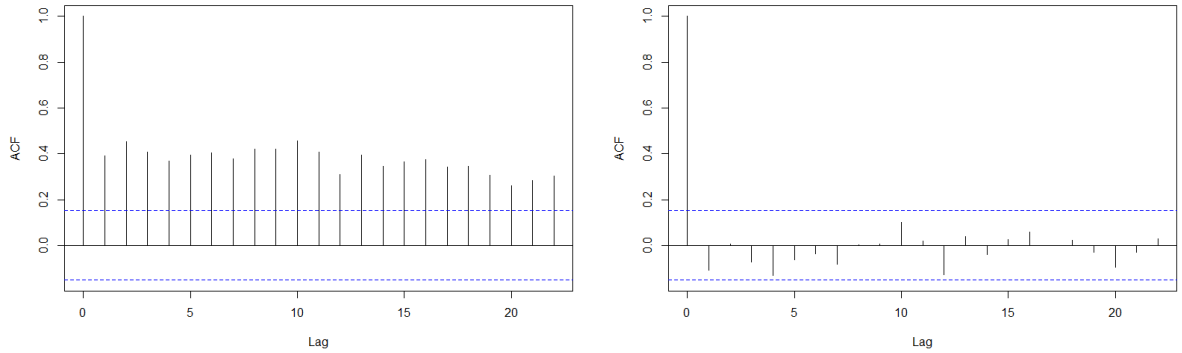
Figure 2.22 shows the residual Q-Q plots for the two investigated modelling approaches given the two sampling scenarios. The plots indicate that the residuals generally follow a normal distribution, with some deviation at either higher (realistic sampling scenario) or lower (complete sampling scenario) quantiles.

The autocorrelation of the residuals was also investigated for individual monitoring wells. Figure 2.23 shows the autocorrelation plots for the residuals obtained from the two investigated modelling approaches for well 1 in the realistic and complete sampling scenarios. Residuals from the trivariate smooth term models show no indication of correlation between the residuals. However, there is strong indication of residual correlation in the separate smooth terms modelling approach, which suggests an issue with the model specification, which could affect the estimated standard errors. It indicates that the temporal smooth term does not capture all of the temporal correlation. However, the focus of this investigation is the comparison of the two mod-

elling approaches and not the precision of inferences from these models. If the separate smooth terms model were to be used for statistical inference, the detected residual autocorrelation would have to be addressed.



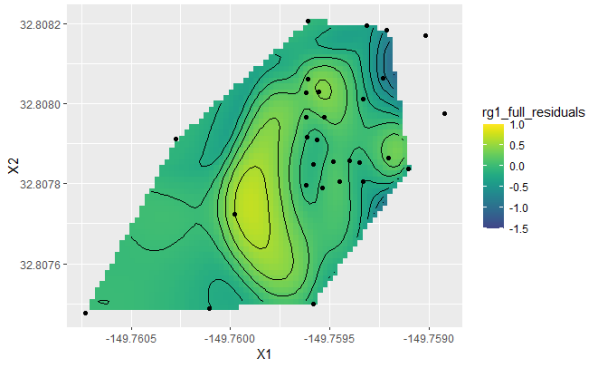
(a) Realistic sampling, separate smooth terms approach (b) Realistic sampling, trivariate smooth term approach



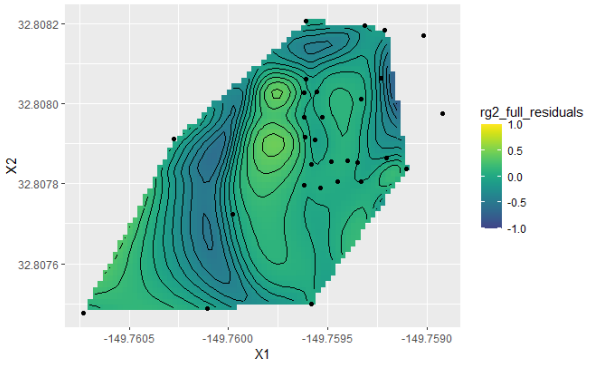
(c) Complete sampling, separate smooth terms approach (d) Complete sampling, trivariate smooth term approach

Figure 2.23: Autocorrelation plots of residuals obtained from the separate smooth terms and trivariate smooth term modelling approaches for monitoring well 1 in the realistic and complete sampling scenarios.

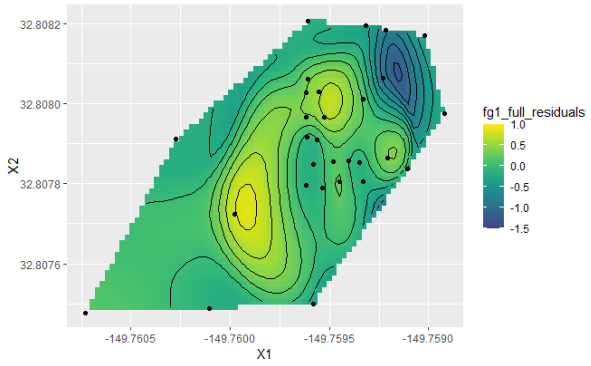
Spatial correlation of the residuals at individual time slices was also investigated. The residual variograms of both sampling scenarios did not exhibit trends for most time points. However, residual spatial correlation was observed at certain time points which suggested that the spatiotemporal smooth did not fully capture these trends in the data (this can also be seen on Figure 2.24). The modelling approaches presented in this thesis aim to capture a description of the spatiotemporal patterns of CoPC concentrations, but to provide rigorous inferences, additional steps should be taken to remove residual spatial correlation. This however, was not the goal of this thesis. Due to the sparsity of data in space (determined by the locations of the monitoring wells), the spatiotemporal smooth is limited in how well it can capture the trends in CoPC concentrations. This leads to the remaining spatial correlation of the residuals.



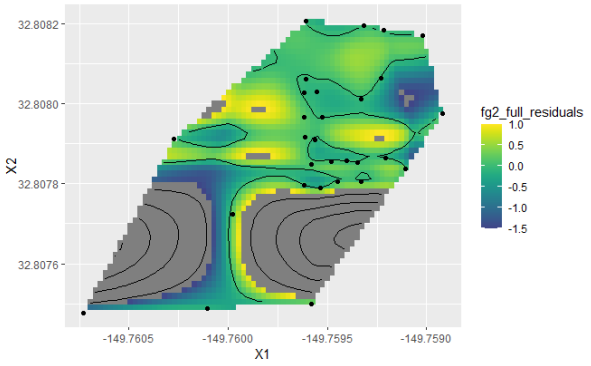
(a) Realistic data, separate smooth terms approach



(b) Realistic data, trivariate smooth term approach



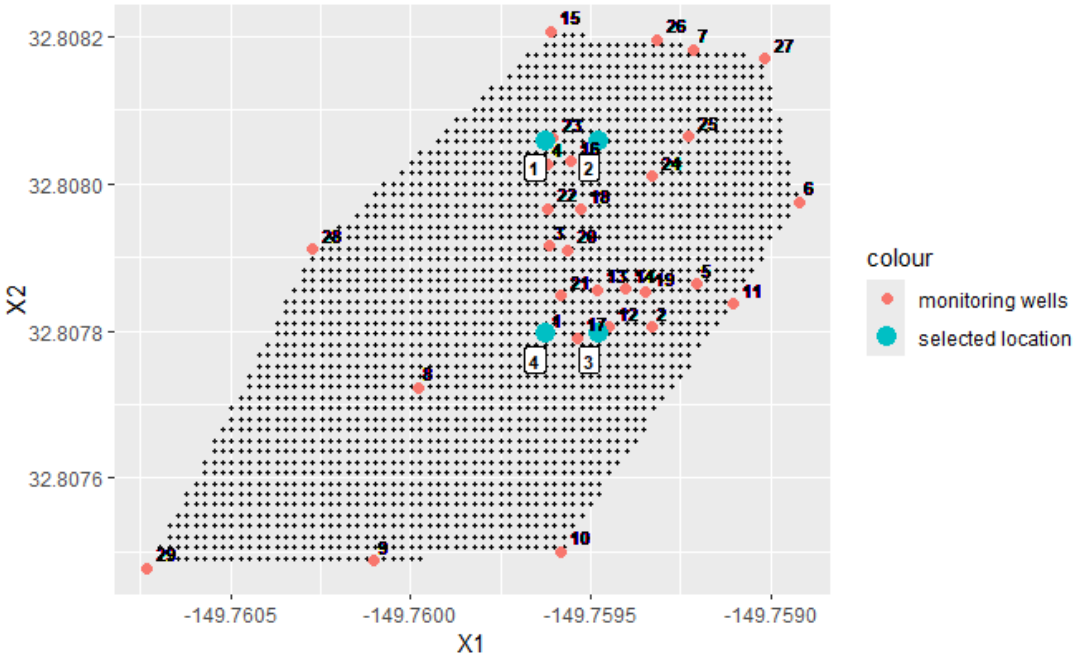
(c) Complete data, separate smooth terms approach



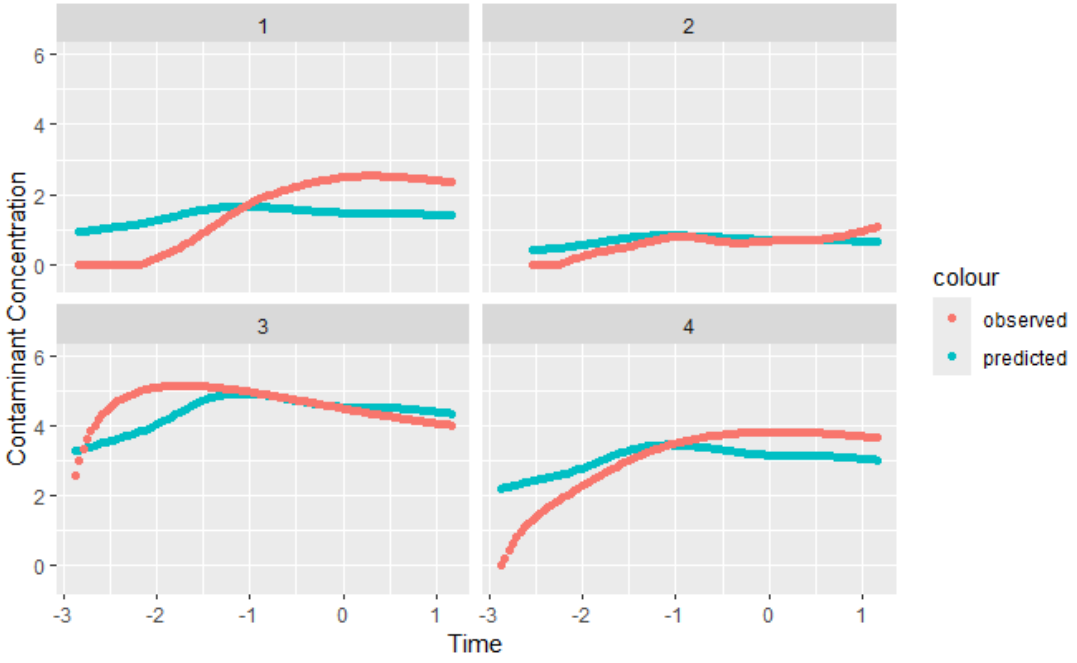
(d) Complete data, trivariate smooth term approach

Figure 2.24: Residual heatmaps indicating over- (blue) and underestimation (yellow) of log-transformed solute concentrations by the different modelling approaches. The grey areas represent regions where the residuals exceed the limits of the displayed range.

Figure 2.24 shows the residuals plotted on heatmaps. These plots help identify where the model over- or underestimates CoPC concentrations. Negative residual values indicate that the model is overestimating CoPC concentrations and positive values indicate the opposite. It is evident from the heatmaps that all models generally underestimate concentrations in the central region of the site. The models however, overestimate concentrations around in the north-east corner of the monitoring network. The grey areas on Figure 2.24d show the ballooning effect (both the over- and underestimation) in the complete data model using the trivariate smooth approach as it was also observed on the estimated concentration surface in Figure 2.20e. The results also support the observation that larger residuals generally appear in regions where no sampling wells are available to provide information on solute concentrations.

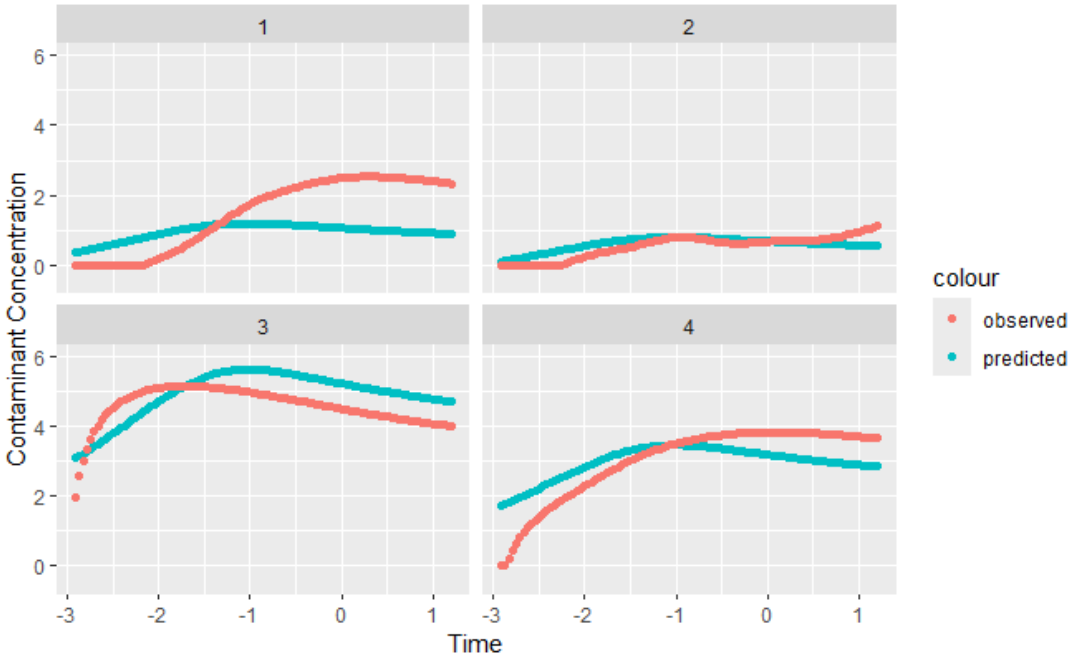


(a) Locations of the selected points for the time-series plots

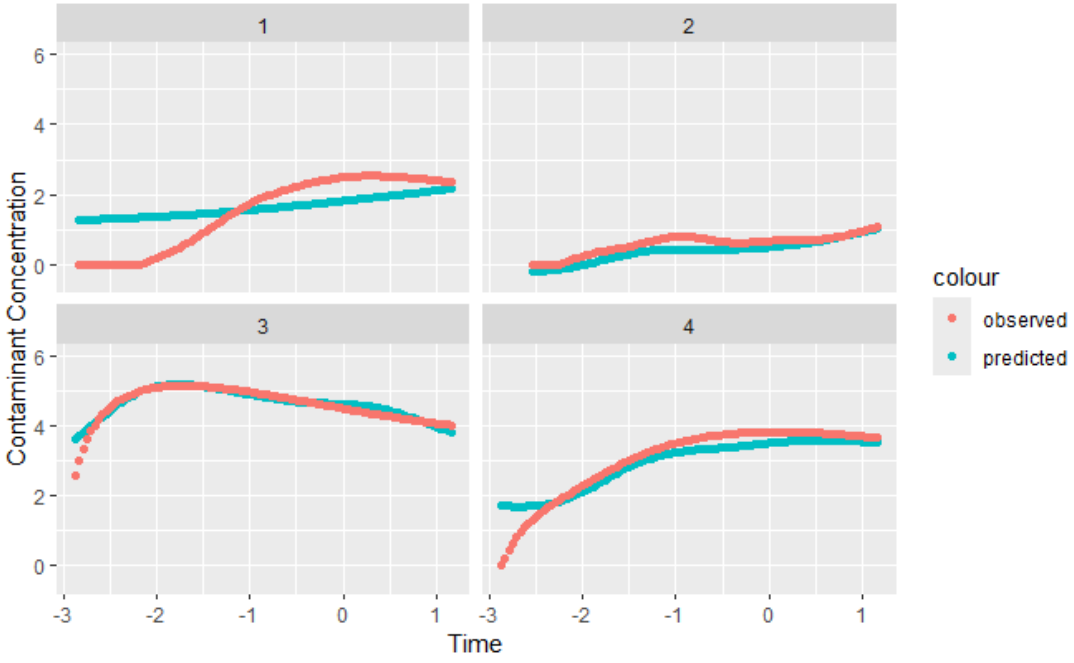


(b) Realistic data, separate smooth terms approach

To further assess the precision of the models, the time-series of CoPC concentrations have been plotted for four locations on the convex hull, selected to be around the central contamination peak. Figure 2.25a shows these selected locations on a map of all available spatial points in the data.



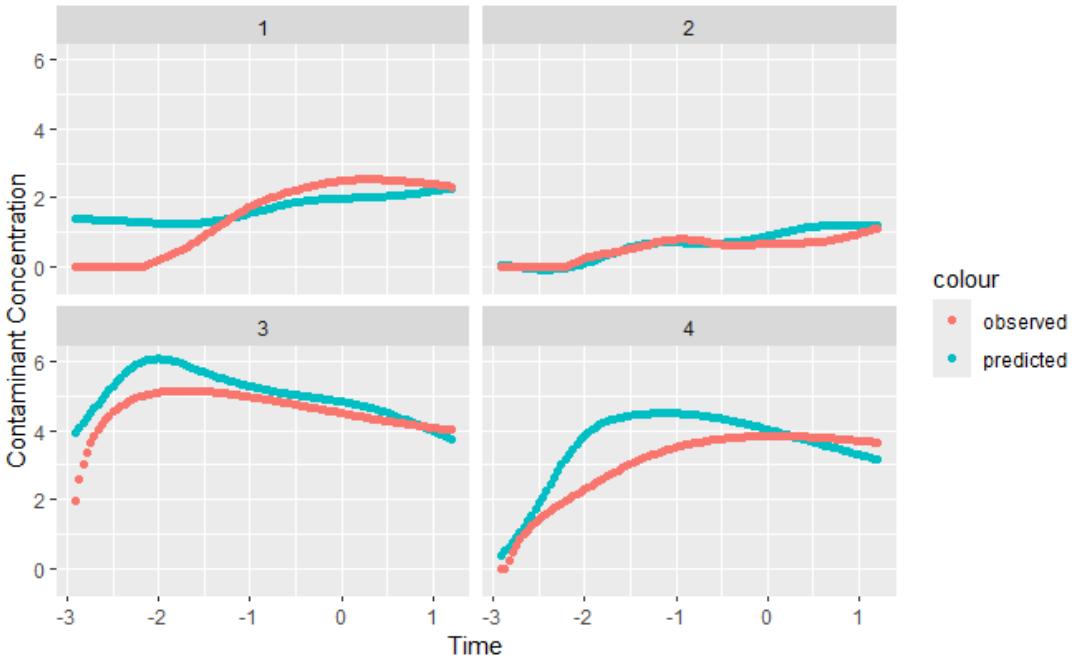
(c) Complete data, separate smooth terms approach



(d) Realistic data, trivariate smooth term approach

The models using approach 1 seem to behave similarly. Their estimates deviate from the actual data at these locations to roughly a similar extent (see Figures 2.25b and 2.25c). The realistic data model using the trivariate smooth term approach matches the actual data more closely at these points (2.25d). This approach also seems to capture the temporal changes adequately in the complete sampling scenario at these locations, suggesting that apart from the regions affected

by ballooning, this model also produces reasonable estimates (see Figure 2.25e).



(e) Complete data, trivariate smooth term approach

Figure 2.25: Time-series plots comparing observed and fitted solute concentrations at four locations and between the four modelling approaches

2.5.5 Model Evaluation

Table 2.3 shows the comparison of AIC, BIC and R^2 values between the investigated models and sampling scenarios. In general, AIC tends to select more complex models than BIC because it penalizes complexity to a lesser extent. In general, AIC is better at model selection when the aim is prediction, but BIC can more reliably select a correct model (Chakrabarti and Ghosh [2011b]). Therefore, looking at both of these values can provide a more balanced approach to model selection. The R^2 values indicate how much of the variation in the predictions is explained by the independent variables.

Table 2.3 shows that both AIC and BIC select the trivariate smooth term models over the separate smooth terms models. The latter approach performed better in the realistic sampling scenario than in the complete sampling scenario despite the latter containing more information. This indicates that more data points could result in overfitting and thus higher prediction errors. This could potentially be mitigated by adjusting model specifications such as the number of basis functions k to increase the smoothness of the model. The R^2 scores also indicate that the trivariate smooth term models explain more of the variation in predictions than the separate smooth terms models. However these high values could also be an indication that the model is over-

fitting, i.e. it matches the observations well, producing a less smooth prediction surface which could lead to ballooning elsewhere.

Table 2.3: Evaluation of the two modelling approaches in the realistic and complete sampling scenarios. RDA1 represents the separate smooth terms approach with the realistic sampling scenario, while RDA2 represents the trivariate smooth term appraoch. Similarly, CDA1 represents the separate smooth terms approach with the complete sampling scenario, while CDA2 represents the trivariate smooth term approach.

Model	AIC	BIC	R^2
RDA1	-521.09	-381.19	0.879
RDA2	-2136.31	-1601.22	0.964
CDA1	1613.43	1833.98	0.804
CDA2	-3356.30	-2289.74	0.931

2.5.6 Assessing Prediction Accuracy

The estimated concentration surfaces over the convex hull over all time points were analysed using root mean square prediction error (RMSPE). The RMSPE values represent the standard deviations of the prediction errors (residuals) of the models at every simulated point on the convex hull. RMSPE was calculated using the following equation:

$$\text{RMSPE} = \sqrt{\frac{\sum_{i=1}^{N}(\mu_i - \hat{\mu}_i)^2}{N}}, \tag{2.9}$$

where N is the number of data points, μ_i is the true value of the log-transformed CoPC concentration and $\hat{\mu}_i$ is the estimated value.

The models using the separate smooth terms approach appear to be the better models for estimating contamination concentrations since they result in lower RMSPE values (see Table 2.4). There is less deviation from the true data overall in these models compared to the others. The RMSPE value obtained for the trivariate smooth term approach with the complete sampling scenario is evidence for the ballooning effects that this particular model produced. In this case including more data points in the model led to ballooning, hence the RMSPE of this scenario is higher than the realistic sampling scenario. In contrast, adding more data points led to a lower RMSPE for the separate smooth terms model. However, based on the AIC and BIC criteria alone, the trivariate smooth model could be chosen as the best model. This discrepancy could be a result of overfitting. Based on the time-series plots (see Figure 2.16), the trivariate smooth model was good at matching the true values at the points of observation (monitoring wells),

however it deviates a lot outside of those points when predicting the surface and especially at the boundaries of the monitoring site. A similar phenomenon can be seen when comparing the two realistic data models. The trivariate smooth model produces better AIC and BIC values but its RMSPE is substantially higher than that of the separate smooth terms model. Overall the latter models seem to be more adequate at estimating CoPC concentration surfaces overall. However, the ballooning of the trivariate smooth models could be mitigated by adjusting the number of basis functions k.

Table 2.4: Root-mean-square prediction errors for the two GAM approaches fitted to the observations from the two sampling scenarios. RDA1 represents the separate smooth terms model with the realistic sampling scenario, while RDA2 represents the trivariate smooth term model. Similarly, CDA1 represents the separate smooth terms approach with the complete sampling scenario, while CDA2 represents the trivariate smooth approach.

Model	RMSPE
RDA1	25.41
RDA2	43.56
CDA1	24.86
CDA2	92.60

2.6 Conclusions

In conclusion, the trivariate tensor product spline model produces a better fit for the synthetic observation data, but suffers from ballooning issues when interpolating the CoPC concentration surfaces over the entire site. The separate smooth term approach in contrast appears to provide more accurate estimates of CoPC concentration surfaces due to the model resulting in a smoother fit. Since the data is mostly dominated by low concentrations, the smoother fit results in a lower RMSPE value. However, the autocorrelation plots highlighted issues with model specification in this case, which should be addressed to obtain a more adequate model. There is also a discrepancy between evaluating the models using information criteria, which support using the trivariate smooth model and prediction errors over the site, which support the separate smooths model. The attractive property of the trivariate smooth approach, including all predictor variables in the same smooth term, did not seem to reliably improve the prediction accuracy of the concentration surface in this case. In fact, it increased the spread of prediction errors substantially due to the ballooning problem. However, this could be mitigated by adjusting the number of basis functions k, to force the model to be more smooth. It should also be mentioned that using tensor product splines to fit the model increased computational costs compared to the separate thin plate regression splines approach when using high k values. More exhaustive investigation into using a trivariate tensor product splines smooth term approach to analyse

groundwater contamination monitoring data should be carried out to provide more insight into potential benefits over the separate smooth terms. A next step in this process could be to adjust the number of basis functions (k) and apply different smoothing parameter estimation and model selection methods to mitigate the negative impact of ballooning. The model specification issue in the separate smooth terms model should also be addressed to remove the correlation between residuals to allow for a fair comparison. The results also highlighted the potential issues with the presence of spatial regions that either do not include sampling locations, or the available sampling locations are not sampled. The presence of such regions could induce the ballooning of concentration estimates, depending on model specifications. Therefore, optimal sampling designs should aim to minimise the presence of information-sparse regions.

This chapter provided an introduction to groundwater contamination monitoring observations from a data analysis perspective. The estimation of CoPC concentration surfaces as demonstrated here will provide the main avenue by which sampling designs will be evaluated throughout the rest of this thesis. Through the application of the GAM framework, various approaches to spatiotemporal modelling have been investigated. The main differences between the investigated approaches were related to how the covariates were represented by smooth functions and how those smooth functions were constructed. As highlighted throughout the chapter, the covariates in groundwater contamination data are commonly spatial coordinates and time. A corresponding GAM model can consist of separate smooth terms for space and time or a joint smooth function for both space and time. The comparison of these two approaches aimed to contribute to a better understanding of the relationship between these covariates and the CoPC concentration estimates. The smooth functions can also be constructed using different spline-based methods such as thin plate regression or tensor product splines. By testing these methods, potential benefits in precision and computational efficiency could be identified.

The comparison between separate spatial and temporal smooth term and spatiotemporal interaction term models provided contrasting results. In terms of AIC, BIC and R^2 criteria, the spatiotemporal interaction models appeared to provide a better fit for the data. In contrast, the total RMSPE of the CoPC concentration surface suggested that the separate spatial and temporal covariates model could provide more precise estimates. It was also observed in the results that in models where the spatial and temporal covariates were represented by separate smooth functions, the choice of which smoothing spline to apply was not critical. Both tensor product and thin plate regression splines produced similar AIC, BIC and R^2 values, suggesting that using tensor product splines might be more efficient due to their generally lower computational cost (S. Wood [2017]). In GAMs where interaction between spatial and temporal covariates was considered via a single smooth term, using tensor product splines led to a better model fit in terms of the considered criteria. The P-splines models fitted using GWSDAT produced similar concentration surface estimates as the GAMs, particularly GAMs with separate spatial and

temporal smooths, thus supporting their application in groundwater contamination data analysis.

The results of the analyses also highlighted the importance of model specification. Using the same modelling approach, the estimated CoPC concentration surfaces could be substantially different based on the number of samples collected. Given the complete sampling data set, some models exhibited overfitting issues, suggesting that specifications should be adjusted based on the size of the available data.

With regards to implications for spatiotemporal sampling designs, the results suggested that ballooning of concentration estimates can occur in information-sparse regions. This implies that an optimal monitoring network should aim to spread sampling wells evenly over the target area to minimise the extent of information-sparse regions. This conclusion supports the findings by McLean et al. [2019] and McLean [2018] regarding ballooning and optimal sampling network designs. In the case of optimising the sampling strategy of existing monitoring networks, the implication of the results is that samples should be spread as evenly as possible in space and time to minimise the appearance information-sparse regions and periods. An evenly spread sampling design could thus improve the overall estimation of the CoPC concentration surface by spatiotemporal statistical models.

A more thorough, theoretical analysis of the spatiotemporal interaction GAMs smoothed via tensor product splines was not in the scope of this thesis, but could shed light on why this approach produced less accurate concentration surfaces. It could also elucidate if the model fit could be improved by adjusting model parameters such as the number of basis functions or smoothing parameters. Furthermore, the GAMs discussed in this chapter are exploratory in nature and aim to find trends and estimate the spatial or spatiotemporal surface of CoPC concentrations. The spatial or spatiotemporal dependence of the observations are considered through the smooth terms. A more detailed inferential treatment of this dependence was not in the scope of this thesis.

The following chapters will explore the challenges of optimising the sampling designs for groundwater monitoring networks, with the explicit aim to support the estimation of CoPC concentration surfaces using statistical modelling techniques such as the ones presented in this chapter.

Chapter 3

Well Influence Analysis

Chapter 2 investigated the spatiotemporal modelling of long-term groundwater contamination data and provided an introduction to its common characteristics. Through the analysis of the simulated data with two different sampling scenarios (realistic and idealistic), it highlighted how the sampling intensity and sampling design can impact subsequent data analysis and prediction of CoPC concentrations over space and time. This chapter aims to delve deeper into the topic of sampling design, with a focus on identifying which monitoring wells in a network contribute more valuable information towards the spatiotemporal estimation of CoPC concentration surfaces. Firstly, the motivation for the presented research will be discussed. Then, an innovative approach will be proposed for the analysis of monitoring well influence.

As discussed in Section 1.2.2, the long-term monitoring of groundwater quality is essential for protecting the health of humans and the environment as well as preserving potential sources of fresh water for the future. Certain aspects of groundwater quality, such as temperature and pH are often monitored remotely via sensors. However, the detection of constituents of potential concern (CoPC) commonly requires direct sampling and subsequent laboratory analyses (Schmidt et al. [2018]). Groundwater samples can be retrieved from natural sources, but more commonly from artificial wells, which provide access to the aquifers. Data of CoPC concentrations collected over time from a network of monitoring wells can subsequently be used to investigate the spatial characteristics and behaviour of CoPC plumes in space and time using statistical models (Meray et al. [2022]). The fieldwork required for the construction and operation of monitoring wells incurs health and safety risks, as well as significant costs. Thus, the spatial and temporal optimisation of groundwater monitoring networks is important for maximising the value of the collected data whilst minimising the number of samples taken. As highlighted in Section 1.2.3, finding optimal locations for new wells in monitoring network designs has been the main focus of much of the proposed optimisation methods (Farlin et al. [2019]).

There is however immense potential in improving monitoring practices for existing networks by optimising corresponding sampling designs.

Sampling design refers to the system that determines which monitoring wells to draw samples from and at what times. In general, sampling all available monitoring wells within a short period of time, or during the same sampling event is not feasible. This is because of logistical reasons such as the time it takes to retrieve a sample, the distance between monitoring wells and the ease with which these distances can be traversed with the sampling equipment. Therefore, the sampling design should select a number of wells to take samples from on each occasion. There are different aspects to consider when selecting sampling locations, including geographic location, accessibility, aquifer characteristics and geological conditions. From a statistical perspective, the selection should aim to maximise the contributions of the samples to achieving monitoring and data analysis objectives. For example, for characterising the spatial and temporal trends of a groundwater CoPC plume using statistical modelling, the sampling designs should aim to improve the precision of model estimates of CoPC concentrations, and not simply provide additional similar (and potentially redundant) information. Thus, a potential statistical approach for the optimisation of sampling location selection in existing monitoring networks is ranking the monitoring wells by their influence on the outcome of the applied statistical analyses, and hence on predicting CoPC concentrations over space and time. More influential wells could then be prioritised when designing future sampling strategies, and the least influential i.e. more redundant wells could be candidates for omission from the design. Monitoring wells can sometimes be abandoned for various reasons such as structural damage, inaccessibility and changes in land use or ownership. In such circumstances, the proposed well influence ranking approach would also provide a way for assessing the impact the omission of such wells would have on the spatiotemporal estimation of CoPC concentrations.

The concept of evaluating the importance of sampling locations and subsequently optimising groundwater monitoring networks also appears in the literature. Hosseini and Kerachian [2017b] combined different statistical criteria and the concept of value of information (VOI) to identify low and high priority monitoring sub-areas and aid the selection of optimal monitoring locations. Hosseini and Kerachian [2017a] used Kriging models and a Bayesian maximum entropy-based methodology to assign a number to each monitoring location indicating their removal priority level. Hosseini and Kerachian [2023] also used VOI to optimally redesign a coastal groundwater quality monitoring network. The idea of estimating well redundancy also appears in the paper by Ohmer et al. [2022], who investigated the use of a data-driven sparse-sensing method via a Python package called PySensors to optimize groundwater level monitoring networks with a focus on spatiotemporal dynamics and well redundancy. Meray et al. [2022] developed a Python package called PyLEnM (Python for Long-term Environmental Monitoring) with a focus on the application of machine learning to analyse spatiotemporal dynamics in groundwater monitoring

networks.

The need for a tool to evaluate well importance in monitoring networks was also echoed by the user base of GWSDAT (Jones et al. [2022]). Version 3.1[1] introduced a new functionality called *well redundancy analysis*, which allowed users to manually omit one or multiple wells from their data sets and repeat the statistical analyses. The primary intent of this feature was to understand which wells may have the most influence and provide supporting evidence that the conclusions of the statistical analyses would not be substantially different with the omission of certain wells. However, groundwater monitoring networks often consist of a large number of wells, which means that the number of combinations of candidate wells to omit is often prohibitively high to exhaustively investigate using such a manual approach. Therefore, it was identified that it would be highly practical to complement the well redundancy analysis feature in GWSDAT with functionality for automatically ranking wells within groundwater monitoring data sets. The ranking would identify the potentially most redundant wells to provide a guideline for where to start manually investigating the impact of well omissions. A method for ranking sampling locations by their influence on predicting CoPC concentrations would not only be applicable in the context of groundwater quality monitoring, but also in wider environmental monitoring surveys.

A potential approach for evaluating well influence is using an iterative process, where in each iteration a well is omitted from the data set and a statistical model is fitted to the remaining data to compute an evaluation metric such as RMSPE (defined in Section 3.4). The larger the RMSPE, the more influence the removed well had on the model estimates e.g. after fitting the P-splines model in GWSDAT as introduced in Chapter 1 Section 1.3.1. This is essentially an automation of the above described manual process and is analogous to performing leave-one-out cross-validation (LOOCV), where instead of individual data points, a collection of data points originating from the same well are left out of the P-splines model in each iteration. In this application, the observations of the removed well represent the test set and the remaining wells represent the training set for cross-validation. A more detailed description of this approach will be given later in this chapter (see Section 3.4). This groundwater monitoring well specific application of cross-validation was referred to as well-based cross-validation (WBCV) by Evers et al. [2015]. However, this approach can become increasingly expensive computationally with more complex statistical models and larger groundwater monitoring networks. This is because in order to produce an initial ranking, the number of times the statistical model has to be fitted equals the number of monitoring wells. Thus given a statistical modelling framework, the computation time of WBCV increases linearly with the number of monitoring wells in the data set. Additionally, if multiple wells are to be omitted from the network, the WBCV process has to be repeated after each new omission, to produce updated rankings for the remaining wells. This is necessary

[1] https://github.com/WayneGitShell/GWSDAT

because removing a well can change the influence of neighbouring wells. The computational inefficiency of the WBCV approach could present a barrier for the more widespread application of sampling location ranking techniques and hence the optimisation of groundwater sampling strategies. Groundwater solute modelling software such as GWSDAT, which already has a substantial user base of groundwater professionals, rely on providing results quickly. Therefore, a candidate approach for the well ranking capability should satisfy these efficiency constraints to allow for better integration. This chapter proposes one such approach, which is now also implemented in the *well redundancy analysis* feature of GWSDAT (Jones et al. [2022]), which allows for the omission of wells, to enhance its utility by ordering the list of wells as they appear for selection.

When using regression models to analyse groundwater monitoring data, a computationally efficient potential approach for ranking wells can be derived by using influential observation detection techniques. These techniques will be collectively referred to as influence analysis (IA) metrics. IA metrics are commonly used to quantify the influences of individual observations on regression estimates, often to identify outliers in the data. Various IA metrics exist in the literature. Merli [2005a] provide an exhaustive discussion on the topic of regression diagnostics for identifying influential data points. IA metrics include leverages (see Section 3.5.1), standardised residuals (Ranganai [2016]), Cook's distance (Cook [1977]), difference in fits (DFFITS) and DFBETAS (Merli [2005a]), covariance ratio (covratio) (Belsley et al. [2005]) and Hadi's measure (Hadi [1992]). In the context of ranking sampling locations, these metrics could be used to quantify the influence of each observation in each well, and then consider the well-specific average influence values as the ranking criteria. Essentially, IA metrics could be used to determine which monitoring wells contain more influential observations. This chapter proposes this method and illustrates that it is more computationally efficient, since the IA metrics can be computed from model estimates and parameters, eliminating the need for an iterative process like WBCV. Thus, if multiple wells are to be omitted from the monitoring network, the IA-metric based approach only requires re-fitting the model once for each well.

This chapter aims to evaluate the use of IA metrics to rank groundwater monitoring wells based on their influence on the estimated CoPC concentrations from spatiotemporal P-splines models. A simulation study was conducted using simulated groundwater contamination data sets 2,3 and 4 (see Section 1.4.2) to compare the accuracy of different IA metrics in approximating well influence rankings computed by WBCV. The WBCV rankings were based on the RMSPEs calculated for the test set (the removed well in each iteration of the process). The simulation study also investigated the effects of 4 design parameters on the IA rankings. These parameters were the spatial complexity of the CoPC plume, the number of monitoring wells, the spatial arrangement of the monitoring network and the assumption of the error type in the observations. The IA metrics investigated in this study were leverages, studentised residuals, Cook's distance,

DFFITS, Hadi's influence measure and covratio, which are defined in Section 3.5. The comparison was then also repeated on case study groundwater contamination data (see case study 2 in Section 1.4.4) to substantiate the results of the simulation study. The primary aim of the study was to identify the IA metric that provides the closest approximation to the WBCV rankings, and to explore how the ranking is affected by the investigated study design parameters.

3.1 Synthetic Groundwater Contamination Data

The simulated data sets of the three hypothetical CoPC plumes used in the simulation study are described in detail in Section 1.4.2. Here, a summary is provided about the generation and characteristics of the simulated data sets. They were generated using a groundwater flow model based on MODFLOW (Harbaugh et al. [2000]) and a solute transport model based on MT3D (Zheng [1990]). Each of the three plumes represented a different level of spatial complexity in CoPC distributions. They were referred to as simple, mid and complex plumes (see Figure 3.1). Each data set contained simulated concentration values of a hypothetical CoPC at 22,500 spatial points on a surface of a 1×1 km^2 grid, recorded over a total of 20 sampling events representing a time span of 10 years with 6-month intervals. This resulted in a total of 450,000 data points per data set. The CoPC concentration levels ranged from 0 to 100.

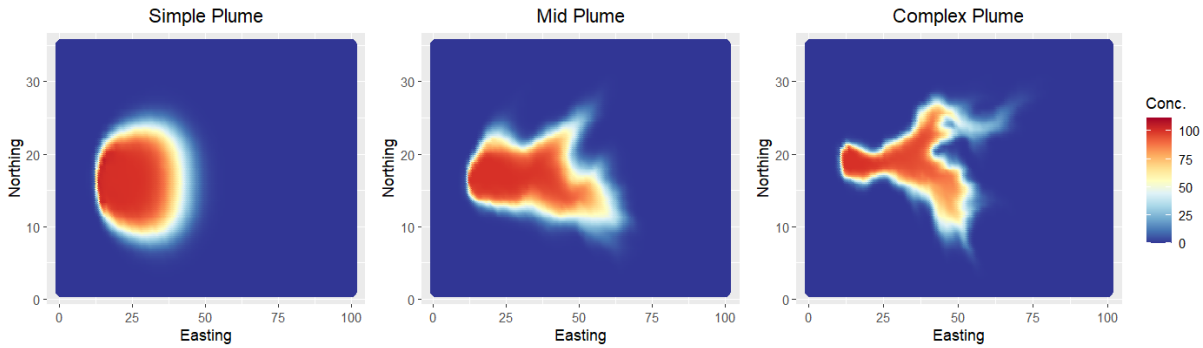


Figure 3.1: Heatmaps showing the CoPC concentrations of the three simulated groundwater CoPC plumes. From left to right: simple plume, medium plume and complex plume.

3.1.1 Hypothetical Monitoring Network Designs

As discussed in Section 1.2.2, groundwater monitoring networks refer to the collection of wells available on the area of interest. For this study, nine monitoring network designs were created using a combination of three types of well arrangements and three different numbers of wells (see Figure 3.2). The first well arrangement type was random, where the locations of wells were selected using simple random sampling on the spatial points of the data. These types of networks

resembled real groundwater monitoring networks, where the well locations are often affected by practical issues that prevent ideal placement, often resulting in seemingly random arrangements (see Section 1.2.3). For the second type, the well locations were selected as the junction points on a rectangular grid. This arrangement intended to minimise the impact of well arrangement on the results by spreading sampling locations evenly over the site. For the third type, the well locations were selected manually based on expert judgement of the hydrogeological conditions such as the groundwater flow direction and the location of the source of contamination, which determine the spatiotemporal characteristics of the plumes. This arrangement represented an idealistic scenario in terms of network design, where well placement was not affected by the aforementioned practical concerns. The number of wells was also varied between 6, 12 and 24 to represent scenarios with different levels of coverage over the site, resulting in observation data with different levels of spatial resolution. The hypothetical monitoring network designs were used to generate well-specific CoPC concentration data, $\mathbf{W}$ from the total simulated plume data sets, $\mathbf{P}$ so that $\mathbf{W} \subset \mathbf{P}$.

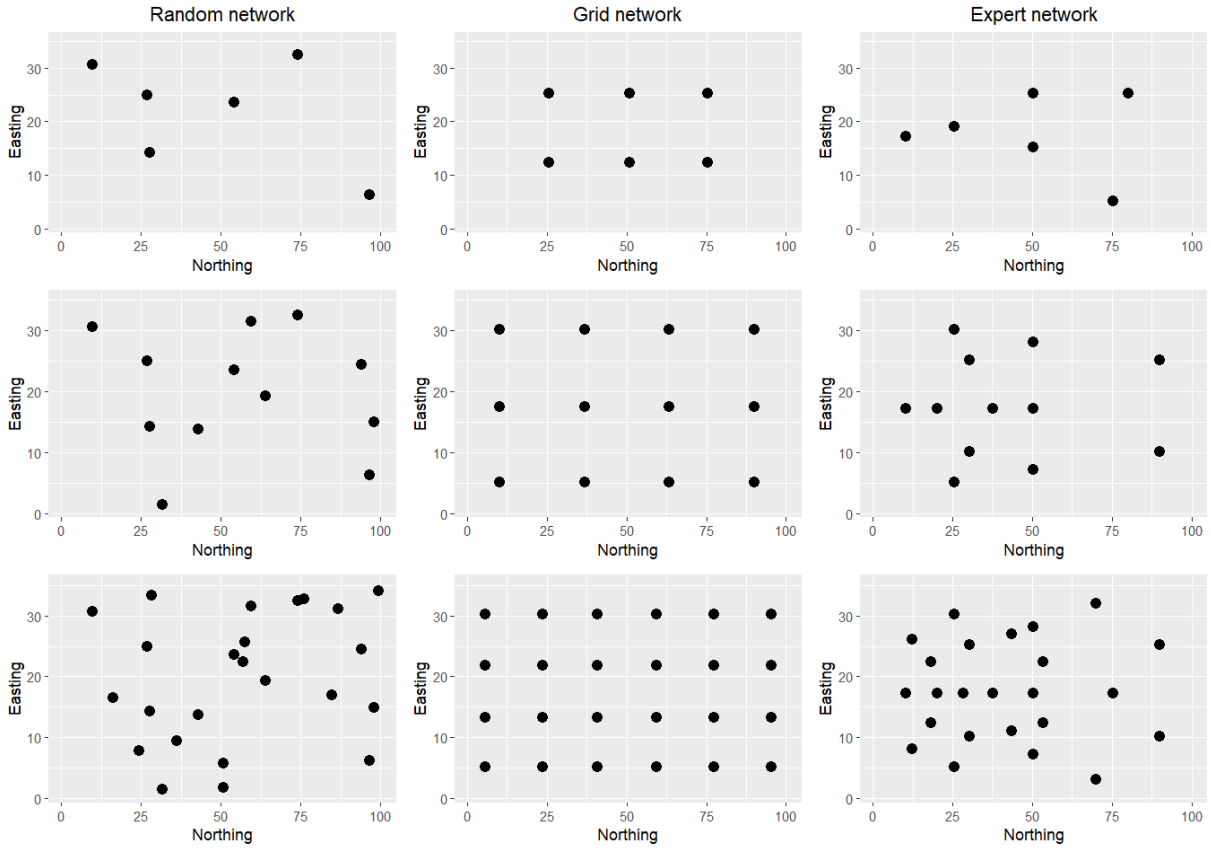


Figure 3.2: All 9 monitoring well network designs used in the simulation study. The black dots indicate the locations of the monitoring wells. The rows of the figure are arranged by the number of monitoring wells (6, 12 and 24), while the columns are arranged by the network design principles (random, grid and expert).

3.1.2 Generating Observations

To mimic groundwater CoPC concentration observation data, random noise representing measurement and analytical errors was applied to the well-specific data. The simulation study was performed using both additive and multiplicative measurement error structures to assess the differences on the performance of influence analysis metrics. The amount of noise applied was 15% in both cases, which was based on sampling and analytical variations in a comparison of blind duplicate samples from a large unpublished groundwater quality data set as discussed by McLean [2018]. In this comparison, the 15% measurement error in solute concentration measurements was observed on the log-scale.

Let y_i be the i-th observation with added measurement noise and z_i be the corresponding well-specific concentration data point without measurement noise, ranging from $0-100$, as described in Section 3.1. The additive measurement noise was added by:

$$y_i = z_i + \varepsilon_i, \tag{3.1}$$

where ε_i is a random variable drawn from $N(\mu, \sigma^2)$, a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.15$ to correspond to 15% measurement error or a signal to noise ratio of $\frac{1}{0.15} = 6.67$. This assumes that the same sampling and analytical techniques and equipment has been used to determine all solute concentrations.

Multiplicative noise was added by:

$$y_i = z_i \times \varepsilon_i, \tag{3.2}$$

where ε_i is a random variable drawn from $N(\mu, \sigma^2)$, a normal distribution with mean $\mu = 1$ to make the mean of the data with noise equal to the mean of the original data, and standard deviation $\sigma = 0.15$ to represent 15% measurement error.

As demonstrated in Chapter 2, the distribution of observed groundwater CoPC concentration data is often right-skewed, since a high portion of the data are low concentration measurements and a relatively low portion of them are high concentration ones (see Figures 2.3, 2.14). To normalise the observation data and avoid the presence of concentration values of 0 within the log term, the observation data were transformed by $y_i' = \log_e(1 + y_i)$, where y_i' represents the transformed observations. The transformation also allows for an additive interpretation of the multiplicative measurement errors on the log-scale.

3.2 Case Study Data Set

The comparison of IA metrics and WBCV was also performed on a case study groundwater contamination monitoring data set, case study data set 2, which is described in detail in Section 1.4.4. The anonymised data comes from the long-term monitoring of a decommissioned petrol station.

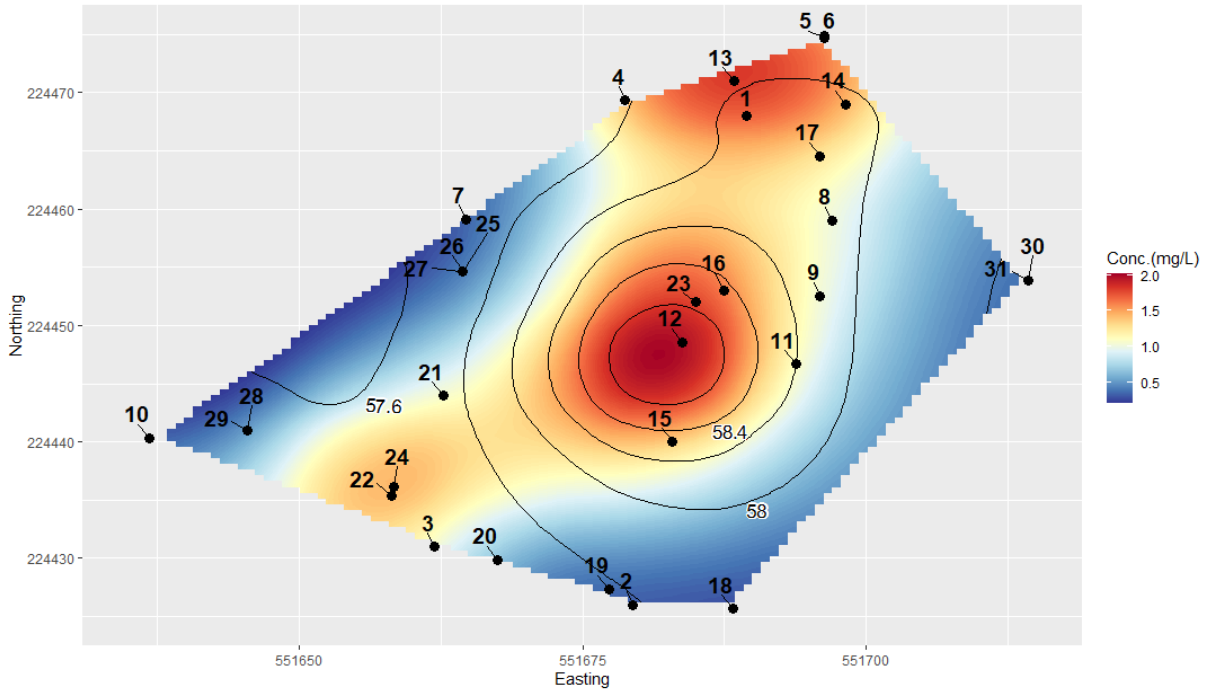


Figure 3.3: Estimated ethylbenzene concentration surface of the convex hull (area enclosed by the wells) at the final time-point in case study data set 2. The concentration surface was estimated using the P-splines modelling framework described in Section 1.3.1).

It contains the concentration measurements of five different solutes in groundwater samples from 32 existing monitoring wells collected over a 4-year period. The five solutes monitored were ethylbenzene, toluene, total petroleum hydrocarbons (TPH), nitrate and sulphate. Figure 3.3 shows estimated ethylbenzene concentrations and groundwater levels around the decommissioned petrol station as well as the layout of the monitoring wells.

3.3 Modelling Approach

The statistical modelling framework used to model the synthetic and case study groundwater CoPC observation data and subsequently calculate RMSPEs for WBCV and the IA metrics, was the spatiotemporal P-splines approach described in detail in Section 1.3.1. The model used in

this study expressed in vector matrix form was:

$$\mathbf{y}' = \mathbf{B}\alpha + \varepsilon, \tag{3.3}$$

where $\mathbf{y}'$ is the vector of responses (i.e. the log-transformed synthetic CoPC concentrations with measurement noise), $\mathbf{B}$ is the matrix of basis functions corresponding to the covariates (coordinates and time of observation), α is the vector of basis coefficients and ε is the vector of random variations. The basis coefficients were estimated using a penalised least squares approach (see Section 1.3.1), where ultimately the basis coefficients were estimated by

$$\hat{\alpha} = (\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D}_d^T\mathbf{D}_d)^{-1}\mathbf{B}^T\mathbf{y}', \tag{3.4}$$

where the matrix $\mathbf{D}_d$ computes the successive d-th order differences across the sequence of α-s in each of the 3 covariate dimensions and λ is a non-negative smoothing parameter that controls the degree of smoothness in the estimate (i.e. smoothing parameter). The value of the smoothing parameter was optimised using the Bayesian approach proposed by Evers et al. [2015], described in detail in Section 1.3.1.

The fitted values were then given by:

$$\hat{\mathbf{y}}' = \mathbf{B}\hat{\alpha}, \tag{3.5}$$

which by substituting $\hat{\alpha}$ with expression 3.4 gives:

$$\hat{\mathbf{y}}' = \mathbf{B}(\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D}_d^T\mathbf{D}_d)^{-1}\mathbf{B}^T\mathbf{y}' \tag{3.6}$$

where the matrix

$$\mathbf{H} = \mathbf{B}(\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D}_d^T\mathbf{D}_d)^{-1}\mathbf{B}^T, \tag{3.7}$$

is called the projection or hat matrix, in which each row represents a data point with a corresponding leverage value located on its diagonal. Leverages will be described in more detail in Section 3.5.1.

As highlighted in Section 1.3.1, selecting the number of B-spline basis functions in each covari-

ate dimension (*nseg* in GWSDAT) provides an additional control on model smoothness besides the the smoothing parameter (λ) which is estimated automatically in the GWSDAT modelling approach. In order to assess the impact of a varying number of basis functions, the simulation study was performed using two different sets of values. In the first set, the number of basis functions was 9 in all three covariate dimensions, and these basis functions were constructed from second order polynomials. These parameter values correspond to the default model settings in the GWSDAT implementation of the framework (Jones et al. [2023]), therefore, these will be referred to as the default parameter set. In the second set, referred to as the custom set, the number of basis functions was 15 for the two spatial covariate dimensions and 10 for time. These B-spline basis functions were constructed using third order polynomials. The second set provides the model more flexibility, especially in the spatial dimensions resulting in a less smooth fit. The second set of parameter values also increases computational complexity substantially.

3.4 Well-Based Cross Validation

As outlined above, the baseline well influence rankings, to which IA metrics would be compared, were computed using the WBCV approach. In each iteration of the WBCV process, the observations of an individual well are removed from the data to be used as the test set, while the remaining data are used as the training set for the model (leave-one-station-out cross-validation). The RMSPEs calculated for the test set then provide basis for the ranking of monitoring wells, with greater RMSPEs indicating that the removed well had greater influence on the model estimates. Thus, the influence of well k, with $k = 1,2,...,w$, where w is the number of monitoring wells in the network, is estimated via the total RMSPE calculated at the coordinates of k, using the model that that does not include any observations from well k. The model fitted to the reduced data set in each iteration is given by:

$$y'_{(-k)i} = \sum_{j=1}^{m} b_j(x_{(-k)i})\alpha_j + \varepsilon_i, \tag{3.8}$$

where $y'_{(-k)i}$ ($k = 1,2,...,w$, where w is the number of monitoring wells) are the observations of CoPC concentrations excluding the responses from well k and $x_{(-k)i}$ are the corresponding coordinates and sampling times. The RMSPEs were then calculated for each model using the observations and estimated concentrations from the removed wells using:

$$\text{RMSPE}_k = \sqrt{\frac{\sum_{i=1}^{n_k}(y'_{ki} - \hat{y}'_{ki})^2}{n_k}}, \tag{3.9}$$

where RMSPE_k is the root mean squared prediction error for the k-th well, calculated using the model which was fitted to a data set that excluded the observations of the k-th well, y'_{ki} is the i-th observation from the k-th well, $\hat{y}'_{ki}$ is the i-th fitted value for the k-th well and n_k is the number of observations from the k-th well. The well influence rankings are then given by the RMSPE values so that the influence of well k, I_k is proportional to the RMSPE value of the corresponding model $I_k \propto \text{RMSPE}_k$. Hence, the wells can be arranged in numerical order of their corresponding RMSPEs. The underlying assumption is that based on this definition of well influence, a well is more influential if its removal from the data set results in a model with higher RMSPE at its coordinates. This assumption, its limitations and potential other approaches will be discussed in more detail in Section 3.11.

3.5 Influence Analysis Metrics

Influence analysis metrics are a set of techniques commonly used to identify influential observations and outliers in regression analysis (Belsley et al. [2005]). An observation is considered influential if its removal from the data would considerably change the outcome of a statistical procedure such as regression analysis. An influential observation can have a high residual, a high leverage or both. An observation with a high residual would be considered an outlier when compared to the other response variables and an observation with high leverage would be considered an outlier when compared to the other independent variables. The identification of outliers is commonly performed as a diagnostic procedure, because influential observations have the potential to substantially affect the fit of statistical models. Various influence metrics have been proposed in the literature. Some of the most commonly applied ones include Cook's distance (Cook [1977]), DFFITS (Belsley et al. [2005]) and covratio (Merli [2005b]). Typically, observations whose influence metric values surpass a certain predefined threshold are considered influential. In the innovative application proposed in this thesis, rather than identifying outliers, influence metric values are used for quantifying the aggregate influence of all observations originating from the same monitoring wells, and use these aggregate values as the basis for ranking.

In the study presented in this chapter, various influence metrics were applied to potentially identify the most suitable ones for well influence ranking. The investigated influence analysis metrics were leverages (J. Li and Valliant [2009]), standardised residuals (Ranganai [2016]), Cook's distance (Cook [1977]), DFFITS (Belsley et al. [2005]), Hadi's influence measure (Hadi [1992]) and covratio (Belsley et al. [2005]).

3.5.1 Leverages

The leverages are the diagonal elements of the hat matrix **H** (derived in Section 3.3) which takes the form:

$$\mathbf{H} = \begin{bmatrix} \mathbf{h_{11}} & h_{12} & \dots & h_{1i} \\ h_{21} & \mathbf{h_{22}} & \dots & h_{2i} \\ \dots & \dots & \dots & \dots \\ h_{i1} & h_{i2} & \dots & \mathbf{h_{ii}} \end{bmatrix}, \tag{3.10}$$

where they are noted by h_{ii} where i represents the size of the matrix . They are a measure of how far away the explanatory variable values of an observation are from those of the other observations. High leverage observations have a tendency to be influential points and thus the model is likely to fit close to these points (J. Li and Valliant [2009]).

3.5.2 Standardised Residuals

Standardised or internally studentized residuals (Weisberg and Fox [2011]) are residuals divided by an estimate of their standard deviation. The main reason for standardising is that the variances of residuals for different input variables may differ, even if the variances of the errors at these values are equal. Standardised residuals are commonly used in regression analysis to detect outliers in the data. The standardised residuals for the observations are computed using the following equation:

$$r_i^s = \frac{r_i}{\sqrt{\frac{\sum_{i=1}^{n} r_i^2}{n-p}}\sqrt{1-h_{ii}}}, \tag{3.11}$$

where r_i^s represents the standardised residual of the i-th observation, r_i is its residual, h_{ii} is its leverage, n is the number of observations and p is the effective degrees of freedom which is given by the trace of the hat matrix:

$$p = tr(\mathbf{H}) = \sum_{i=1}^{n} h_{ii}. \tag{3.12}$$

3.5.3 Cook's Distance

Cook's distance (Cook [1977]), is commonly used to estimate the influence of a data point in least-squares regression analysis. It is a measure of the change in regression estimates if an observation is deleted. It can be expressed using the standardised residuals r_i^s, the effective degrees of freedom p and the leverages h_{ii} using:

$$CD = \frac{1}{p}(r_i^s)^2 \frac{h_{ii}}{1-h_{ii}}. \tag{3.13}$$

3.5.4 DFFITS

DFFITS (Belsley et al. [2005]) is a diagnostic meant to estimate the influence of a data point in regression analysis. Similarly to Cook's distance, it measures the change in predicted values if an observation is deleted. DFFITS values can be calculated by:

$$\text{DFFITS} = r_i^e \sqrt{\frac{h_{ii}}{1-h_{ii}}}, \tag{3.14}$$

where r_i^e is the externally studentized residual (Weisberg and Fox [2011]) calculated by:

$$r_i^e = \frac{r_i}{\sqrt{\frac{\sum_{j\neq 1}^{n} r_j^2}{n-p-1}}\sqrt{1-h_{ii}}}, \tag{3.15}$$

where $j \neq i$, r_i represents the i-th residual and the first term in the denominator is an estimate of the standard deviation of the i-th residual based on all but the i-th residual. In contrast, for internally studentised residuals (3.11) the standard deviation is estimated using all observations.

3.5.5 Hadi's Influence Measure

Hadi's influence measure (Hadi [1992]) is another measure of data point influence based on the assumption that an observation can be influential in the space of response variables, explanatory variables or both. Hadi's influence measure is calculated by:

$$H_i^2 = \frac{pa_i^2}{(1-(1-h_{ii})a_i^2)} + \frac{h_{ii}}{1-h_{ii}}, \tag{3.16}$$

where, a_i is adjusted residual of the i-th observation given by:

$$a_i = \frac{r_i}{\sqrt{1-h_{ii}}}. \tag{3.17}$$

3.5.6 Covratio

Covratio (Belsley et al. [2005]) quantifies the influence of an observation by measuring the change in the determinant of the covariance matrix of the estimates when deleting the i-th observation. Covratio can be calculated using the following equation:

$$\text{COVRATIO} = \frac{1}{[\frac{n-p-1}{n-p} + \frac{r_i^{e2}}{n-p}]^p (1-h_{ii})}, \tag{3.18}$$

where n is the number of observations, p is the effective degrees of freedom, r_i^e is the externally studentized residual of the i-th observation and h_{ii} is its leverage. The influence of an observation is proportional to the distance of its covratio value from 1. In order to allow for the interpretation of covratio to be the same as the other metrics in the study (i.e. influence is proportional to the the metric value) the covratio values were transformed by $|COVRATIO - 1|$.

3.6 Ranking via Influence Analysis Metrics

In the framework proposed here, observations are grouped by which monitoring well they originate from, and their average influence is considered to represent the influence of the well. Hence the influence of a well is proportional to the average influence metric value of its observations. Well influence ranking can then be done via the average influence metric values. The median was selected as the averaging function for the IA metrics in the proposed approach to provide a more robust average that mitigates the effects of skewness and potential outliers (due to measurement errors) in the data. For standardised residuals, the median absolute deviation (MAD) was calculated, since residuals could also take negative values. To provide an example, let L_k be the ordered list of influence metric values of the observations from the k-th well in the data set. The influence of well k, I_k is then given by

$$I_k = \begin{cases} L_k[\frac{n_k+1}{2}], & \text{if } n_k \text{ is odd} \\ \\ \frac{L_k[\frac{n_k}{2}]+L_k[\frac{n_k}{2}+1]}{2}, & \text{if } n_k \text{ is even} \end{cases}, \tag{3.19}$$

where n_k is the number of observations in well k.

3.7 Comparing Well Influence Rankings

The performances of IA metrics in approximating WBCV results were quantified by calculating a difference score D, that measured the aggregate differences in the placements of each monitoring well between the IA- and WBCV-based rankings. D was calculated using the following equation:

$$D = \sum_{i=1}^{w} |o_i^{wbcv} - o_i^{ia}|, \tag{3.20}$$

where w is the number of monitoring wells in the network, o_i^{wbcv} is the position of the i-th well in the WBCV ranking and o_i^{ia} is the position of the same well in the IA-based ranking. The number of wells, w, can be different in the scenarios, therefore D was normalised to allow for scenario-independent comparisons. This was accomplished by dividing D by the maximum value D can take, which is a function of the number of wells given by

$$D_{max} = \frac{w^2}{2}. \tag{3.21}$$

Notably, D_{max} can be reached by multiple different configurations of rankings. Then, the normalised difference score D_n was calculated by

$$D_n = \frac{D}{D_{max}}. \tag{3.22}$$

D_n is a measure of how precisely monitoring wells were ranked by the IA metric compared to WBCV, with a value of 0 indicating that the rankings are equivalent, and a value of 1 meaning maximum difference in rankings.

3.8 Simulation Study Design

To summarise, 54 different scenarios were generated in total by the combination of the study design parameters. Table 3.1 shows a summary of the variations and possible combinations of these design parameters. The three hypothetical CoPC plumes were the simple, mid and complex plumes (see Figure 3.1). The number of wells ranged between 6, 12 and 24, and the arrangement schemes used to select their locations were random, gird and expert judgment (see Figure 3.2). Finally, either additive or multiplicative measurement noise was used to generate synthetic concentration observations.

Table 3.1: Design parameters and their number of variants, providing the number of possible combinations to generate synthetic scenarios for the simulation study.

Design parameter	**Number of variants**
CoPC plume	3
Number of wells	3
Well arrangement schemes	3
Type of noise	2
Total possible combinations	**54**

The simulations consisted of 100 runs for each of the 54 scenarios. The application of random measurement noise as described in Section 3.1.2 provided variability across the scenarios. In each simulation run, WBCV was performed to establish the baseline in terms of the influence rankings of monitoring wells. Well influence rankings were then computed using each investigated IA metric. These IA-based rankings were compared to the WBCV-based rankings using the normalised difference scores D_n derived in Section 3.7. Finally the performances of the investigated IA metrics were compared to each other via their corresponding D_n values, in order to determine which metrics perform better at approximating the WBCV results.

Figure 3.4 shows a process flow diagram summarising the main stages of the simulation study. Some IA metrics will be abbreviated in the results section as shown in Table 3.2.

Table 3.2: Influence analysis metric abbreviations

IA metric	**Abbreviation**
Mean absolute deviation of standardised residuals	MAD
Cook's distance	CD
Hadi's measure (or potential)	HP

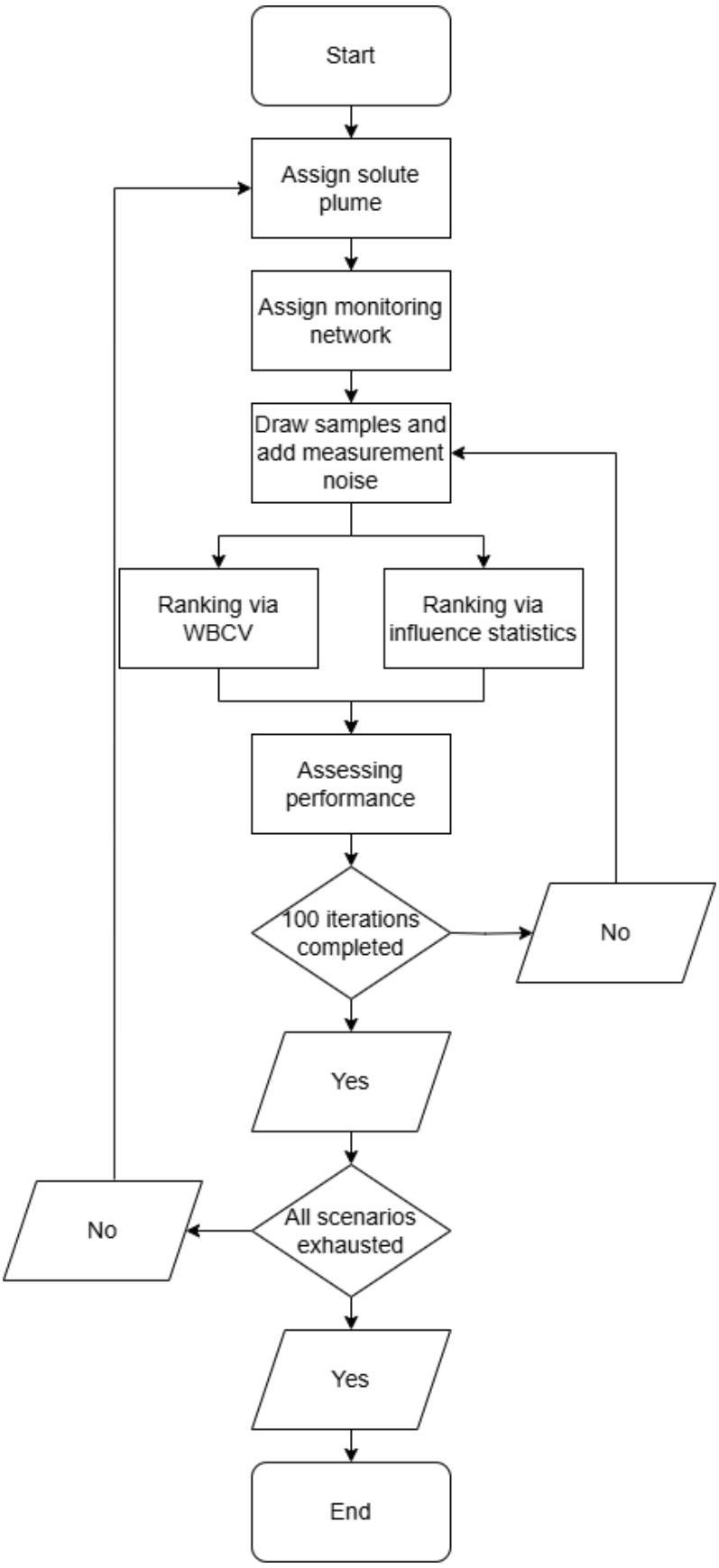


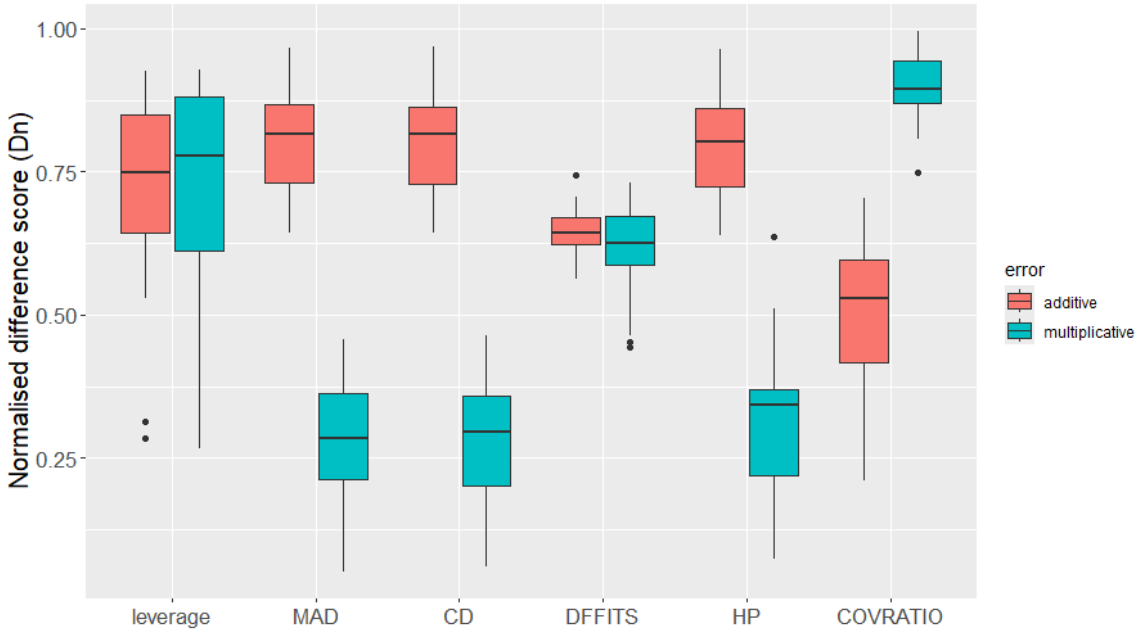
Figure 3.4: Flowchart of the well influence analysis simulation study

3.9 Simulation Study Results

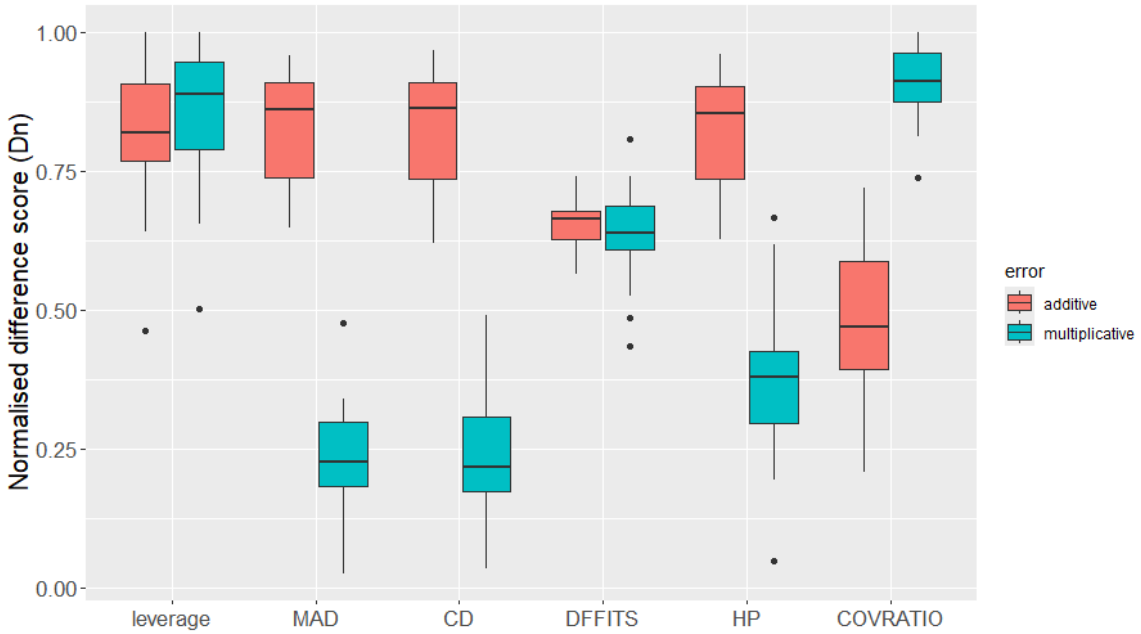
Figure 3.5 shows a summary of the results of the simulations using the two sets of basis function parameter settings, the default set and the custom set (see Section 3.3). Figure 3.5a shows the results for the default, more smooth parameter settings, while Figure 3.5b shows the results for the more flexible custom settings. No substantial difference can be observed in the outcomes of the two model parameter settings in terms of how well the IA metrics performed relative to each other. There is only a slight variation in the distribution of D_n values due to the different model flexibilities and random variation introduced by the measurement noise. This indicates that the models constructed by the two sets of basis functions were sufficiently similar so as to not affect the IA metrics in approximating WBCV rankings.

Figure 3.5, also shows the results based on which measurement error structure was applied on the observation data prior to modelling. Using data with the multiplicative error structure resulted in much better estimations of WBCV rankings when using the MAD of standardised residuals or Cook's distance as influence measures. The CD-based analysis approximated WBCV rankings with the custom model parameter settings by a median difference of $D_n = 0.22$ or 22%. This means that across the 27 scenarios that used data with multiplicative errors, CD and WBCV influence ranking results only differed by 23% on average. Using the MAD of standardised residuals in these scenarios resulted in a difference of $D_n = 0.23$ or 23%, followed by Hadi's influence measure with $D_n = 0.38$, DFFITS with $D_n = 0.64$ and leverages with $D_n = 0.89$. Well influence rankings estimated by covratio were farthest from WBCV results with $D_n = 0.91$ or 91%. By contrast, using data with additive error structure resulted in covratio providing the closest approximations with $D_n = 0.48$ or 48%, which was still less precise than CD and MAD in the multiplicative error scenarios. With additive errors, CD, MAD and leverages least accurate results compared the WBCV rankings.

Figure 3.6 shows a comparison of the changes in difference scores (D) achieved by the IA metrics across 100 simulation runs using the custom model parameter settings for the scenario with complex plume, an expert network design and 24 monitoring wells. The figure shows the results for the data with additive (3.6a) and multiplicative (3.6b) noise. These results substantiate the difference seen between additive and multiplicative errors in Figure 3.5. Different variability can be observed for different IA-metrics, which also depends on the error type added to the modelled observations. For example, covratio produces highly variable results in the additive error scenario but very low variability for data with multiplicative noise. The other methods appear less sensitive to the errors. DFFITS seems the least affected by the error type, having similarly high variability in both cases. The variability of leverages and HP also decreases somewhat in the multiplicative error scenario compared to the additive one. It can also be observed that MAD and CD produce very similar results in both cases.

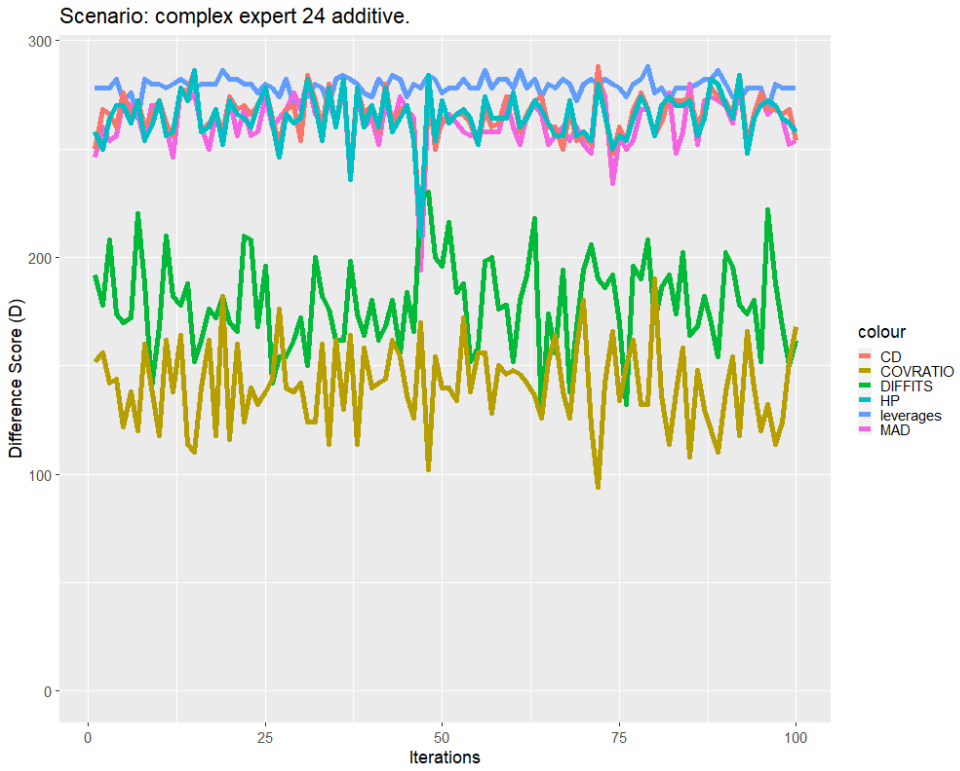


(a) Default model parameters



(b) Custom model parameters

Figure 3.5: Box plots summarizing the normalised difference scores (D_n) achieved by IA metrics in 100 simulations across all 54 scenarios. The results are grouped by the type of measurement noise in the observations. Lower D_n scores indicate better approximation to WBCV rankings. 3.5a shows the results for the default basis function settings (9 quadratic basis functions in all 3 dimensions), while 3.5b shows the results for the custom basis function settings (15 cubic basis functions for each spatial dimension and 10 for time). See Section 3.3 for the description of the two settings.



(a) Additive noise

(b) Multiplicative noise

Figure 3.6: Results showing the difference scores D achieved by the different IA metrics across 100 iterations of the simulation for the complex plume with expert monitoring network design and 24 wells. 3.6a shows the results when the observations contain additive noise, while 3.6b shows the results when they contain multiplicative noise.

This could be due to the fact that the value of CD depends highly on the standardised residuals (see Section 3.5.3). CD, MAD and HP appear to be sensitive to the error structure and produce more accurate approximations to WBCV if the observation data has multiplicative errors.

As mentioned in Section 3.1.2 groundwater quality monitoring data are commonly assumed to contain multiplicative noise associated with errors in the sampling procedure or the analytics used to measure solute concentrations (McLean et al. [2019]). The results of the simulation study indicate that given this assumption, IA metrics such as CD or MAD could approximate WBCV rankings reasonably well. Additionally, the data set contains a higher ratio of low concentrations. When additive errors are applied to this data, the ratio of noise on these low concentrations will be much larger than in the multiplicative case. This high amount of noise can result in unrealistic model estimates, that affect the ratio of influential observations, making the proposed approach less effective. Therefore, when looking at the impact of study design parameters on the results, the focus will be placed on the scenarios where observations containing multiplicative errors were modelled.

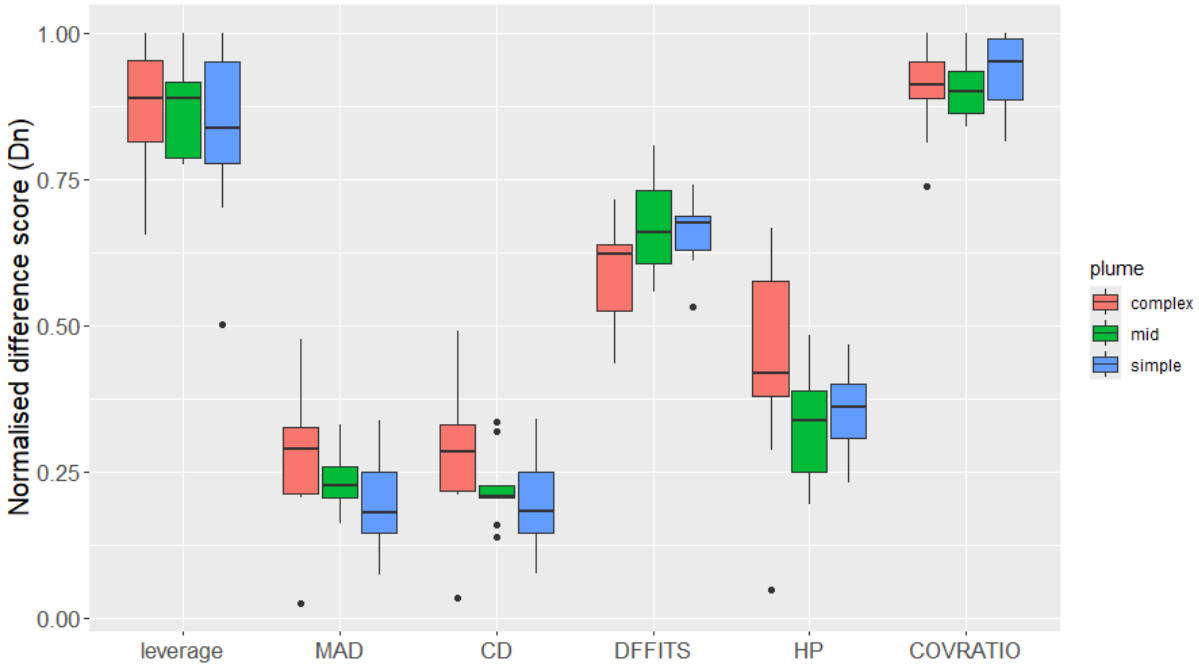


Figure 3.7: Box plot showing the normalised difference scores (D_n) achieved by IA metrics in 100 simulation runs across the 27 scenarios with multiplicative noise. The results are broken down by plume scenario (simple, mid and complex). Lower D_n values indicate better approximation of the WBCV ranking.

Figure 3.7 shows the normalised difference scores (D_n) of well influence rankings produced by the different IA metrics grouped by the spatial complexity of the CoPC plume. The IA metrics with the lowest average D_n, namely CD and MAD, show an increasing accuracy in approximating the WBCV rankings, with decreasing plume complexity.

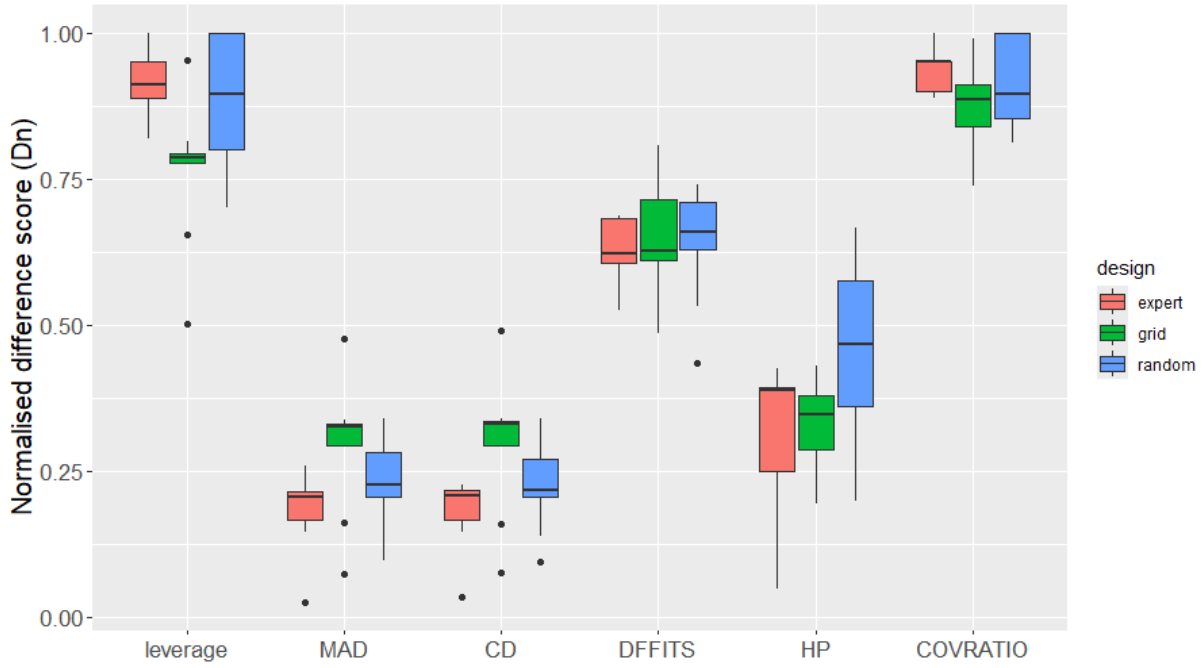


Figure 3.8: Box plot showing the normalised difference scores (D_n) achieved by IA metrics in 100 simulation runs across the 27 scenarios with multiplicative noise. The results are broken down by monitoring network design (random, grid, expert). Lower D_n values indicate better approximation of the WBCV ranking.

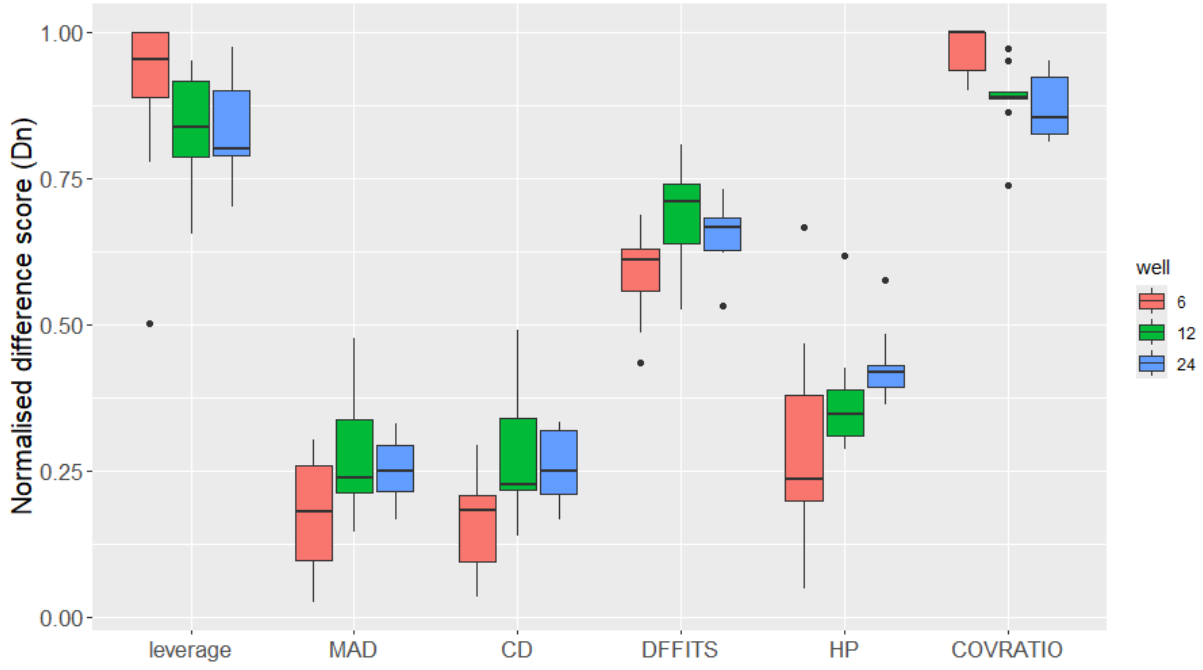


Figure 3.9: Box plot showing the normalised difference scores (D_n) achieved by IA metrics in 100 simulation runs across the 27 scenarios with multiplicative noise. The results are broken down by the number of monitoring wells (6, 12, 24). Lower D_n values indicate better approximation of the WBCV ranking.

The median D_n score achieved by CD increases from 18% to 28% between the simple and the complex CoPC plume scenarios. This suggests that well influence analysis on a more spatially complex CoPC plume could result in a less precise estimation of the CoPC concentration surface. A possible explanation for this outcome is that the model fit results in larger residuals for CoPC plumes with more complex spatial structures. This in turn can affect the influence metric values and increase differences between the IA and WBCV-based well rankings.

Figure 3.8 shows the results based on the monitoring network design in the scenario. For CD, the expert design scenarios resulted in the lowest D_n scores with a median of 0.21 or 21%, while the grid design scenarios resulted in the highest score at a median of 29%. Overall, most IA metrics provided the most precise approximations to the WBCV rankings in the expert monitoring network design scenarios. Only covratio and leverages performed better under the grid design scenarios. These results indicate that the arrangement of sampling locations can influence the ranking process when using IA metrics.

Figure 3.9 shows the results based on the number of monitoring wells used in the scenario. The lowest D_n values were reached in scenarios with 6 monitoring wells when using CD. These scenarios resulted in a median D_n value of 0.18. The highest median D_n value was obtained in the 12 monitoring wells scenarios. These results could be explained by the fact that when ranking fewer wells there are fewer possible permutations for the two rankings to differ, thus resulting in lower D_n scores on average. The difference between using 12 and 24 wells is much smaller in terms of D_n values, potentially indicating that the effect of fewer permutations is minimised once a sufficiently large number of wells are introduced.

3.10 Case Study Results

Table 3.3: Normalised difference scores D_n achieved by different influence metrics approximating the WBCV rankings of the sampling locations based on the monitored constituents in the case study data.

Constituent	Leverage	Standardised residuals	Cook's distance	DFFITS	Hadi's measure	COVRATIO
Ethylbenzene	0.43	0.38	**0.36**	0.63	0.38	0.60
Toluene	0.48	0.30	**0.25**	0.64	0.34	0.86
Nitrate	0.57	0.54	**0.18**	0.76	0.73	0.77
Sulphate	0.36	0.62	**0.23**	0.72	0.29	0.47
TPH	0.57	0.54	**0.32**	0.48	0.37	0.63
Median	0.48	0.54	**0.25**	0.64	0.37	0.63

Table 3.3 shows the D_n scores achieved by the different influence statistics in the case study. The best performance was achieved by Cook's distance with a median D_n of 0.25, which is close to the result achieved in the simulation study. The other influence statistics did not perform well in the case study with Hadi's measure reaching a median D_n of 0.37, and leverages and studentised residuals reaching 0.48 and 0.54 respectively. Thus, Cook's distance provided the most favourable results in both the simulation study and the case study. Moreover, the computation time using WBCV for the initial ranking for a single solute was 4.66 minutes, whereas with the Cook's distance-based redundancy metric it was only 14.33 seconds, which is a reduction of around 95%.

Table 3.4: Comparison of well ranking by influence in increasing order computed using WBCV and Cook's distance in the case study for the solute Nitrate. The values in the three columns represent the identification numbers of sampling wells and their influence ranks in increasing order.

Rank	**WBCV**	**CD**
1	13	13
2	18	18
3	10	14
4	14	15
5	15	10
6	19	19
7	9	7
8	7	2
9	3	9
10	2	3
11	8	4
12	11	8
13	5	17
14	1	16
15	6	1
16	4	11
17	16	6
18	17	5
19	12	12

Table 3.4 shows the influence rankings computed by WBCV and Cook's distance for the solute nitrate, which received the lowest D_n value (0.18). The table shows that the Cook's distance-based influence metric approximated the WBCV ranking very closely in this case.

Table 3.5: Comparison of well ranks by influence in increasing order computed using WBCV and Cook's distance in the case study for the solute Ethylbenzene. The values in the three columns represent the identification numbers of sampling wells and their influence ranks in increasing order.

Rank	WBCV	CD
1	27	22
2	28	23
3	21	24
4	22	27
5	24	28
6	26	26
7	23	21
8	25	25
9	14	15
10	5	7
11	4	2
12	16	1
13	12	12
14	1	16
15	3	4
16	13	19
17	15	6
18	8	5
19	6	18
20	10	14
21	18	8
22	20	17
23	9	11
24	11	20
25	17	10
26	2	9
27	19	13
28	7	3

Table 3.5 shows the same comparison for ethylbenzene, which in the case study received the highest D_n value (0.36). The table shows that despite the higher difference score, most wells hold similar positions in both rankings, especially towards the top of the table. Hence, the proposed influence metric identified the same wells as being the most redundant as WBCV, with

only minor differences in rank positions. The two tables indicate that the influence metric tends to approximate WBCV results better for sampling locations with lower influence (towards the top of the table).

This does not diminish the utility of the approach since the aim of the analysis is to identify potentially redundant sampling locations. The wells estimated to be more redundant in Tables 3.4 and 3.5 are either located in well clusters, i.e. in close proximity to each other, such as wells 27, 28, 26, 22, 24 in Table 3.5 or at the boundaries of the network such as wells 13, 18, 10, 14 in Table 3.4 (see Figure 3.3 for well locations). Wells located in clusters can be less influential because they provide observations from the same spatial location (assuming similar sampling depths). The influence of boundary wells largely depends on the characteristics and extent of the solute plume. Boundary wells can be important in constraining the model estimates. However, if they are estimated to be more redundant, it might indicate that the solute plume is far from the boundaries of the network, making boundary observations less influential.

3.11 Discussion

The results indicated that Cook's distance could be used as a computationally efficient approximate to WBCV to rank groundwater monitoring wells by their influence on statistical model estimates, given that the assumption of multiplicative noise in the data holds. The well influence rankings computed using the Cook's distance values of the observations were on average only 22% different from ones computed using WBCV in the simulation study and 25% different in the case study. As shown in Tables 3.4 and 3.5, the differences in ranking were mostly due to small differences in rank placements by the two investigated methods, rather then large deviations. In general, there were no substantial disagreements between the two approaches on whether a monitoring well was placed in the lower or upper part of the influence spectrum. Given that the proposed application of IA-metrics is intended to provide a starting point for a more exhaustive investigation of the impact of removing wells, the obtained results are adequately close to WBCV. The saving in computation times in the case study is also substantive.

There was a very apparent difference in the performance of well influence analysis depending on whether the groundwater quality monitoring data had multiplicative or additive measurement errors. The influence statistics could not approximate the WBCV results as closely if the monitoring data had additive noise. This might be explained by the scale of the noise in the additive case. Multiplicative errors are scaled by the observation they are added to, meaning small CoPC concentration measurements will have smaller corresponding errors. This is not the case with additive errors, which are independent of the concentration measurements, and their range is determined by the chosen distribution (see Section 3.1.2). Therefore, the residuals of

small concentration estimates in these cases can be substantially larger than with multiplicative noise. Since IA metrics commonly depend on the residuals, this can result in an increase in the IA metric values of certain observations, thus affecting the ranking procedure. Notably, the investigated data sets contained a higher ratio of low concentrations. Thus, when additive errors are applied to this data, the ratio of noise on these low concentrations will be much larger than in the multiplicative case. This high amount of noise can result in unrealistic model estimates, that affect the ratio of influential observations, making the proposed approach less effective. Since groundwater quality monitoring data are commonly assumed to have multiplicative noise, this phenomenon is not expected to have a large impact on rankings in real data sets. This is substantiated by the analysis on the case study data set. Thus, the analysis of the results in this chapter focused on the scenario where multiplicative measurement noise was applied. However, the robustness of the proposed well influence ranking approach could be improved by analysing the error structure of the monitoring data prior to ranking or by improving the method to include safeguards against these effects.

The results showed that the geometry of the CoPC plume also affects the performance of the CD-based well influence analysis. Using well influence analysis on a more spatially complex CoPC plume will generally result in a less precise estimation of the CoPC concentration surface. The model can produce larger residuals given a CoPC plume with a more complex spatial structure. This in turn can affect the influence metric values and increase differences between the IA and WBCV-based well rankings. The negative effects of a complex plume geometry on well influence analysis can somewhat be mitigated by allowing more flexibility in the model, e.g. by increasing the number of B-spline basis functions used to construct the P-spline model.

The spatial arrangement of monitoring wells also seemed to affect the results through the precision of model estimates similarly to spatial plume complexity. Expert designs incorporating hydrogeological information such as groundwater flow direction and source of contamination produced better outcomes than randomly or regularly placed monitoring wells, suggesting that the more appropriately a network design can capture the spatial heterogeneity of the concentration surface, the better the estimation of well influence will be. It should be noted that the difference between the results in the different scenarios (different plumes and network designs) is substantial, suggesting that the differences are not purely the result of random variation, but indicate systemic effects.

The lowest normalised difference scores were reached in the scenarios with the lowest amount of wells (6), despite having less data about the underlying CoPC plume. These were followed by the ones with the highest (24) number of wells. The difference in scores between the 12 and 24 well cases was substantially smaller than between the 6 and 24 well cases. This outcome may be the result of fewer possible permutations in rankings between WBCV and the IA-metrics given

fewer monitoring wells. Since there are fewer ways of ordering the monitoring wells, there is simply less room for disagreement between the two methods. However, given a sufficient amount of wells that are able to capture enough of the underlying contamination pattern, this effect seems to level out as indicated by the small difference in scores between the 12 and 24 well cases. Similarly to the other study design parameters (plume and network design), the difference in outcomes is substantial but the impact of random variation could not be ruled out. Repeating the simulation study with a higher number of runs, could help substantiate the results.

When comparing two ranked lists, such as done in this study, the largest possible difference score can be achieved by multiple different permutations, i.e. there is not one set of arrangements that corresponds to maximum difference. One of the permutations that achieves this is where one ordered list is the reverse of the other. In these cases it could be argued that the worst performing influence metric could approximate the WBCV results better if its reverse order is considered. However, the number of permutations resulting in the maximum difference score increases with the number of items in the lists, i.e. the number of monitoring wells. Therefore, given a large dataset it is highly unlikely that the worst performing method will get close to the reverse well influence order. This makes applying this form of the well influence analysis method unreliable. Thus, it is still a better strategy to rely on the best performing influence diagnostics such as Cook's distance.

Another point worth mentioning is that there are several possible ways of comparing two ranked or ordered lists besides the one used in this study. One of these methods is called the Kendall tau distance (Kendall [1938]). In short, the Kendall tau distance is the sum of scores assigned to each pair of elements in a set, based on their respective order in two lists that are being compared. If the pair follow the same order in both lists, they are assigned a score of 0, if their order is reversed, they are assigned a score of 1. The scores of each element pair are added up and this gives the Kendall tau distance. In principle, this approach is similar to the one applied in this study.

The proposed innovative approach using Cook's distance to rank groundwater monitoring wells by their influence as described in this chapter is currently implemented in the well redundancy analysis feature of version 3.2 of GWSDAT (Jones et al. [2022]). As highlighted in this chapter, the implementation of this functionality is based on feedback from the GWSDAT user base. The well redundancy analysis feature allows users to omit observations of selected monitoring wells from their networks to assess the impact of leaving out certain wells on the model estimates. The selection of wells in this feature appears in order of estimated influence by the CD-based well influence analysis. The estimated well influence ranking informs the user of the approximate importance of the monitoring wells, thus improving the efficiency of more exhaustive, manual analyses.

There were a few limitations in performing and interpreting the results of this simulation study. One of the limitations was using the WBCV results as the indicator for the true influence of monitoring wells. Since in evaluating the IA metrics, the outcomes depend on how the importance of a well is defined, there could be different ways of determining what the actual well ranking could be based on. Such a metric could be the precision in estimating the entire CoPC concentration surface, instead of the concentrations only at the location of the removed well. In the approach proposed in this study, WBCV provides an indication as to how difficult it is to estimate the observations of a well based on the observations of the other wells. Given a sufficiently flexible model and the other observations held constant, changes in estimates are expected to occur at the location of the removed well. Moreover, averaging across the entire CoPC concentration surface could result in micro-scale differences in estimates being unidentified. Therefore, using this metric as an indicator of true well influence is an appropriate approach. In the future, different approaches, such as considering the entire concentration surface could be explored to substantiate the obtained results.

It should also be noted that there are a variety of other aspects to consider before omitting or decommissioning sampling wells in long-term groundwater monitoring networks and the results of the influence analysis alone are not sufficient justification. It is important to consider the position of the monitoring wells with regards to groundwater flow direction, hydrological features of the site and the distribution of the CoPC plume before committing to the omission of any well from upcoming sampling campaigns. Thus, while the output of the influence analysis is independent of hydrogeological site conditions, the knowledge of these in combination with the output helps identify potentially redundant wells in monitoring networks.

Another limitation to consider is the reliability of the influence diagnostics given their innovative use in this case. Normally, influence diagnostics are used to identify influential observations such as outliers or high leverage data points, by checking whether they cross a certain threshold value. Data points that cross the threshold warrant a second look to determine whether they should be removed from the model to improve the fit. In this case, it was assumed that the influence of an observation is proportional to its influence metric value, and thus the well-wise averages of these metrics could be used to order the monitoring wells. The impact of using different averaging functions or potentially considering the sum of influences instead, could be investigated as well.

A potential area of improvement in well redundancy ranking is the implementation of more computationally efficient cross validation approaches. Cross validation could provide a more accurate view of the changes that occur in model estimates when certain wells are omitted from the data set. Reducing its high computation times would allow better integration in groundwater quality data analysis software such as GWSDAT. In the R package *mgcv* (S. Wood [2021])

for example, neighbourhood cross validation (NCV) can be used to optimize smoothing parameters in GAMs. In order to reduce the computational cost of this optimisation procedure, S. Wood [2024] developed a quadratic approximation to the NCV criterion (QNCV). This approach makes the computational cost of NCV comparable to a single model fit, which could represent an $O(n)$ saving when modelling n data. If this approach could be generalised and implemented in the context of well redundancy ranking, it could provide a computationally comparable alternative to the influence metrics. The efficiency, accuracy and explainability of such an approach would have to be investigated.

The work presented in this chapter has been peer reviewed and published in the proceedings of the 37th International Workshop on Statistical Modelling (Radvanyi et al. [2023]), and has been submitted to the journal of Environmental and Ecological Statistics[2] for review. The code used to perform the simulation study and the case studies was written in the open-source statistical programming language R (R Core Team [2020]), and is available on GitHub[3]. A *shiny* (Chang et al. [2021]) web application has also been developed to allow for the reproduction of the simulation study results[4]. The application allows for the selection of the three different CoPC plume data sets, monitoring network design types, the number of monitoring wells and the measurement error structure, as well as the amount of noise added to the data, the number of segments for the three covariate dimensions (easting, northing and time) in the P-splines model and the degrees of polynomials in smoothing spline function. The application displays the difference score results for six different influence metrics (Cook's distance, covratio, DFFITS, Hadi's measure, leverages and mean absolute deviation of studentised residuals) and shows the actual obtained rankings from these metrics as well as the ranking obtained via WBCV. This allows for a more direct comparison between the approaches.

In conclusion, the influence statistics-based method outlined in this study was intended to provide an approximate, computationally efficient solution for indicating how important different groundwater monitoring wells are in terms of their influence on the statistical modelling of the data. The results show that the proposed well influence analysis approach could be a convenient tool for this purpose, and hence, could aid professionals involved in groundwater monitoring in designing monitoring networks and sampling campaigns.

[2]https://link.springer.com/journal/10651
[3]https://github.com/peterradv/Well-Influence-Analysis
[4]https://peterradv.shinyapps.io/well-influence-analysis/

Chapter 4

Spatially Balanced Sampling Designs for Groundwater Monitoring

In Chapter 3 a framework was proposed for ranking wells in a groundwater monitoring network based on their influence on estimating a spatiotemporal CoPC concentration surface. The ranking aimed to provide information about which monitoring wells could potentially be left out of future sampling designs, and about the potential impact of needing to decommission a given well. The well influence analysis framework is a particular approach to optimising sampling intensity for existing monitoring networks that focuses on the removal of redundant sampling locations, but does not provide suggestions for sample selection in space and time. This chapter will investigate this more general aspect of sampling designs to consider and propose spatiotemporal sample selection methods that could be used to inform future sampling campaigns.

As established in Chapter 2, optimal groundwater quality monitoring sampling designs should aim to minimise the presence of information sparsity in terms of space and time within the collected data. A class of probability sampling techniques called spatially balanced sampling designs aim to spread samples evenly in two or more dimensions whilst maintaining a stochastic component. Hence, these techniques can help minimise data sparsity when selecting sampling locations from a pool of pre-existing potential locations. The aim of this chapter is to explore current literature on spatially balanced sampling design algorithms, compare different methods and analyse their potential utilisation in long-term groundwater contamination monitoring well networks to support the descriptive, spatiotemporal modelling of groundwater CoPC concentration levels. At the end of the chapter, different spatially balanced sampling techniques are compared in drawing balanced samples in three dimensions over space and time. As a result of the literature review and the comparison, a proposal is made for which approach appears more beneficial for designing a groundwater monitoring network sampling strategy that is balanced

over space and time. To the best of the authors knowledge, this provides a novel contribution in the field of long-term groundwater contamination monitoring.

4.1 Introduction

As highlighted in Chapter 1.2.2, groundwater is generally difficult to sample because it is difficult to reach. Access to groundwater aquifers requires the installation of wells, which can be a very costly and hazardous endeavour. Therefore, the positioning of these access points in relation to surface and hydrogeological features requires thorough consideration. There is extensive literature on the optimal design of groundwater quality monitoring networks, with the problem of finding optimal locations for installing new monitoring wells receiving most of the attention (Farlin et al. [2019]). Methods for solving such optimisation problems range from purely data-driven to purely process based approaches. Farlin et al. [2019] provides a summary of the mathematical and statistical tools that have been applied to this problem, which include geostatistical modelling, principal components and cluster analysis, entropy, Bayesian optimisation and groundwater flow models.

Developing sampling strategies for already established well networks received much less attention, especially in a spatiotemporal framework (Farlin et al. [2019]), despite it being important for reducing the costs, workload and health hazards to personnel induced by sampling campaigns. Notably, McLean [2018] introduced two objective functions (variance of plume mass and integrated prediction variance) that could be used to optimise the sample selection in existing monitoring networks in both space and time. However, these are computationally more demanding, iterative approaches based on the minimisation of the variances of different population estimates, and thus result in non-probability sampling designs. Additionally, these methods have not been tested in terms of their precision in estimating the CoPC concentration surface.

In long-term groundwater contamination monitoring projects, the concentrations of constituents of potential concern (CoPCs) are measured over time, often via the frequent sampling of monitoring wells. The collected data along with hydrogeological parameters can subsequently be used to estimate the characteristics of the CoPC plume (such as size, mass, velocity and flow direction) through the use of mathematical and statistical modelling tools, and to identify trends in concentration levels. The main objectives of long-term monitoring operations are to verify the stability of the remediation process and to detect anomalies and changes in the spatial distribution of CoPC concentration levels. Sustainable remediation (SR) has become an important concept since its emergence in the late 2000s (U.S. Sustainable Remediation Forum [2009]). SR addresses some of the key problems with traditional remediation approaches, such as post-treatment, low-level residual contamination and the environmental impacts (waste production,

noise, traffic, air pollution, ecological disturbances, energy use and greenhouse gas emissions) associated with the remediation technologies themselves (Meray et al. [2022]). It promotes the adoption of sustainable practices in remediation projects and the transition from intense, active soil treatment approaches to more passive ones such as monitored natural attenuation (MNA; Meray et al. [2022]). MNA relies on the observation of natural processes to achieve remediation objectives rather than direct intervention. Optimising when and which monitoring wells to sample in such systems is essential for increasing the sustainability of monitoring activities.

There are several aspects to consider when developing a sampling strategy for an existing network of monitoring wells[1]. Such aspects include the monitoring objectives, the incurred costs and health hazards of sampling, the hydrogeological conditions of the site, land use, accessibility, the target type of contamination, seasonal changes in groundwater levels and the distance to potential sources of contamination. Thus, groundwater sampling designs for existing monitoring well networks are often determined by non-probabilistic methods such as expert judgement or regulators' opinions (Meray et al. [2022]). Therefore, the resulting catalogue of observations can be difficult to analyse statistically and can result in biased estimates of CoPC plume characteristics (Kermorvant et al. [2019b]).

Probabilistic sampling designs (that involve a stochastic component) can aid experts in developing sampling strategies that provide samples that capture more of the spatial and temporal variation in solute concentrations, can be analysed statistically and are more economical. A particular set of such techniques are called spatially balanced sampling designs, which aim to select a set of sampling locations that are evenly spread over the target area, which in this case includes the CoPC plume (Benedetti et al. [2017]). The rationale behind such designs is to maximize sample coverage over the target of interest to reduce bias and capture more of the spatial characteristics of the population for statistical modelling and consequently reduce sampling costs (Kermorvant et al. [2019b]). Well-spread samples are usually also preferred if the target variable is spatially stratified (i.e. its value has significant variation between subregions) and if it exhibits a cluster structure (Benedetti et al. [2017]). In such cases, a single sample can be enough to characterize the rest of the cluster, and so it makes sense to spread out the remaining samples. As explored in the previous chapters, these characteristics commonly apply to groundwater contamination monitoring data, due to the spatial heterogeneity of the CoPC plume.

If the objectives of an environmental survey include measuring changes in the target variables over time, then the sampling design has to address this dimension as well. A reasonable option is to draw spatially balanced samples repeatedly. In this case, the time between sampling campaigns can be determined independently and based on for example, expert judgment. Another option is to extend the sampling design algorithm to spread the samples in space and time si-

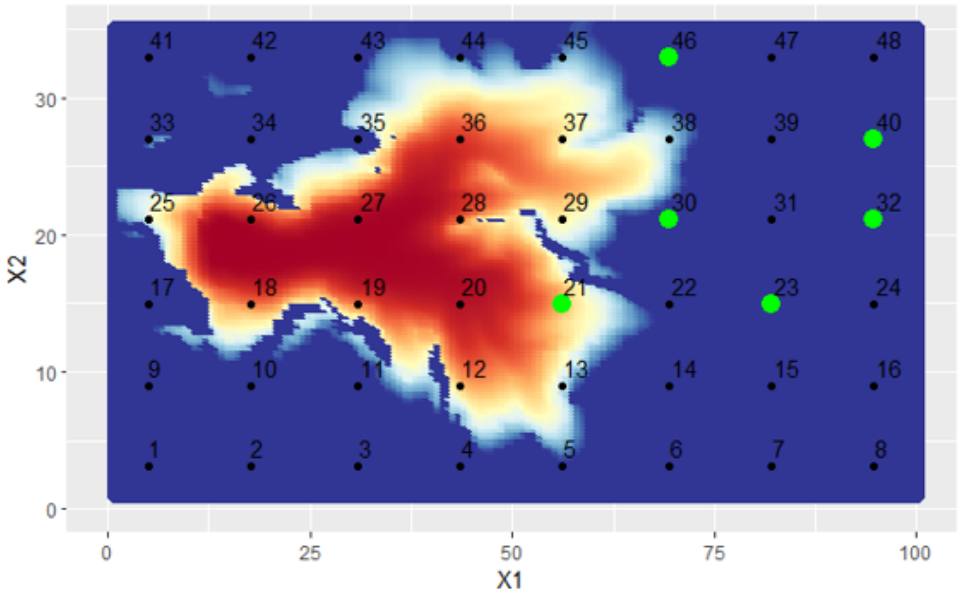
[1] https://semspub.epa.gov/work/HQ/100001800.pdf

multaneously. In this case the spatial coordinates and time are treated as a three-dimensional space that can be continuous or made up of discrete points. A pre-defined number of samples can then be evenly spread in this space to provide a spatiotemporally balanced set of samples. Yet another option is to divide the available sampling locations into two or more spatially balanced groups and sample them in an alternating manner from campaign to campaign. Similarly to option one, in this case the time between sampling campaigns can be determined independently. The advantage of the spatiotemporal approach is that it minimises data sparsity in the temporal domain as well, while in the repeated spatial approaches the sampling frequency is an additional parameter to be optimised. Many of the spatially balanced sampling algorithms can be extended to more than two dimensions (see for example Robertson et al. [2018] or Grafström et al. [2012]). Initially, the motivation for the extension in the literature was to address the fact that many environmental resources or variables of interest are distributed in three spatial dimensions (e.g. depth below ground surface, depth of water column, etc.), but time can also be treated simply as an extra spatial dimension in this case.

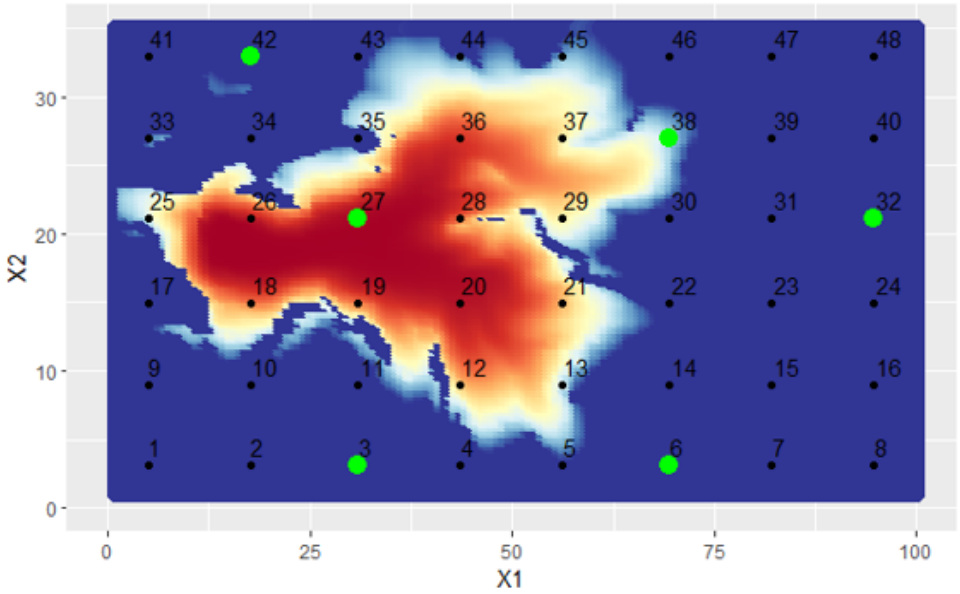
Another aspect to consider in terms of the temporal distribution of the samples is the trade-off between the frequency of sampling campaigns and the number of samples collected per campaign. Sampling more frequently can increase the temporal resolution of the data but has economical drawbacks and can have detrimental environmental impact due to the required transportation and machinery. Less frequent but more intensive sampling is usually more economical due to the costs associated with transportation. Moreover, collecting fewer samples per campaign also decreases spatial resolution, which may result in a poorer estimation of the spatial structures associated with the environmental variable of interest.

There are limited examples in the literature of spatially balanced sampling designs being applied to groundwater quality monitoring problems. One such example comes from the USEPA (Olsen [2023]), where they used a spatially balanced design algorithm called generalized random-tessellation stratified (GRTS; Stevens and Olsen [2004]) on existing private and municipality-owned wells to estimate the water quality of groundwater aquifers in a subregion of Florida. Additionally, the study only considered the balance of samples in the spatial domain and not in the temporal domain. There are however, several examples of spatially balanced sampling designs being applied more generally in the field of environmental science, especially ecology (see Kermorvant et al. [2019b], Abi et al. [2017], Kermorvant et al. [2019a] and Koski and Eidsvik [2024]). In their review Kermorvant et al. [2019b] encourage all environmental scientists to use spatially balanced sampling designs for problems that require inferences on spatially distributed populations from samples. Stevens and Olsen [2004] have developed the GRTS design as a general framework for sampling natural resources. The monitoring of groundwater CoPC plumes falls into this category of problems. Although these techniques only started receiving more attention in the past two decades (since the proposal of GRTS), the relatively few examples for the

application of spatially balanced sampling designs, esecially in a spatiotemporal context, indicates a gap between long-term groundwater contamination monitoring and other environmental science disciplines.



(a) Simple random sampling



(b) Spatially balanced sampling

Figure 4.1: Spatial samples of $n = 6$ for a hypothetical grid monitoring network of 48 wells using a simple random and a spatially balanced sampling design. The green dots represent the selected monitoring wells. The underlying image shows the log-transformed CoPC concentration values of the complex simulated plume (see Section 1.4.2) at the final time point.

Moreover, there is a general gap in the application of these techniques in the spatiotemporal

sense. Therefore, the assessment and proposals made in this chapter and more broadly in this thesis are contributions towards filling these identified gaps.

In the case of groundwater contamination, the spatiotemporal distribution of the plume is determined by the location of the CoPC's source, the rate of its release, its chemical properties and the hydrological and hydrogeological conditions (aquifer composition, recharge rate, hydraulic conductivity, etc.), since they determine groundwater flow patterns. Sampling strategies for groundwater contamination monitoring networks should aim to capture as much of the spatial and temporal structure of the plume as possible, to improve the statistical estimation of its characteristics. As shown by Benedetti et al. [2017], a spatially balanced technique is guaranteed to provide a more balanced sample than simple random sampling and thus, it generally does better at capturing the spatial heterogeneity of the target population. To illustrate the difference between simple random sampling (SRS) and spatially balanced samples, Figure 4.1 shows a comparison between the two in selecting a sample of $n = 6$ out of a hypothetical, idealised, grid-like monitoring network consisting of 48 wells. The underlying image shows the log-transformed CoPC concentration values of the complex simulated groundwater plume data (see Section 1.4.2). The comparison shows that the wells selected randomly by SRS are all in close proximity of each other, whereas the wells selected by the spatially balanced method, while still random, are well spread over the given area.

The following sections will describe spatially balanced sampling designs in more detail and introduce different approaches. Several studies from the literature will also be discussed that compare different spatially balanced sampling methods based on the spatial balance of the obtained samples and the estimation of population characteristics. Finally, an exploratory study will be presented that compare some of the investigated techniques in drawing spatiotemporally balanced samples. Based on the collected information from the literature and the conducted comparison, a proposal will be put forth on which spatially balanced sampling algorithm would be most suitable to implement for long-term groundwater quality monitoring networks.

4.2 Simple Random Sampling

The simplest probabilistic sampling design is simple random sampling (SRS), in which there is no special objective to be achieved with regards to the spatial distribution of the samples. SRS simply selects randomly from the set of possible sampling locations it is given. SRS in the context of spatial environmental sampling is usually done without replacement, since there is no need to collect multiple samples from the same point location at a given time. Data collection over time is commonly done using repeated spatial sampling, where locations can be selected again at different times.

In long-term groundwater contamination monitoring, the monitoring wells represent potential sampling locations. As mentioned before, the spatial arrangement of these wells is defined by many different natural and anthropogenic factors and is subject to expert judgement, planning and optimisation (see Section 1.2.3). This constrains probability sampling methods in terms of where samples can be drawn from in space. However, as discussed in Chapter 1.1, probability samples commonly have several advantages over non-probability samples as they reduce bias, ensure the generalisability of the results and allow for inference where the uncertainty can be characterised by the random sample design. Since SRS does not take into account the distance between neighbouring sampling points, the resulting repeated spatial samples can be prone to clustering both in space and time. As highlighted in Section 4.1, clustered samples provide less information on the target population parameters than more spread out samples (Benedetti et al. [2017]). Thus, there is a high variability associated with SRS samples in terms of how well they can represent the spatial structure of the CoPC plume. Although this is constrained by the locations of the monitoring wells in relation to the plume, SRS samples in general are still less reliable in this respect than spatially balanced samples (Benedetti et al. [2017]).

4.3 Accounting for Spatial Dependence

Spatial dependence represents the degree to which the value of a particular response variable z_i of spatial unit i is related to the response z_j of a nearby unit j. In spatially correlated data, nearby units are expected to be more similar to each other than to more distant units. Spatially balanced sampling designs generate sampling strategies that account for the spatial dependence commonly present in data that characterizes a population distributed in space, and thus minimise the expected variance of population estimators. Thus, a design-based inference can be obtained, which allows for the presence of spatial dependence, and its uncertainty is characterised by the randomised sample design .

Benedetti et al. [2017] derived how the minimisation of the variance of a population estimator leads to a spatially balanced sample. Let $\mathbf{z}$ be responses from a finite, spatially distributed population and $\mathbf{X}$ a matrix of some available auxiliary information. Assume that based on prior knowledge, the finite population can be viewed as a sample from an infinite superpopulation, and there is a model ξ that defines the characteristics of the superpopulation, which are also valid for the current survey period. The sampling design is then defined on the basis of the expected moments of $\mathbf{z}$ given $\mathbf{X}$. Let $\hat{Z}$ be the Horvitz-Thompson (HT) estimator (Horvitz and Thompson [1952]) of the population total of $\mathbf{z}$ defined as

$$\hat{Z} = \sum_{i=1}^{N} \frac{z_i}{\pi_i},$$

where N is the number of responses, z_i is the i-th response and π_i the corresponding inclusion probability. An optimal sampling design should minimise the expected variance of $\hat{Z}$, under an appropriate superpopulation model for $\mathbf{z}$. This can be defined via the expected variance of $(\hat{Z} - Z)$ under the sample s and the superpopulation model ξ

$$\hat{V}(\hat{Z} - Z) = E_\xi \{E_s[(\hat{Z} - Z)^2]\} - [E_\xi \{E_s(\hat{Z} - Z)\}]^2,$$

where E_ξ represents the expectation with respect to the superpopulation model and E_s denotes the expectation with respect to the sampling design. Assuming a linear superpopulation model, which is common for designing samples for scalar survey variables (Benedetti et al. [2017]), the model takes the form

$$\begin{cases} y_i = \mathbf{x}_i^t \beta + \varepsilon_i \\ E_\xi(\varepsilon_i) = 0 \\ Var_\xi(\varepsilon_i) = \sigma_i^2 \\ E_\xi(\varepsilon_i \varepsilon_j) = 0 \quad i \neq j \end{cases},$$

where i and j represent the spatial units, β is a vector of regression coefficients and ε_i is a random variable. Assuming that $Var_\xi(\varepsilon_i) = \sigma_i^2$ and $Cov_\xi(\varepsilon_i, \varepsilon_j) = \sigma_i \sigma_j \rho_{ij}$, where ρ_{ij} is the dependence parameter, the expected variance of the HT estimator for the population total of $\mathbf{z}$ given $\mathbf{X}$ under the above superpopulation model is

$$\hat{V}(\hat{Z}_{HT} - Z) = E_s\left[\left(\sum_{i \in s} \frac{X_i}{\pi_i} - \sum_{i \in U} X_i\right)^Z \beta\right]^2 + \sum_{i \in U} \sum_{j \in U} \sigma_i \sigma_j \rho_{ij} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j},$$

where U is a finite population, π_i and π_j are the sample inclusion probabilities for spatial units i and j and π_{ij} is their joint inclusion probability. The above expression is minimised when ρ_{ij}, the dependence parameter between each pair of units i and j is equal to zero. Assuming that the value of the dependence parameter decreases as the distance between the units increases, the sample selection strategy that can minimise the expression is one where the distance between the selected units is maximised, i.e. they are spatially well-spread.

Thus, spatially balanced samples are advantageous if the underlying spatial dependence is posi-

tive, as is most commonly the case in geographic data (Cressie [1993]), while clustered samples could be beneficial in case of a negative spatial dependence. Selecting a spatially balanced sample is also advantageous if the value of the observed population response $\mathbf{z}$ varies significantly in space, or if the population exhibits a cluster structure (Benedetti et al. [2017]). In such cases, the cluster can be characterised by one or a few samples and thus, further sampling would not provide additional benefits. Instead, sampling effort can be redirected to other, under-sampled regions.

The spatial structure of a groundwater CoPC plume is highly dependent on the characteristics of the aquifer medium (i.e. the stratum that confines the groundwater), the chemical properties of the CoPC and the groundwater flow. These features can lead to a variety of different spatial structures in terms of stratification and the presence of clusters. A common feature of these plumes however, is that the concentration of the CoPC in the groundwater tends to decrease with distance from the source and with time from the start of release, due to the effects of diffusion and advection. In other words, water soluble CoPCs diffuse into groundwater, which then carries it in the general direction of groundwater flow (advection). This leads to spatial and temporal correlation in the CoPC concentration data collected from the monitoring network over time. Without prior knowledge of the underlying processes that determine the spatial structure of the plume, it is a reasonable assumption that a spatially well-spread sample will result in the best population estimates. However, as outlined in Section 4.1, access to the groundwater is commonly only possible through fixed points called monitoring wells. The placement of the monitoring wells is already subject to optimisation that commonly takes into account the underlying hydrogeological processes in the form of process-based models, as well as other practical considerations such as accessibility. Therefore, the selection of spatially well-spread samples from these existing monitoring wells is constrained by their predetermined spatial arrangement. Given a network that has already been sampled in the past, the collected data could be analysed to obtain an estimate of the spatiotemporal trends characterising the plume. With such an estimate available, the balanced sampling design could be updated to assign higher weights to wells that provide more information on these trends. Hence, one of the questions that this thesis set out to answer and will address in more detail in Chapter 5 using empirical results, is whether a sampling design that sacrifices spatial balance, but accounts for the spatial structure of the CoPC plume via predictions from past observations, can result in better population estimates.

4.4 Spatially Balanced Sampling Designs

In spatially balanced sampling designs, spatial spread is typically maximized by avoiding the selection of neighbouring sampling locations. As explained by Benedetti et al. [2017], on the

most basic level the concept is to draw samples by solving a combinatorial optimisation problem that maximizes the distances between the selected sampling locations. However, the authors point out that this approach lacks the stochastic component and leads to fixed, unique samples, hence it cannot be used to generate randomised sampling designs. Spatially balanced sampling design algorithms have different ways of incorporating a stochastic component while balancing the maximisation of distances between selected samples. These will be described in more detail below.

A common feature of the probabilistic spatial sampling algorithms is that each geographical unit is assigned an inclusion weight, that determines the probability of it being selected as a sampling location at a given time point. These inclusion weights are determined mostly by the distances between the sampling points. Closer units normally receive smaller inclusion weights than units that are farther apart. Most spatially balanced sampling design algorithms allow for tuning the inclusion weights. This provides an opportunity for tuning the algorithm to prioritise certain sampling locations over others based on additional information about the target site or the underlying processes that determine the distribution of the population. For example, access to certain sampling locations can often be restricted, or there might be a need to concentrate sampling effort on a certain sub-region. Benedetti et al. [2017] provides an overview of the state of the art in spatially balanced sampling designs and compares several different methods to a baseline of a simple random sampling (SRS) design without replacement. The proposed spatially balanced algorithms in the literature include GRTS (Generalized Random Tessellation Stratified; Stevens and Olsen [2004]), CUBE (Deville and Tillé [2004]), DUST (Dependent Areal Units Sequential Technique; Arbia [1993]), CPS (Correlated Poisson Sampling; Bondesson and Thorburn [2008]), SCPS (Spatially Correlated Poisson Sampling; Grafström [2012]), DBSS (Doubly Balanced Spatial Sampling; Grafström and Tillé [2013]), LPM 1 & 2 (Local Pivotal Methods; Grafström et al. [2012]), BAS (Balanced Acceptance Sampling; Robertson et al. [2013]) and HIP (Halton Iterative Partitioning; Robertson et al. [2018]). The authors conclude that a well-spread sample can be advantageous in identifying the spatial structures of the population. Their results showed that if the data contain spatial heterogeneity, spatially balanced methods are able to capture this and provide a remarkable reduction in mean squared error compared to SRS. The following sections will provide a more detailed description on some of the existing spatially balanced design algorithms to demonstrate different approaches. The spatially balanced design techniques that will be looked at in more detail in this chapter were selected based on the following reasons. Firstly, they were chosen to represent different ways of approaching balanced sample selection, from partitioning the area (GRTS, HIP) to iterative processes that maximise the distance between samples (LPM) to using quasi-random number sequences to generate well-spread points (BAS, HIP). Secondly, they were selected to represent the methods that were found to be applied the most in the environmental science literature. Thirdly, the presented methods can be extended into three dimensions, allowing them to produce spatiotemporal designs. Lastly,

there are dedicated R (R Core Team [2020]) packages for the investigated methods that enable their application in comparison studies. Additionally, some of the methods are proposed by the same authors as improvements or modifications on previous designs. For example, CPS, SCPS and LPM were proposed by the same author, with LPM being the most recent with the most examples of application. Similarly, HIP and BAS were proposed by the same author with HIP having been built on the concept of BAS but improves its computational efficiency substantially. Ultimately, the spatially balanced designs that will be discussed in this chapter are GRTS, BAS, HIP and LPM.

4.4.1 GRTS

The GRTS design was proposed by Stevens and Olsen [2004] and it is the most influential and widely used approach (Benedetti et al. [2017]). It is based on mapping the two-dimensional space of the target area onto one dimension while preserving information on spatial proximity relationships. It uses a quadrant-recursive function to partition the two-dimensional target area into finer and finer nested quadrants. Each quadrant is assigned an ID number from 1-4 randomly and thus the nested cells can be referenced by a string of these ID numbers.

The finest subdivisions can be arranged numerically by these addresses in reverse hierarchical order, which ensures that a spatially balanced sample is selected when systematic sampling is applied. The addresses of potential sampling locations can then be mapped onto a one dimensional line, where each line segment corresponds to a cell in two-dimensional space. The line can then be divided into a number of equal length segments based on the number of desired samples and systematic sampling is applied to draw samples on the line. The inclusion weights of the potential sampling locations can be controlled by adjusting the length of the line-segment that represents their addresses in the one-dimensional space (see Figure 4.2). The resulting samples will be spatially balanced, since the sampling procedure will select samples located in different quadrants. The lengths of the line-segments i.e. their inclusion probabilities can be determined by an inclusion-probability function that represents different aspects of the potential sampling locations (e.g. importance, size, elevation, species distribution, etc.).

The GRTS design can also be used to draw sampling locations for one-dimensional resources, for example along river networks. According to Benedetti et al. [2017] GRTS is currently the most widely used spatially balanced sampling design algorithm. Its advantages include the intuitive handling of unequal inclusion probabilities and its ability to address common problems in environmental surveys. These problems include site inaccessibility, irregular spatial patterns of the population and missing data. The GRTS design allows for handling different levels of sample density within the same master-frame and dynamically adding additional sampling loca-

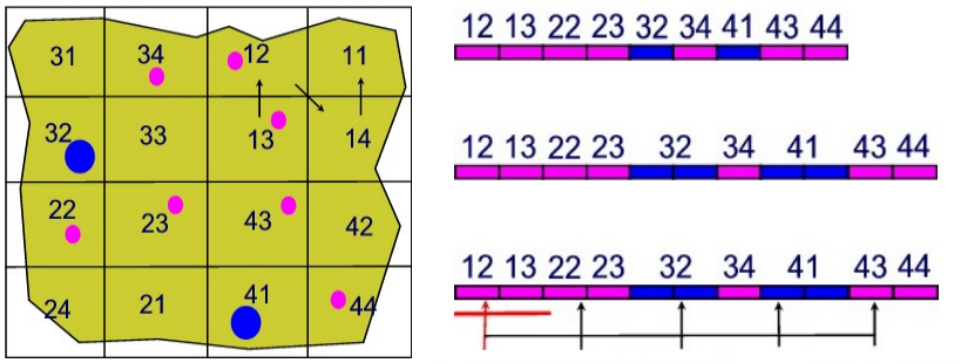


Figure 4.2: From the publicly available presentation by Olsen [2004]. The mapping of a two-dimensional surface of potential sampling sites onto one-dimension using quadrant-recursive partitioning. The inclusion probabilities of the blue sampling points are increased by doubling their corresponding line segment. Lastly, 5 samples are drawn systematically by dividing the line into equal-length segments.

tions from a spatially balanced over-sample as inaccessible locations are discovered (Robertson et al. [2018]). It also allows for the generation of panel-based sampling structures, in which sampling locations are divided into different panels that are sampled over time in an alternating fashion. The statistical advantages of applying GRTS on continuous surfaces are well demonstrated, but there is no empirical evidence on the gain in efficiency when sampling from discrete populations (Benedetti et al. [2017]). One drawback of the GRTS design when sampling discrete populations is that during the mapping of sampling locations from two-dimensional space onto one dimension, locations that are very close to each other in reality could end up far apart in one dimension. Moreover, the GRTS design is normally restricted to two-dimensional space (Robertson et al. [2013]), but can be used to generate spatiotemporal designs through the used of panels and rotating panels, which essentially produce repeated spatial GRTS designs.

To obtain estimates of population characteristics for a GRTS sample, the Horovitz-Thompson (HT), its continuous population analogue or the Yates-Grundy-Sen (YG) estimator can be used. For example, an estimate for the population total of a response z can be obtained by:

$$\hat{Z}_T = \sum_{s_i \in R} \frac{z(s_i)}{\pi(s_i)},$$

where, $\hat{Z}_T$ is the population total estimate, $z(s_j)$ is a sample from the i-th random cell and $\pi(s_i)$ is the corresponding inclusion intensity function that specifies the target number of samples from that cell. R is the domain of the population to be sampled and s_i is the i-th cell.

However, Stevens and Olsen [2004] explain that the variance estimator based on the approximations to the HT or YG estimators tends to be unstable for GRTS samples. This is because GRTS

achieves spatial balance by forcing the pairwise inclusion probability to tend towards 0 as the distance between the pair of points approaches 0. Moderate-sized samples tend to have at least a few pairs of points that are close together with correspondingly small inclusion probabilities. In the HT and YG variance estimators, these pairwise inclusion probabilities appear as divisors, which can consequently inflate the value of the variance estimate. Stevens and Olsen [2004], based on a previous work of theirs (Stevens and Olsen [2003]) proposed a contrast-based variance estimator for the GRTS design, that averages several contrasts over a local neighbourhood of each sample point and takes the following form:

$$\hat{V}_{NBH}(\hat{Z}_T) = \sum_{s_i \in R} \sum_{s_j \in D(s_i)} \omega_{ij} \left(\frac{z(s_j)}{\pi(s_j)} - \sum_{s_k \in D(s_i)} \omega_{ik} \frac{z(s_k)}{\pi(s_k)} \right)^2, \tag{4.1}$$

where $D(s_i)$ is a local neighbourhood of the i-th cell, s_i. ω_{ij} are weights that reflect the behaviour of the pairwise inclusion function for GRTS and are constrained so that $\sum_i \omega_{ij} = \sum_j \omega_{ij} = 1$.

The GRTS design has seen extensive application in environmental surveys. Kermorvant et al. [2019a] investigated the appearance of GRTS in the literature between 2001 and 2018 in the form of reports and publications. In total, the authors found 600 separate documents citing the GRTS design, mostly from different fields of environmental science such as ecology, environmental chemistry and environmental statistics. Only two publications appeared from other fields, one from economics and one about maintenance management quality assurance. The GRTS method is implemented in the R package Kincaid and Olsen [2016].

4.4.2 BAS

The BAS design was proposed by Robertson et al. [2013] and it is based on the quasi-random number sequence called the Halton sequence (Halton [1960]), which can be used to generate spatially well-spread random points. The BAS design can be used to draw spatially balanced samples from continuous or discrete populations in any number of dimensions.

The Halton sequence is based on van der Corput sequences, which are created by reversing the base p ($p \geq 2$) representation of the sequence of natural numbers (Robertson et al. [2013]). For example, consider the van der Corput sequence with base $p = 2$. The binary expansion of the natural number $k = 2$ is 10, and its radical inverse is $\phi_2(2) = 0.01$. $\phi_2(2)$ in the decimal systems gives $x_2 = 0.25 = 1/4$ which is the 2nd term in the van der Corput sequence with base 2. The sequence can be used to partition the unit interval with respect to the chosen base at values of $1/p, 1/p^2, 1/p^3$, etc. Thus, the partitioning generates numbers evenly on the unit interval. The

d-dimensional Halton sequence is based on d van der Corput sequences (one sequence for each dimension), where each base of the van der Corput sequences are pairwise co-prime (Robertson et al. [2013]). The Halton sequence on $[0,1]^d$ is

$$\mathbf{x}_k = (\phi_{p_1}(k), \phi_{p_2}(k), ..., \phi_{p_d}(k)), \quad k = 0, 1, 2, ...$$

where $\phi_{p_j}(k)$ is the van der Corput sequence of base p_j, where p_j is the j-th prime number. This Halton sequence, however, is deterministic. By selecting a contiguous Halton subsequence starting from a random seed, stochasticity can be introduced without affecting the even distribution of numbers. Robertson et al. [2013] defines the random start Halton sequence as follows. Let $\mathbf{u} = (u_1, ..., u_d)$ be a vector whose members are independent random integers with uniform distribution on $[0, \mathbb{U}]$, where $\mathbb{U}$ is any sufficiently large integer. The sequence $\{\mathbf{x}_k\}_{k=0}^{\infty}$ in $[0,1]^d$ defined by

$$\mathbf{x}_k = (\phi_{p_1}(u_1 + k), \phi_{p_2}(u_2 + k), ..., \phi_{p_d}(u_d + k)),$$

is called a random-start Halton sequence. The resulting sequence is the Halton sequence after skipping the first u_1 terms in the first dimension, the first u_2 terms in the second dimension and so on. In this implementation, the uniformity of the sequence is preserved, since the subsequences that define the dimensions are still uniform (Robertson et al. [2013]).

To implement the BAS design, the user needs to specify the dimensions of a hyperrectangle that encompasses the study area. In two-dimensional space this is a rectangle that encloses a geographic region. For continuous populations, a random-start Halton sequence is used to generate points within the hyperrectangle until the required number of samples are drawn from the target area. For discrete populations, the study area is partitioned into cells until each discrete unit is captured in a single cell. This can be performed using Voronoi type partitioning. Then, a unit is selected if the Halton point falls within the cell that contains the unit. The procedure is continued until the required number of samples are captured.

The exact inclusion probabilities for discrete sampling units can also be calculated in the above described implementation. For a given BAS design, there are $\mathbb{U}^d$ possible random-start Halton sequences. Each sample corresponds to one of these being selected randomly by the vector $\mathbf{u}$. Thus, for any value of $\mathbb{U}$, the inclusion probability for unit i is the fraction of sequences that select unit i. As Robertson et al. [2013] defines, this is given by:

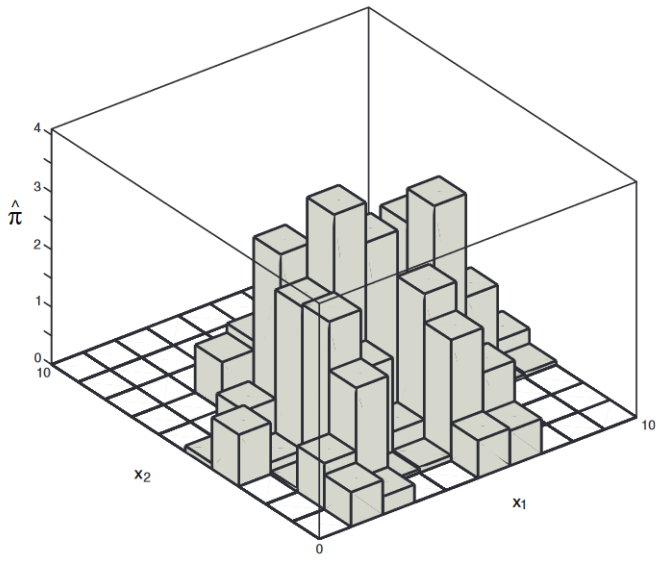


Figure 4.3: From Robertson et al. [2013] in accordance with JSTOR Terms & Conditions. Example of a two-dimensional discrete population with a specified target inclusion density function. Generating random-start Halton points in the box and accepting points that fall under the shaded volume is equivalent to selecting spatially well-balanced points over the study area with respect to a target inclusion density function.

$$\pi_i = \frac{1}{\mathbb{U}^d} \sum_{j=1}^{\mathbb{U}^d} \mathbf{I}(\{\mathbf{x}_k^{(j)}\}_{k=1}^{v}), \tag{4.2}$$

where $\mathbf{I}(\cdot) = 1$ if unit i is selected by the jth random-start Halton sequence, $\mathbf{x}^{(j)}$, and 0 otherwise. According to Robertson et al. [2013], using equation 4.2 can be computationally expensive at large values of $\mathbb{U}$, but the calculation is highly parallelizable.

A target inclusion probability (density) function can also be provided to control the prioritization of sub-regions or units within the target area. This will result in an unequal probability BAS design. A density function can be incorporated into the BAS design by adding an extra dimension to the sample domain, generating random-start Halton sequence points and accepting them as sampling locations if they fall within the volume or sub-region defined by the density function (see Figure 4.3). The relative error between the target inclusion probabilities and the actual inclusion probabilities can also be calculated by:

$$e_i = \frac{|\pi_i - \hat{\pi}_i|}{\pi_i}, \tag{4.3}$$

where π_i is given by equation 4.2 with $\mathbb{U}^{d+1}$ random-start Halton sequences and $\hat{\pi}_i$ is the target inclusion probability of unit i (Robertson et al. [2013]).

Since the BAS design produces samples with specified first order inclusion probabilities given by equation 4.2, the HT estimator or its continuous analogue can be used to obtain population characteristic estimates (Robertson et al. [2013]). The population total estimate for a response z can be obtained by

$$\hat{Z} = \sum_{i=1}^{N} \frac{z_i}{\pi_i} \mathbf{I}(i), \tag{4.4}$$

where $\mathbf{I}(i) = 1$ if unit i is in the sample and 0 otherwise. The variance of the HT estimator for a given sample size can be defined as

$$V(\hat{Z}) = -\frac{1}{2} \sum_{i,j(i \neq j)} (\pi_{ij} - \pi_i \pi_j) \left(\frac{z_i}{\pi_i} - \frac{z_j}{\pi_j} \right)^2,$$

and can be estimated from a sample using the YG estimator

$$\hat{V}(\hat{Z}) = -\frac{1}{2} \sum_{i,j(i \neq j)} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{z_i}{\pi_i} - \frac{z_j}{\pi_j} \right)^2 \mathbf{I}(ij), \tag{4.5}$$

where π_{ij} is the second order inclusion probability and $\mathbf{I}(ij) = 1$ if both unit i and unit j are included in the sample and 0 otherwise. As Robertson et al. [2013] states, the second order inclusion probabilities can be calculated using equation 4.2 by setting $\mathbf{I}(\cdot) = 1$ if both i and j are selected by the jth random-start Halton sequence and $\mathbf{I}(\cdot) = 0$ otherwise. However, as mentioned above in the case of GRTS, the YG estimator can become unstable if there are units close to each other in space with very low values of second order inclusion probabilities. This is commonly the case with spatially balanced sampling designs. Hence, the authors suggest that the local mean variance estimator derived for GRTS (see equation 4.1) can be used instead, which in this case is given by:

$$\hat{V}_{NBH}(\hat{Z}_T) = \sum_{i=1}^{N} \sum_{j \in D_i} \omega_{ij} \left(\frac{z_i}{\pi_i} - \bar{z}_{D_i} \right)^2 \mathbf{I}(ij), \tag{4.6}$$

where D_i is a neighbourhood to unit i that contains at least four observed units and $\bar{z}_{D_i}$ is the total in the neighbourhood of unit i. The weights, ω_{ij}, decrease as the distance between i and j increases. Moreover, it satisfies the condition that $\sum_j \omega_{ij} = 1$.

The BAS design has two drawbacks when sampling discrete populations. First, the required

acceptance/rejection sampling step might not achieve the targeted inclusion probabilities, which inflates the variance estimators Robertson et al. [2018]. Second, as mentioned above, BAS partitions the target area into cells until each sample unit is contained within a cell. However, if the population is large or lacks regular grid structure, the BAS partitioning becomes inefficient or computationally prohibitive because of the large number of required cells.

There are a few examples in the literature of the BAS design being applied for environmental sampling problems. Abi et al. [2017] compared the BAS design to a systematic sampling design in a case study of a crab species from an intertidal marine zone in Qatar with favourable results. Koski and Eidsvik [2024] compared SRS, GRTS and BAS for optimising the cost efficiency of a long-term monitoring program of manila clams in the Archacon Bay in France. The study yielded equally good results for GRTS and BAS. Dam-Bates et al. [2018] used the BAS design to construct a master frame to help and increase coordination between agencies monitoring biodiversity in New Zealand.

4.4.3 HIP

HIP was proposed by Robertson et al. [2018] as an alternative to or an improvement on their previous BAS design. According to the authors, HIP shares the desirable properties of BAS while eliminating its drawbacks when sampling discrete target populations. Rather than using the points generated by the Halton sequence, HIP exploits a structural, quasi-periodic property of the sequence to partition the target area into nested boxes, which are subsequently sampled in a specific order to obtain a spatially balanced sample. The HIP design has a low computational complexity of $O\,(N\,log\,N)$. It can be applied in any number of dimensions for continuous or discrete populations and achieves target inclusion probabilities. It can also be used to create spatially balanced over-samples, which can be useful if inaccessible sampling locations are discovered and need to be replaced.

The HIP design uses Halton indices to partition the target area into boxes. According to Robertson et al. [2018] the Halton index of a point can be derived in the following way. The *i*th coordinate of the *k*th point in a Halton sequence is given by:

$$x_k^{(i)} = \phi_{b_i}(k) = \sum_{j=0}^{\infty} \left\{ \left\lfloor \frac{k}{b_i^j} \right\rfloor \mod b_i \right\} \frac{1}{b_i^{j+1}}, \tag{4.7}$$

where $\lfloor x \rfloor$ is the floor function that gives the largest integer less than or equal to x. For instance, the 4th point in the three-dimensional Halton sequence is then:

$$\mathbf{x}_4 = (\phi_2(4), \phi_3(4), \phi_5(4)) = \left(\frac{1}{8}, \frac{4}{9}, \frac{4}{5}\right).$$

The Halton index is the subscript of each point, which denotes the integer that is mapped to a point in $[0,1)^d$ using equation 4.7. If $B = \prod_{i=1}^{d} b_i^{J_i}$, where J_i is any non-negative integer, it can be shown that B consecutive points from a Halton sequence with bases b_i will have exactly one point in each Halton box defined by:

$$\prod_{i=1}^{d}\left[m_i b_i^{-J_i}, (m_i+1)b_i^{-J_i}\right), \tag{4.8}$$

where m_i is an integer satisfying $0 \leq m_i \leq b_i^{J_i}$, for all $i = 1,2,...,d$. If $\mathbf{x}_k$ falls within a particular Halton box, then k must take a specific mod B value from the set $0,1,...,B-1$ (Robertson et al. [2018]). Therefore, points $\mathbf{x}_k$, $\mathbf{x}_{k+B}$, $\mathbf{x}_{k+2B}$, ..., etc. must be contained within the same Halton box. Thus, each box is associated with a Halton index (mod B). This shows that the Halton sequence is quasi-periodic in this way with period B. According to Robertson et al. [2018], the Halton index k of a box with lower bounds $l_1,...,l_d$ can be obtained by solving the system of d congruences

$$k = a_i \pmod{b_i^{J_i}},$$

where

$$a_i = \sum_{j=1}^{J_i}\left\{\lfloor l_i b_i^J \rfloor \mod b_i\right\} b_i^{j-1}$$

and $i = 1,2,...,d$. For example, the Halton index for $[1/4,1/2)\times[1/3,4/9)$ with $B = 36$ requires solving

$$k = 2 \pmod 4$$
$$k = 1 \pmod 9,$$

which gives $k = 10 \pmod{36}$. Another aspect of this partitioning based on Halton indices is derived from equation 4.8, which indicates that small Halton boxes are nested in larger parent boxes. The indices of the nested boxes must be congruent mod B, where B is the number of

parent boxes, because each box has a specific mod B index (Robertson et al. [2018]).

As Robertson et al. [2018] explains, the HIP design then draws a sample of size n from a target population by iteratively partitioning the area into $B = \prod_{i=1}^{d} b_i^{J_i} \geq n$ boxes, that have the same nested structure as the Halton boxes. The size of each box is designed to keep the inclusion probability the same, i.e. boxes will tend to be small in areas with high inclusion probabilities and larger otherwise. Each box is indexed by the Halton index of its corresponding Halton box and the sample is obtained by drawing n points from consecutively numbered boxes. The nested partitioning can be generated for continuous and discrete populations with equal or unequal inclusion probabilities in any number of dimensions.

Robertson et al. [2018] derives the method for continuous resources the following way. The indices of the boxes to be sampled are determined before partitioning the area:

$$S_k = \{k, (k+1) \bmod B, ..., (k+n-1) \bmod B\},$$

where k is a random integer from the set $0, 1, ..., B-1$. The partition is then created iteratively to find boxes with the indices specified in S_k. Initially, $[0,1)^2$ is split into two boxes along x_1 at $0 < q < 1$

$$H_0 = [0,q) \times [0,1) \quad \text{and} \quad H_1 = [q,1) \times [0,1), \tag{4.9}$$

such that the inclusion densities in the boxes are equal. These boxes are then partitioned using x_2 to give six boxes of the form

$$\begin{aligned} H_0 &= [0,q) \times [0,r) \quad H_3 = [q,1) \times [0,r') \\ H_4 &= [0,q) \times [r,s) \quad H_1 = [q,1) \times [r',s') \\ H_2 &= [0,q) \times [s,1) \quad H_5 = [q,1) \times [s',1), \end{aligned} \tag{4.10}$$

where H_k denotes the Halton box with index k. Each index is calculated using its corresponding Halton box by setting $q = 1/2, r = r' = 1/3$ and $s = s' = 2/3$. The values $r < s$ and $r' < s'$ are chosen so that the inclusion probability in each box is constant

$$\int_{H_k} \pi(\mathbf{x}) d\mathbf{x} = \frac{n}{6},$$

for $k = 0, 1, ..., 5$. The method is then repeated on each box in equation 4.10. The iterative partitioning continues until the number of boxes is greater than n. Afterwards, only those boxes are split that contain sub-boxes with indices in S_k. The sampling design is obtained by sampling the boxes indexed by S_k. For each $k \in S_k$, one point is selected from box H_k using acceptance/rejection sampling with the inclusion probability function

$$f_k(\mathbf{x}) = \begin{cases} \frac{B}{n}\pi(\mathbf{x}) & \text{if } \mathbf{x} \in H_k \\ 0 & \text{otherwise.} \end{cases}$$

If the population domain is $\omega = [0,1)^2$ and $\pi(\mathbf{x})$ is the uniform inclusion probability function, a HIP sample will be similar to n consecutive points from a Halton sequence. Otherwise, more samples will be drawn from regions with higher inclusion probabilities.

For discrete populations (i.e. point resources) such as groundwater monitoring well networks, a sample of size n is drawn from N points in $[0,1)^2$, where the ith point has an inclusion probability $0 < \pi_i < 1$ such that $\sum_{i=1}^{N} \pi_i = n$. As Robertson et al. [2018] discusses, the partitioning happens similarly to the continuous population case, but rather than keeping the inclusion probabilities in the boxes constant, in this case they are optimised to be as similar to each other as possible. First, the points are partitioned into two boxes H_0 and H_1 using the expression 4.9 where q minimises

$$\left(\sum_{i:\mathbf{x}_i \in H_0} \pi_i - \sum_{j:\mathbf{x}_j \in H_1} \pi_j \right)^2.$$

These boxes are then partitioned with splits along x_2 to give six boxes (as in expression 4.10) such that

$$\sum_{k \in 0,...,5} \left(\sum_{i:\mathbf{x}_i \in H_k} \pi_i - n/6 \right)^2$$

is minimised. The method is repeated on each nested box to define a sequence of B boxes so the total inclusion probabilities within the boxes are as similar as possible and ≤ 1. The optimal number of boxes is determined by considering candidate values and selecting the one with the best average spatial balance. Each box H_k is then assigned an inclusion interval

$$I_k = \left[\sum_{i=0}^{k-1} d_i, \sum_{i=0}^{k} d_i \right) \quad (k > 0),$$

where $I_0 = [0, d_0)$ and

$$d_k = \sum_{i:\mathbf{x}_i \in H_k} \pi_i,$$

for $k = 0, 1, ..., B-1$. The boxes whose intervals contain a point from

$$\{(s + \lambda\alpha) - n\lfloor (s + \lambda\alpha)/n \rfloor : \lambda = 0, 1, ..., n-1\},$$

have one point drawn from them using selection probabilities δ_i, where s is randomly chosen from $\in [0, n), \alpha = \max\{d_k\}$ and

$$\delta_i = \pi_i / d_k : \mathbf{x}_i \in H_k.$$

Since $\alpha \geq d_k$ for all k and $B \geq n$, n different boxes have a point drawn from them. The inclusion probability of $\mathbf{x}_i \in H_k$ is $d_k \delta_i = \pi_i$.

According to Robertson et al. [2018], for a point resource, $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^d$ the HT estimator can be used to estimate the population total for a response z using

$$\hat{Z} = \sum_{i \in S} \frac{y_i}{\pi_i},$$

where $S \subset \{1, 2, ..., N\}$ is a sample and π_i is the inclusion probability of the ith point. The variance of $\hat{Z}$ could be estimated using the YG estimator, but similarly as in the case with the BAS design, this estimator tends to be unstable for spatially balanced sampling designs due to the presence of second order inclusion probabilities with values approaching zero. Therefore, the local mean variance estimator can be used for HIP designs instead, which is given by

$$\hat{V}_{NBH}(\hat{Z}) = \sum_{i \in S} \sum_{j \in D_i} \omega_{ij} \left(\frac{y_j}{\pi_j} - \hat{Z} D_i \right)^2, \tag{4.11}$$

where D_i is a neighbourhood containing at least four nearest neighbours to the ith point, $\hat{Z}_{D_i}$ is an estimate of the population total on D_i and ω_{ij} are weights.

At the time of writing, there are no published examples of the application of the HIP sampling design in practice. There is however an implementation of the method in the statistical program-

ming language R (R Core Team [2020]). Moreover, Kermorvant et al. [2019b] compared several spatially balanced sampling designs including HIP for environmental surveys. The author of the BAS and the HIP methods also developed an R package which includes an implementation of both. The package is currently available from GitHub[2].

4.4.4 The Local Pivotal Methods

The Local Pivotal Methods (LPM1 and LPM2), introduced by Grafström et al. [2012] are spatially balanced sampling design methods that rely on step-wise updating rules to avoid the selection of neighbouring units. They can be used to draw samples from continuous or point resources in any number of dimensions and allow for equal or unequal inclusion probabilities.

LPM is based on the pivotal method (PM) introduced by Deville and Tillé [1998], which is a sampling technique where the inclusion probabilities for two units are updated so that the sampling outcome is decided for at least one of those units. Given a population of N units, a sample can be obtained in N updating steps. The updating is described by Grafström et al. [2012] as follows. π_i' is the possibly updated inclusion probability of unit i. Unit i is finished if $\pi_i' = 0$ or $\pi_i' = 1$, and once it is finished it will not be chosen again. Let i and j be two random units with inclusion probabilities π_i and π_j. The reduced vector (π_i, π_j) is updated by the following rule. If $\pi_i + \pi_j < 1$, then

$$(\pi_i', \pi_j') = \begin{cases} (0, \pi_i + \pi_j) & \text{with probability } \frac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) & \text{with probability } \frac{\pi_i}{\pi_i + \pi_j} \end{cases},$$

and if $\pi_i + \pi_j \geq 1$, then

$$(\pi_i', \pi_j') = \begin{cases} (1, \pi_i + \pi_j - 1) & \text{with probability } \frac{1-\pi_j}{2-\pi_i-\pi_j} \\ (\pi_i + \pi_j - 1, 1) & \text{with probability } \frac{1-\pi_i}{2-\pi_i-\pi_j} \end{cases}.$$

The updating procedure is repeated with the updated inclusion probabilities for randomly chosen units until all units are finished. The LPM design improves the spatial balance of the obtained pivotal method sample by locally keeping the sum of the updated inclusion probabilities as constant as possible. Hence, the main difference is that the two units chosen for the updating rule will be nearby units.

[2]https://github.com/tmcd82070/SDraw

The LPM1 design starts by randomly selecting a unit i, then its nearest neighbour, unit j. If multiple units have the same distance to unit i, then one of them is randomly selected with equal probability. If unit i is also unit j's nearest neighbour, then their inclusion probabilities are updated according to the PM updating rule, otherwise the first step is repeated an another random unit is selected. This procedure continues until all units are finished. In the LPM2 design, the inclusion probabilities of units i and j are updated regardless of whether they are mutually nearest neighbours.

Thus, LPM2 is a simpler, faster implementation of LPM1, which is spatially more balanced but has a has a larger computational cost. The number of computations for LPM1 is in the worst case proportional to N^3 and in the best case proportional to N^2, whereas for LPM2 it is always proportional to N^2. Lisic and Cruze [2016] further reduced computational costs for LPM2 in their implementation which uses a k-dimensional tree-based nearest neighbour search. The number of computations in their implementation is proportional to $Nlog(N)$. The LPM design also allows for using unequal inclusion weights.

Given a response z with value z_i for unit i, the HT estimator can be used on an LPM sample to estimate the population total:

$$\hat{Z} = \sum_{i \in U} \frac{z_i}{\pi_i} I_i,$$

where U is a spatial population of N units, π_i is the inclusion probability of the ith unit and I_i is the inclusion indicator for unit i. $I_i = 1$ if the unit is contained in the sample and $I_i = 0$ otherwise. The variance for the HT estimator for a fixed sample size is given by

$$V(\hat{Z}) = -\frac{1}{2} \sum_{i,j \in U} (\pi_{ij} - \pi_i \pi_j) \left(\frac{z_i}{\pi_i} - \frac{z_j}{\pi_j} \right)^2,$$

where π_{ij} is the second-order inclusion probability, indicating the probability that both units i and j are included in the sample. This variance can be estimated using the YG estimator. However as explained in previous sections and pointed out by Grafström et al. [2012], due to the presence of near-zero or zero valued second-order inclusion probabilities in spatially balanced samples, the YG estimator (see equation 4.5) can be unstable, and result in a variance estimate of zero despite the true variance being large. This situation can arise for example if the units to be sampled are clustered in small groups within which the inclusion probabilities sum to one. In these cases one unit is selected from each group independently from other groups, therefore, $\pi_{ij} = \pi_i \pi_j$.

Grafström et al. [2012] introduces different variance estimators instead of the YG estimator. The first one being an estimator that works under the assumption that the sample was taken using independent random sampling (IRS). This estimator is given by

$$\hat{V}_{IRS}(\hat{Z}) = \frac{n}{n-1}\sum_{i\in s}\left(\frac{z_i}{\pi_i} - \frac{1}{n}\sum_{j\in s}\frac{z_j}{\pi_j}\right)^2, \tag{4.12}$$

where s is the sample and n is the sample size. Another option is to use the Hajek-Rosen estimator originally proposed for the Conditional Poisson (CP; Hájek and Dupač [1981]) sampling:

$$\hat{V}_{HR}(\hat{Z}) = \frac{n}{n-1}\sum_{i\in s}(1-\pi_i)\left(\frac{z_i}{\pi_i} - \frac{\sum_{j\in s} z_j(1-\pi_j)/\pi_j}{\sum_{j\in s}(1-\pi_j)}\right)^2. \tag{4.13}$$

However, Grafström et al. [2012] suggest using the local mean variance estimator, also used successfully for the GRTS, BAS and HIP designs (see equations 4.1, 4.6, 4.11). Similarly to GRTS and other spatially balanced designs, the LPM designs also produce spatially well-spread samples, therefore the local mean variance estimator is expected to perform well. The local mean variance estimator is given by:

$$\hat{V}_{NBH}(\hat{Z}) = \sum_{i\in s}\sum_{j\in D_i}\omega_{ij}\left(\frac{z_j}{\pi_j} - \bar{z}_{D_i}\right)^2, \tag{4.14}$$

where D_i is a neighbourhood to unit i containing at least four units, $\bar{z}_{D_i}$ is a neighbourhood total and ω_{ij} are weights that decrease as the distance between units i and j increases, and they satisfy $\sum_j \omega_{ij} = 1$.

In a later publication Grafström and Schelin [2014] suggested a new variance estimator for the LPM design,

$$\hat{V}_{LPM}(\hat{Z}) = \frac{1}{2}\sum_{i\in S}\left(\frac{z_i}{\pi_i} - \frac{z_{j_i}}{\pi_{j_i}}\right)^2, \tag{4.15}$$

where, $j_i \in S$ is the index of the nearest neighbour to the ith point in the sample. This estimator does not rely on the definition of neighbourhoods or the computation of weights, therefore, it is more efficient than the local mean variance (NBH) estimator.

Based on the analysis by Benedetti et al. [2017], the LPM designs perform well in terms of

estimating the sample mean compared to simple random sampling and providing highly spatially balanced samples. The authors recommend using LPM particularly when strong spatial dependence is believed to be present in the data. In a simulation study, Robertson et al. [2018] also showed that LPM2 provided the best spatial balance when compared to GRTS and HIP. The LPM design has seen some application in environmental surveys, especially in the field of forest management. For example, Räty et al. [2019] used LPM sampling designs to investigate the effect of cluster configurations in sampling campaigns for the planning of a national forest inventory in Finland. Saad et al. [2016] and later Räty et al. [2020] compared LPM to systematic sampling methods for forest management planning with favourable results. Roberge et al. [2017] evaluated a new method for forest damage inventories where auxiliary data are used for sample selection with the LPM. Notably, Lisic, Jonathan and Grafström, Anton [2018] created an R package that includes an implementation of LPM1 and LPM2 with the computationally efficient k-dimensional tree nearest neighbour search method introduced by (Lisic and Cruze [2016]), which helped the more widespread application of the LPM framework in environmental sampling problems.

4.5 Comparison of Spatially Balanced Sampling Designs in the Literature

Benedetti et al. [2017] conducted a simulation study to evaluate the strengths and weaknesses of a number of spatially balanced sampling designs against a baseline of SRS. The compared designs were GRTS, CUBE, DUST, SCPS, LPM and DBSS. The authors used the statistical computing language R (R Core Team [2020]) with packages sampling (Tillé and Matei [2015]), survey (Lumley [2014]), spsurvey (Kincaid and Olsen [2016]) and BalancedSampling (Grafström and Lisic [2016]) to conduct the study on two different datasets, both containing significant geographical trends. The authors took samples using the different spatially balanced sampling design techniques and used the samples to estimate population totals using the HT estimator. The sampling procedure was repeated $10,000$ times for different sample sizes. The metrics used to compare the designs were the relative efficiency of the sample mean and the mean of the spatial balance indices. The former is given by the ratio between the mean squared error (MSE) achieved with the spatially balanced sample and the SRS sample: MSE/MSE_{SRS}. The latter is the mean of a measure that indicates the spatial balance of the sample, and is based on Voronoi polygons. For a set $\mathbf{P}$ of points in n-dimensional Euclidean space, the Voronoi partition $V(\mathbf{P})$ of the space is such that each point in $\mathbf{P}$ has a region which is closer to it than any other point in the set (Ito [2015]). The regions associated with the points in this way are referred to as Voronoi polygons or cells. The Voronoi-based spatial balance measure can be defined as:

$$v_k = \sum_{i \in VP(k)} \pi_i, \tag{4.16}$$

where v_k is the sum of the first-order inclusion probabilities of the units of the population in the k-th Voronoi polygon. For any sample unit the expected value of v_k, $E(v_k)$ will be one. For a spatially balanced sample, all of the v_k values should be close to 1. Thus, the variance $Var(v_k)$ is an indicator of the spatial balance of a sample, where a lower value indicates a spatially more balanced sample.

Prentius and Grafström [2024] have recently proposed a new measure, the local spatial balance measure to assess the spatial balance of a sample. This measure is based on the balancing equation which captures the balance between sample units and their neighbours. The evaluation of the new measure against commonly used ones such as the Voronoi spatial balance presented in equation 4.16, indicated that it can differentiate between sampling designs with similar Voronoi partitions, since it measures balance within the polygons. Thus the measure's efficacy was comparable to the Voronoi approach with some additional benefits. It can effectively be used to compare different sampling designs before committing to any particular campaign.

The results of the simulation study for the GRTS and LPM designs showed that LPM1 and LPM2, the two implementations of the LPM design, produced identical outputs. They also produced spatially more well-spread samples than the GRTS design, which resulted in larger improvements to the MSE values on average. Based on the results, the authors suggested that samples with better spatial distribution obtained by the distance-based methods, such as LPM, increase the efficiency of the sampling designs. They also propose some recommendations regarding best practices for spatial sampling. In case of the presence of a strong spatial dependence in the data, the authors recommend using methods based on the distance between sampling units, such as LPM.

Grafström et al. [2012] also conducted five simulation studies using different types of datasets to compare the two LPM implementations and GRTS. The authors compared samples drawn using the different design methods using a measure of spatial balance and different variance estimators. In the first simulation study, the results showed that the LPM methods produced the most spatially balanced samples. LPM 1 slightly outperformed LPM2 which was explained by the difference between the two implementations given in Section 4.4.4.

The second simulation study performed by Grafström et al. [2012] aimed to compare the LPM to the GRTS design using the variance of the HT estimator. The variance of the HT estimator was computed by

$$\hat{V}_{Sim}(\hat{Z}) = \frac{1}{m}\sum_{s}(\hat{Z}(s) - Z)^2, \tag{4.17}$$

where, m is the number of simulated samples, s. The variance estimated by this equation was compared to the IRS estimator (see equation 4.12), the HR estimator (see equation 4.13) and the NBH estimator (see equation 4.14) as well as the exact variance computed throughout the simulation under SRS. The results indicated that the LPM and the GRTS samples show similar behaviour and produce similar variance estimates with all estimators, including the NBH estimator which is well-established for GRTS samples.

Grafström et al. [2012] performed three more simulation studies similar to the one described above using different datasets. The third simulation had the same objective and used the same estimators as the second simulation, but was conducted on a small, spatially stratified, synthetic population. Due to the stratified nature of the dataset, a stratified SRS sampling design was also used for comparison. The results showed that the LPM designs performed better on this population than the stratified SRS sample, due to the spatial trends within the three strata, and the fact that SRS did not spread the samples well enough within the clusters. The LPM designs also performed better than the GRTS design, due to the inability of the GRTS design to recognise and account for the stratification of the population. All estimators, including the NBH estimator produced similar estimates for the LPM designs and GRTS. All variance estimates were conservative (overestimating the variance) due to the small population and sample sizes compared to the strong spatial trends.

In the fourth simulation study the target variable was from a synthetic continuous population. The results of this study showed a similar trend as in the other simulations. The LPM designs performed better than the GRTS and SRS designs in terms of $\hat{V}_{Sim}$. The LPM and GRTS designs also produced similar variance estimates with all of the estimators.

The results of the fifth simulation study also substantiated that the mean variance estimates were lower for the LPM designs than for GRTS. As the sample size was increased, the difference between the sampling designs naturally decreased as well. The variance estimates for SRS were substantially higher, due to the fact that it does not balance the spatial distribution of the samples. The NBH estimator worked as well for the LPM designs as for the GRTS designs.

Another comparison of spatially balanced design algorithms comes from Robertson et al. [2013], who ran a simulation study to analyse the performance of their BAS design against GRTS, LPM1, LPM2, CP, SRS and stratified SRS. In their analysis, the authors compared the variance of the HT estimator (see equation 4.4) for two populations. The variance of the HT estimator was estimated using equation 4.17. The first population the authors used came from the fourth

simulation study conducted by Grafström et al. [2012]. The results showed negligible difference between BAS, LPM1 and LPM2. They all performed as well as, or better than GRTS. Their simulated variances were similar or lower than the exact variance for the stratified SRS designs. However, the mean local variance estimator showed a large bias for this problem, and consistently overestimated the variance. The second simulation study yielded similar results, with all spatially balanced sampling designs performing well. The mean local variance estimator also yielded good estimates of the simulated variances.

Robertson et al. [2018] also conducted simulation studies to compare the HIP design with other spatially balanced sampling algorithms such as LPM2 and GRTS. First they compared the spatial balance of samples drawn by the different methods using Voronoi polygons, as described in equation 4.16. The measure of spatial balance was expressed as the mean squared error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(v_i - 1)^2, \tag{4.18}$$

where n is the number of samples, and v_i is the sum of first-order inclusion probabilities of the units in the ith Voronoi polygon. The results of the analysis showed that regardless of sample size, GRTS, LPM2 and HIP achieved a higher degree of spatial balance than SRS. LPM2 and HIP produced more balanced samples than GRTS, but LPM2 produced the most balanced samples consistently. The authors also pointed out that LPM2 is extremely effective with smaller populations. Moreover, the LPM2 sampling algorithm with k-dimensional trees implementation had the same computational complexity as HIP, $O(N \log N)$.

Robertson et al. [2018] also performed a simulation study to compare the spatially balanced sampling designs GRTS, LPM2, HIP with SRS in terms of using the obtained samples to estimate the variance of the HT estimator. The results showed that the spatially balanced sampling designs consistently outperformed SRS in estimating the population total. The spatially balanced designs were most beneficial for the population with the strongest spatial trend. HIP and LPM2 performed better than GRTS, but there was negligible difference between HIP and LPM2. LPM2 produced samples with slightly better spatial balance than HIP but this did not translate to better precision in estimating the variance of the population total. The NBH estimator mostly produced conservative estimates, while the LPM estimator used for the HIP and LPM2 samples performed better, but also produced conservative estimates for smaller sample sizes.

In summary, the comparative studies show that the LPM designs, including the computationally more efficient LPM2 with the k-dimensional trees implementation, provides a constant, good performance in selecting spatially well balanced random samples, and thereby resulting in better population estimates. Additionally, as established in Section 4.4.4, it can be used to draw bal-

anced samples in more than two dimensions, and incorporate both equal and unequal inclusion probabilities. These factors make it an ideal candidate for creating spatiotemporal groundwater contamination sampling designs with the option to incorporate historic information in the form of an inclusion density function. In the following section, an additional test will be presented that investigates the spatial balance values of spatiotemporal samples drawn using the LPM1 and LPM2 sampling algorithms.

4.6 Comparison of Spatially Balanced Sampling Designs in Three Dimensions

The comparative studies found in the literature only considered the 2-dimensional, spatial application of balanced sampling designs. To substantiate the results of these studies, and to determine the most appropriate spatially balanced design algorithm to apply in a spatiotemporal setting, an additional outline comparison was carried out between SRS (see Section 4.2), LPM1 (see Section 4.4.4), LPM2 (see Section 4.4.4) and GRTS (see Section 4.4.1). The LPMs were selected based on the results of the comparative studies, that found that they consistently performed better than other approaches at selecting spatially balanced samples. Additionally, there was a lot of information found in the literature on assessing LPMs against other approaches. GRTS was selected for the comparison because it is the most widely used spatially balanced design in the literature.

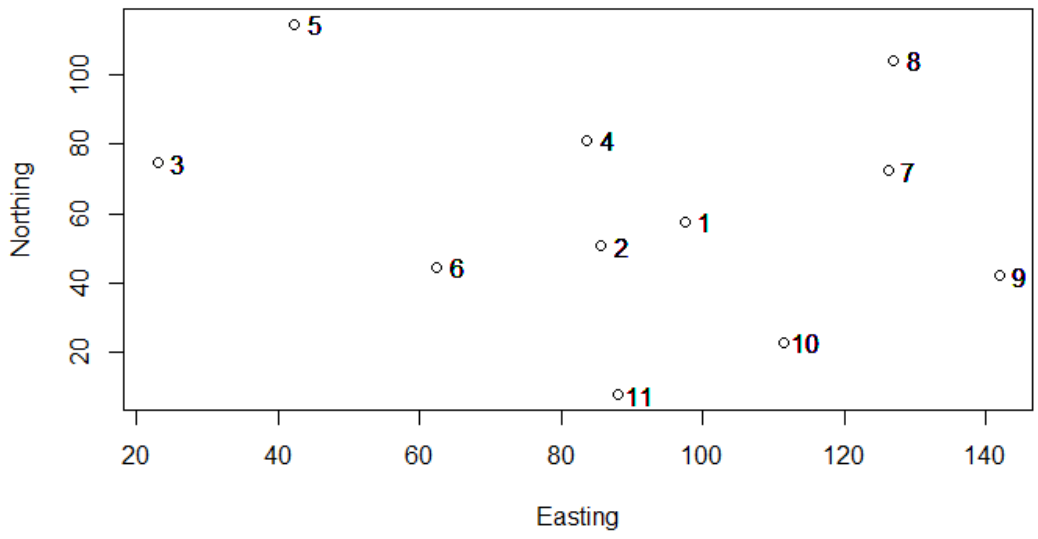
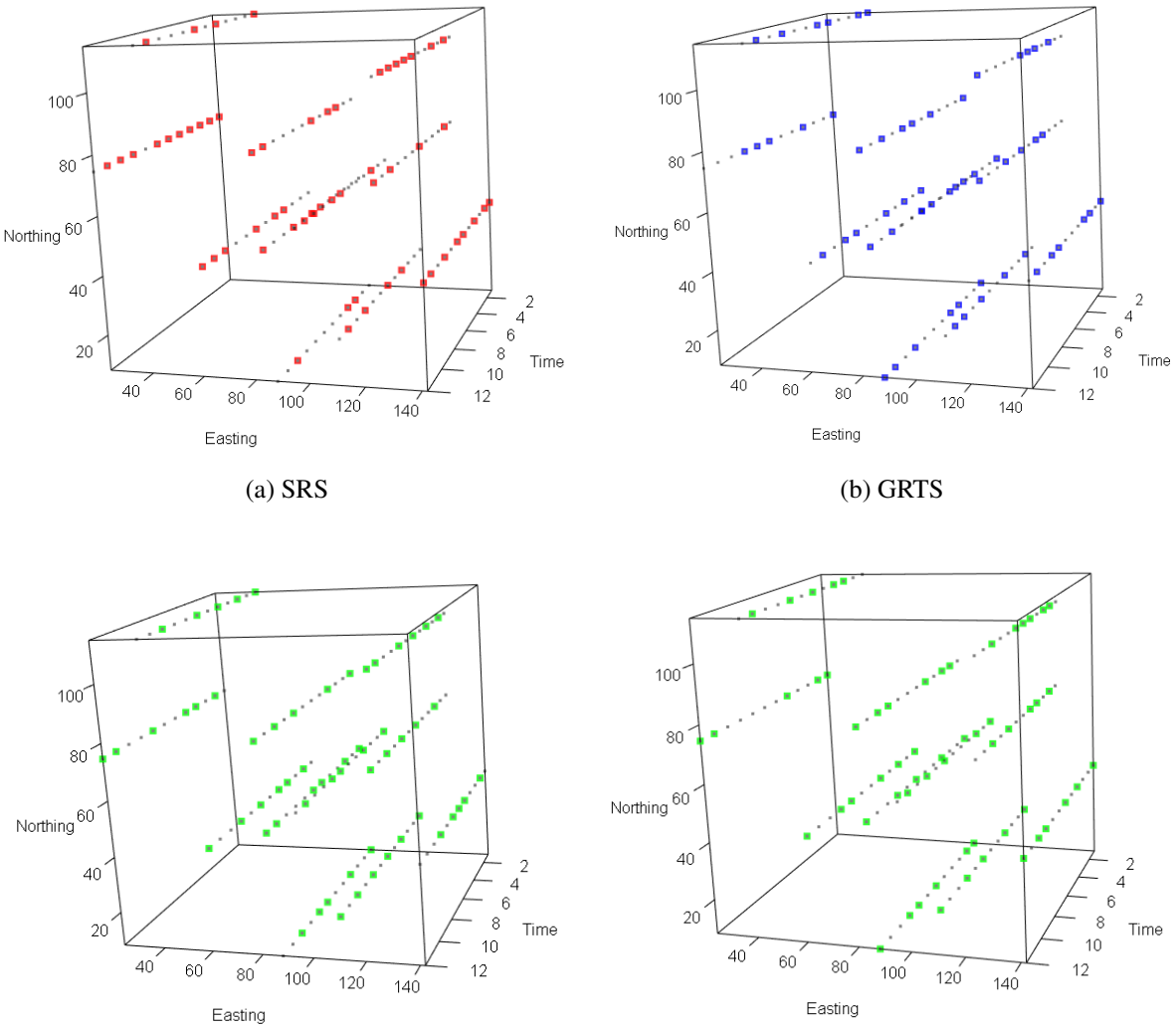


Figure 4.4: Spatial coordinates of the existing monitoring wells in case study data 1.

Moreover, both LPMs and GRTS have established ways of extending sampling designs to more than two dimensions and these multi-dimensional implementations are available in their corresponding R packages. This promotes the application of these methods for designing environmental surveys. All of these factors were considered, when selecting candidate sampling techniques for the spatiotemporal comparison. The scope of the comparison was not to perform an extensive analysis, but to outline that the results obtained in the two-dimensional settings in the literature, hold when the sampling design is extended to higher dimensions.



(a) SRS

(b) GRTS

(c) LPM1

(d) LPM2

Figure 4.5: Spatiotemporal samples of $n = 60$ drawn from the space of potential samples with $N = 132$ using SRS, GRTS, LPM1 and LPM2.

4.6.1 Comparison of Spatial Balance

The comparison was performed on the monitoring network introduced in case study data set 1 (see Section 1.4.3), which contained concentration measurements of benzene, toluene and xylene (BTEX) from 11 monitoring wells collected over a 4-year monitoring period. The data were analysed in detail in Chapter 2. In this comparison, only the monitoring network design was used, since the target of the comparison was the degree of spatial balance achieved with the investigated sampling designs. Figure 4.4 shows the monitoring network design that generated the case study data.

In this comparison, 12 hypothetical sampling events were considered. Using the monitoring network and the evenly spaced hypothetical sampling events, a three-dimensional space was generated with the dimensions being the easting and northing coordinates and time. Within this space, the potential samples were represented by the available sampling locations at the 12 sampling events, which resulted in a total of $N = 132$ potential samples. The potential sample space is shown on Figure 4.5. The sampling methods SRS GRTS, LPM1 and LPM2 were used to draw 3-dimensional, spatiotemporally balanced samples from the space of potential samples with a sample size of $n = 60$. The spatial balance values of the selected sampling designs were calculated using the Voronoi polygon-based method as described in 4.16. This relies on using a Voronoi diagram of the target area where the expected value of the sum of inclusion probabilities in each polygon is 1. The spatial balance of samples can be evaluated based on how much these values deviate from the expectation using the variance of the polygon-wise sum of inclusion probabilities. In spatially balanced samples all of these values should be close to the expectation. The spatial balance values were calculated with respect to the positions of the existing monitoring wells and the frequency of sampling events. The lower the value of this spatial balance metric, the more well spread a sample is in the three-dimensional space, given the positions of the potential samples. The spatial balance calculations were implemented using the *BalancedSampling* R package (Grafström and Lisic [2016]).

Figure 4.5 shows a comparison of the resulting spatiotemporal samples. SRS produced a sampling design that is visibly less well-spread and clustered. GRTS produced a more balanced sample, but from the visual analysis, LPM1 and LPM2 produced the most well-spread samples. Table 4.1 shows the results of the comparison in terms of the degree of spatial balance and it substantiates the conclusions made via visual analysis. More balanced samples were obtained using any of the spatially balanced design methods than SRS. LPM1 and LPM2 generated similar spatial balance values which were lower than GRTS. These results confirm the findings reported in the literature and serve as further evidence that the LPM2 algorithm reliably selects spatially highly balanced samples, even in a three-dimensional setting.

Table 4.1: Comparison of the degree of spatial balance of samples drawn using SRS, GRTS, LPM1 and LPM2 in the synthetic case study.

	SRS	GRTS	LPM1	LPM2
Degree of spatial balance	0.329	0.098	0.078	0.078

4.7 Conclusions

In conclusion, designing sampling schemes for spatially distributed populations using methods that include a stochastic component allows for a less-biased statistical analysis of population characteristics than using systematic sampling. From the available literature it is also evident that drawing spatially balanced i.e. well-spread samples can provide additional benefits such as more accurate estimation of these population characteristics and their spatial trends. Several spatially balanced sampling design algorithms have been proposed in the literature that have different ways of ensuring the selection of a spatially evenly spread sample. GRTS is the most widely used method and is based on iteratively partitioning the area that contains the population into quadrants and mapping these onto a one dimensional line that is sampled systematically. The LPM method is based on maximising the distance between the selected samples by iteratively examining the distance of neighbouring units from a randomly selected unit. The BAS and HIP methods are based on a pseudo-random series called the Halton sequence that can generate evenly spaced random points. The simulation studies presented indicate that the more recent designs, LPM, BAS and HIP select spatially more balanced samples than GRTS (Robertson et al. [2018]), which also leads to the more accurate estimation of the population total and its variance. The difference between the three best performing designs in terms of estimation however, is marginal. Therefore, selecting the most suitable method will come down to other factors. The BAS design has a higher computational complexity especially for discrete or point-based populations, which makes it prohibitive in these cases. The default implementation of the LPM design, or LPM1 is computationally more expensive than the HIP design ($O(N^2)$ at best and $O(N^3)$ at worst compared to $O(N \log N)$). However, the more efficient implementation, LPM2 using k-dimensional trees (Lisic and Cruze [2016]) matches the computational complexity of HIP without compromising its performance. Additionally, LPM2 was shown to more reliably select samples with higher degrees of spatial balance than HIP. Both the LPM2 and HIP methods have available implementations in the statistical computing language R (R Core Team [2020]), in the *SamplingBigData* (Grafström and Lisic [2018]) and the *SDraw* (McDonald et al. [2016]) packages respectively, but in its current implementation, the *SDraw* package only allows sample selection via HIP in two dimensions, while the LPM2 implementation in *SamplingBigData* can handle any number of dimensions. Thus, the LPM2 algorithm was used in the following chap-

ter for testing the proposed novel method for determining target sample inclusion weights for spatially balanced groundwater sampling designs.

In summary, spatially balanced sampling aims to maximise the distance between samples whilst allowing for random variation to occur and it can be applied in practice in various circumstances. As highlighted in this chapter, it can be used to establish new spatially well-spread networks within a given geographic region by selecting a number of locations or to select new well locations for existing monitoring networks via the use of master-frames. Given existing monitoring wells (which is the focus of this thesis) a monitoring time-frame, and a number of samples to be taken, spatially balanced sampling can be used to draw samples that maximise spatial or spatiotemporal coverage. These can be done independently from whether historic observations are available for the monitored site or not, since the objective function of the sampling is to maximise the spread of the samples. Historic observations for an existing network however, can then be used to estimate the spatiotemporal dependence, examine the stability of estimates in time and select optimal sampling locations accordingly. Moreover, applying spatially balanced sampling on historic data can also help assess the adequacy of the historic sampling intensity and the trade-off between high volume spatial or high frequency temporal sampling. These applications are explored in Chapters 5 and 6.

As mentioned in Section 4.4, a sample inclusion density function that determines the target inclusion weights allows for incorporating additional relevant information or auxiliary variables about the monitoring site into spatially balanced sampling designs. While Grafström et al. [2012] performed a simulation study using both equal and unequal target inclusion weights (proportional to the basal area of the tree), the focus of the analysis was the spatial balance of the samples. Therefore, it is unclear from the literature examined above if the application of an inclusion density function that is proportional to the process that drives the spatial and temporal distribution of the target population can improve the estimation of population characteristics. Thus, the aim of the following chapter was to describe the development of a novel method of determining target sample inclusion weights for spatially balanced sampling designs of groundwater monitoring networks, and to test whether it provides benefits for modelling CoPC plumes and for estimating population characteristics over designs with equal target inclusion weights.

Chapter 5

Data-driven Tuning of Inclusion Weights in Spatiotemporally Balanced Sampling Designs

It has been highlighted throughout this thesis that sampling strategies for groundwater quality monitoring networks should aim to capture as much of the spatiotemporal patterns of CoPC plumes as possible to improve the statistical estimation of their characteristics. As shown in Chapter 4, when relying on probability sampling, spatially balanced designs do better at capturing spatial structures than simple random sampling (SRS) due to their better coverage of the target area. It was also discussed how these techniques can be extended to an additional dimension to draw spatiotemporally balanced samples. As mentioned in Section 4.1, balanced sampling design algorithms also allow for the incorporation of target inclusion density functions, that determine the inclusion weights of the available sampling locations in subsequent designs (see discussion around Figure 4.3 in Chapter 4.4.2). This provides an opportunity for tuning the algorithm to prioritise certain sampling locations over others based on additional information about the site under investigation or the underlying processes that determine the distribution of the population. An example that illustrates this application is provided by Grafström et al. [2012] on a data set containing measurements regarding a population of trees within a geographic region. The inclusion probabilities of the sampling design were chosen to be proportional to the basal areas of the tree trunks. In other words, the sampling design aimed to collect more samples from larger trees than smaller ones. In the groundwater contamination monitoring context, access to certain sampling locations can often be restricted, or there might be a need to concentrate sampling effort on a certain sub-region. These factors can be integrated into the sampling design by specifying an inclusion density function.

However, the main reason for tuning inclusion weights in the case of long-term groundwater quality monitoring is to try and track the progression of the CoPC plume more closely on the basis of historical information. As highlighted in Chapter 4, the spatial distribution and temporal evolution of groundwater CoPC plumes is determined by several factors such as the source and chemical characteristics of the CoPC, the rate of its release, the hydrological conditions (precipitation, recharge rate, etc.) and geological parameters (aquifer composition, hydraulic conductivity, etc.). If available for a given contaminated site, historical CoPC concentration data could be used to produce an initial estimate of spatiotemporal plume characteristics, such as the changes in its extent over time. These initial estimates can then be used to update the sampling strategy by tuning the target inclusion weights to prioritise monitoring wells that might be more useful in capturing the spatiotemporal structure of the plume. This modification to the spatially balanced sampling algorithm could lead to improved accuracy in the estimation of certain plume characteristics in the future. This approach provides the benefits of probability sampling and spatially balanced sampling whilst allowing more control over the selection. Being data-driven, the method also allows for integration in groundwater monitoring data analysis software such as GWSDAT. As will be shown throughout Chapters 5 and 6, the proposed approach could be used to evaluate historic data in terms of its sampling design and suggest new designs for the future, optimised to support subsequent data analyses.

Thus, in this chapter a data-driven methodology will be proposed for generating proportional target inclusion weights for spatially balanced groundwater contamination sampling designs. The proposed method aims to produce spatiotemporal samples that capture the spatial features of groundwater CoPC plumes to a greater extent than simple random sampling designs and spatially balanced sampling designs with equal target sample inclusion weights. The sample inclusion weight tuning approach was evaluated via the estimation of CoPC plume characteristics in the synthetic groundwater contamination data introduced in Section 1.4.2. The method generates target sample inclusion weights for groundwater monitoring wells at given sampling times, that are proportional to their predicted distance from the CoPC plume at the given time. Wells closer to the plume receive higher inclusion weights than wells farther from the plume. The well-plume distances were predicted using generalised linear models (GLMs) on time-series data derived from past observations of CoPC concentration in groundwater.

5.1 Data-Driven Tuning of LPM Inclusion Weights

Rather than maximising the spatial balance of the sample distribution, the goal of the target inclusion density function is to produce sampling designs that allocate more effort to sampling locations that are predicted to be closer to the solute plume during future sampling campaigns.

The inclusion weight tuning parameters for future sampling campaigns are determined using a data driven approach. First, a spatiotemporal, P-splines model of the historic solute concentrations is used to generate time-series of the Euclidean distances between the solute plume and each sampling well. The time-series data are then extrapolated to the future sampling campaigns using GLMs. Finally, a half-normal kernel function (see Section 5.1.6) is used to determine the value of the inclusion weight tuning parameter for each well at each future sampling time based on its predicted distance from the solute plume. The half-normal kernel function is scaled automatically by the number of sampling wells in the data set and the standard errors of the predicted distances by the GLMs. The reason why future distances between the plume and monitoring wells were estimated from the time-series using GLMs is because they could not be extrapolated directly from the spatiotemporal P-splines model of the historic observations. The P-splines modelling approach is appropriate for interpolation but not for the prediction of future results. However, the aim of this work was to develop and test a data-driven sampling design algorithm that can be used with the spatiotemporal P-splines modelling approach.

5.1.1 Spatiotemporal P-splines model of historic data

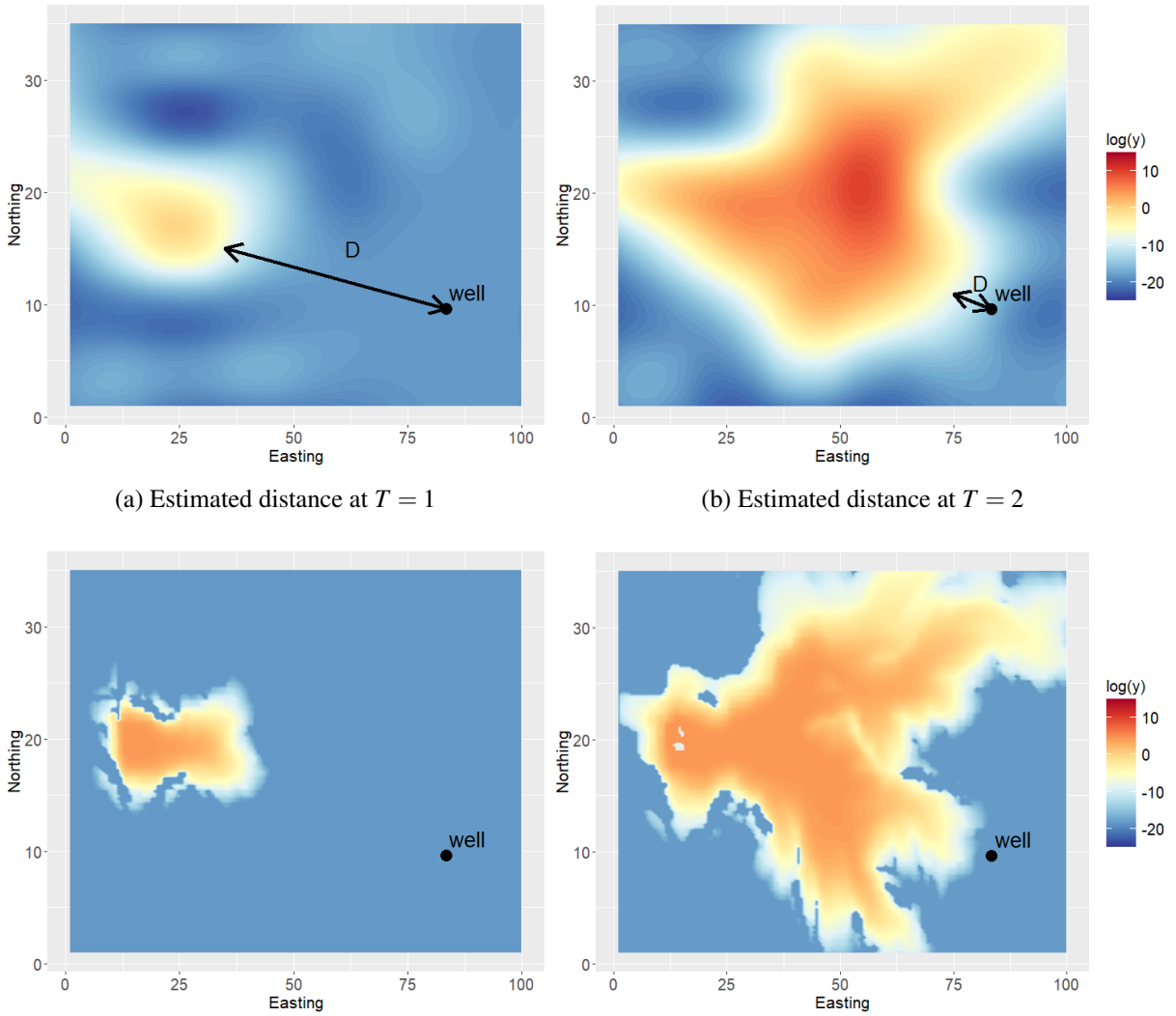
As the first step in the novel sample inclusion weight tuning method, the groundwater CoPC concentration surface has to be modelled from historic concentration measurements to allow for the estimation of the distances between the monitoring wells and the plume. These historic concentration measurements could be modelled using any modelling technique that estimates the solute concentration surface over the given period. In this implementation they were modelled using a spatiotemporal P-splines approach, which is based on the work of Evers et al. [2015], and was described in detail in Section 1.3.1. The model is described by equation 1.6 and can be represented in vector matrix form as:

$$\boldsymbol{y} = \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is the vector of solute concentration observations, $\boldsymbol{B}(\boldsymbol{x})$ is the matrix of basis functions evaluated at each data point, $\boldsymbol{x}$ is the matrix of covariates (sampling well coordinates and time points), $\boldsymbol{\alpha}$ is the vector of basis coefficients and $\boldsymbol{\varepsilon} \sim N(\mu, \sigma^2)$ describes random variation. The spatiotemporal dependence of $\mathbf{y}$ is described in the systematic component of the model by the spatiotemporal smooth $\boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha}$.

Throughout the investigations in this chapter, the P-splines model parameters that were used were referred to as the default settings in Chapter 3, and were described in Section 3.3. The number of B-spline basis functions in each covariate dimension (space and time) were set at 9

and they were constructed of second order polynomials. These parameter values correspond to the default values used in the modelling framework in GWSDAT (Jones et al. [2023]).



(a) Estimated distance at $T = 1$

(b) Estimated distance at $T = 2$

(c) True distance at $T = 1$

(d) True distance at $T = 2$

Figure 5.1: Evolution of distance between monitoring well and CoPC plume between two time points $T = 1$ and $T = 2$ for the model estimate of the concentration surface of the synthetic data set described in Section 1.4.2 (sub-figures a) and b)). D represents the distance between the monitoring well and the plume. The true concentration surface is shown for reference at the given time points.

5.1.2 Delineating the CoPC plume

In order to be able to calculate distances between the plume and the monitoring wells, the boundary of the plume had to be defined. In general, the diffusion of solutes in groundwater causes concentration levels to decrease continually with increasing distance from the source. Thus, a

functional boundary for the plume can be defined by a concentration limit. There are national and international directives that define recommended permissible concentration limits for most constituents of potential concern depending on the value of the water body. These limits can be used to delineate the solute plume and identify its functional boundaries. For example, the recommended permissible level (PL) of the organic CoPC benzene in groundwater is 0.01 mg/L according to the World Health Organization Guidelines for Drinking Water Quality [1]. Therefore, the functional boundaries of a benzene plume in groundwater would be identified at this concentration level.

The plume delineation in the simulation study was performed using the above approach. Let $\hat{y}_{ti}$ be the estimated solute concentration corresponding to the coordinate (x_i, y_i) $(i = 1, 2, ..., I)$ on the site at historic sampling time t, where $t = 1, 2, ..., 10$. The spatial extent of the plume was given by the coordinates that correspond to values of $\hat{y}_{ti} \geq 0.01\ mgL^{-1}$. These coordinates will be represented by matrix $\boldsymbol{P}_t$:

$$\boldsymbol{P}_t = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ . & . \\ x_I & y_I \end{bmatrix},$$

where $x_1, x_2, ..., x_I$ are the easting, $y_1, y_2, ..., y_I$ are the northing values and $t = 1, 2, ..., 10$ are historic sampling times.

5.1.3 Calculating Euclidean distance between each monitoring well and the plume

The shortest distance between each sampling well and the plume was then calculated for each historic sampling event. Let us call the set of monitoring well coordinates $\boldsymbol{W}$. To determine the Euclidean distance d_{wt} of each well w $(w = 1, 2, ..., W)$ to the plume at sampling time t, we need to find (x_i, y_i) in $\boldsymbol{P}_t$ for each (x_w, y_w) in $\boldsymbol{W}$ such that:

$$d_{wt} = \min_{(x_i, y_i) \in \boldsymbol{P}_t} \sqrt{(x_i - x_w)^2 + (y_i - y_w)^2}. \tag{5.1}$$

Figure 5.1 shows an example of the calculation of the plume-well distance over time.

[1] https://www.who.int/publications/i/item/9789240045064

5.1.4 Monitoring well distance to plume time-series

Using the results from equation 5.1 for each monitoring well (w), a time-series of the Euclidean distances (d) to the plume was established. The time-series showed the change in estimated distance between the monitoring well and the plume throughout the historic sampling times. Figure 5.2 shows an example of the well-plume distance time-series data as well as the fitted GLM used for extrapolation to future sampling events.

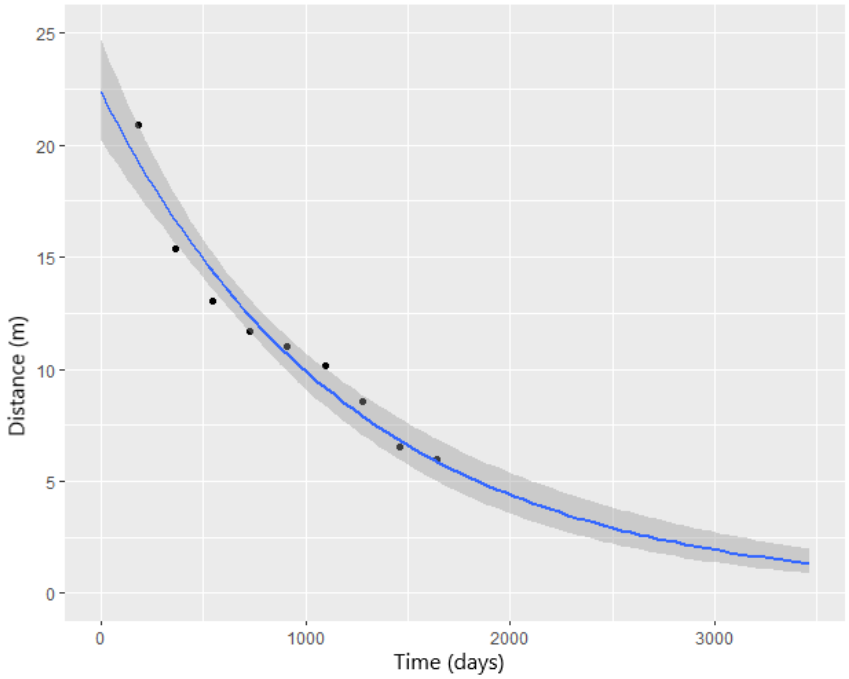


Figure 5.2: Time-series data of estimated historic plume distances in meters from a monitoring well and the corresponding GLM with log link function (blue line) used for the prediction of plume distances at future sampling times.

5.1.5 Modelling distance time-series

The aim is to predict well-plume distances at future sampling times using the time-series data. To extrapolate these results, the historic time-series were modelled using a GLM framework. The GLMs took the following form for each monitoring well:

$$log(d_t + 1) = t_t \beta, \tag{5.2}$$

where d_t are the predicted distances at time $t = 1, 2, ..., 10$, t_t are the sampling times and β are corresponding coefficients. One was added to the response variables (distances) to avoid the

presence of zeros within the link function. The logarithmic link function was used because the historic well distance time-series data generally showed non-linear, exponentially decreasing behaviour as shown on Figure 5.2. This also makes sense from a hydrogeological perspective. The spread of the solute in groundwater is mainly limited by diffusion (assuming the effects of advection are negligible in comparison), which is a temporally non-linear process. The future distances were predicted by extrapolating the model to t values between $10 < t \leq 20$. These fitted values ($\hat{d}_t$) were obtained by the expression:

$$\hat{d}_t = log^{-1}(t_t\beta) - 1. \tag{5.3}$$

Thus, an estimated distance between each monitoring well and the solute plume was obtained for each future sampling time point. These predicted distances were used to tune the sample inclusion weights proportionally through the use of a kernel function. Figure 5.2 shows an example of the fitted GLM used for extrapolation to future sampling events.

5.1.6 Tuning parameters through the kernel function

A kernel function was used to establish the relationship between the predicted future distances and the tuning parameters that were subsequently used to adjust the target inclusion weights. The selection and scaling of the kernel function was crucial in determining how large the variation in the values of the target inclusion weights would be. The amount of variation determines the strength of the prioritization among the monitoring wells. With too little variation, the resulting sampling design is close to a spatially balanced sample. On the other hand, with too much variation the sampling designs are too biased towards the plume, and consequently omit too much information from areas of lower solute concentrations.

Two types of kernel functions were investigated: a linear and a half-normal function. A comparison of the two functions is shown on Figure 5.3. The linear function established an inverse linear relationship between the predicted distances and the tuning parameters such that $f(d) = md + c$, where m represents the gradient and c represents the intercept. Thus larger distances corresponded to smaller tuning parameter values and consequently, smaller target inclusion weights (see Figure 5.3a). The function decreased linearly from a predicted distance of 0 m to the maximum predicted distance. The range of the tuning parameters in this case could vary between any two arbitrary numbers larger than one. This is because the actual values of the tuning parameters did not matter, only their ratios to each other, since they would be used to scale the inclusion weights as described in equation 5.5. However, it was identified that a linear kernel function did not produce large enough variation in the target inclusion weights and the resulting sampling de-

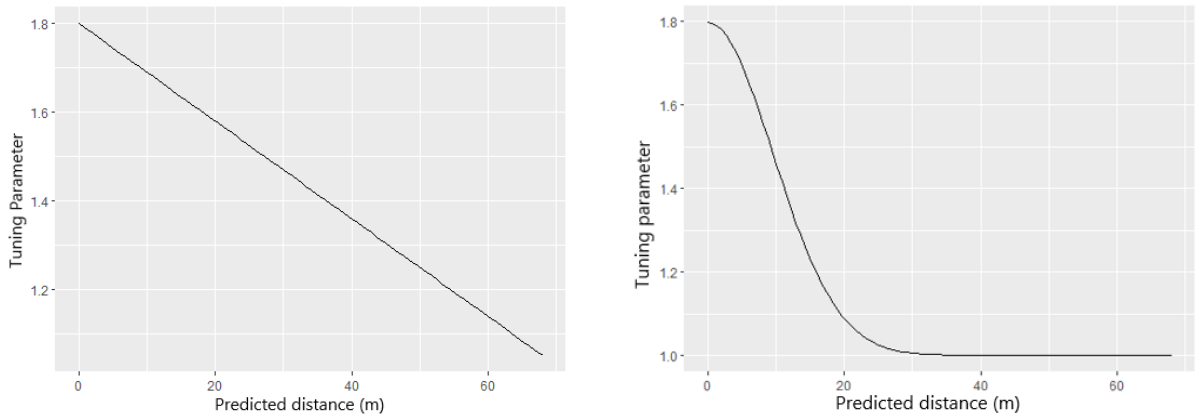
signs were close to equally weighted spatially balanced designs. Section 5.3.1 shows the results related to this investigation.

Thus, it was determined that the the tuning parameter values needed to decrease more sharply at smaller distances and eventually level off towards the the end of the distance range. This approach would induce stronger prioritization towards monitoring wells in close proximity to the plume and converge on equal inclusion weights for wells farther away. Therefore, the probability density function of a half-normal distribution was introduced as the kernel function. The shape of this function makes it a suitable kernel function for the selection of inclusion weight tuning parameters. The maximum value of the function occurs at a predicted distance of 0. From here the function decreases sharply over a short distance before it levels out and approaches the minimum value of 1. The function shown on Figure 5.3b is described by the expression:

$$f(d;\sigma) = 1 + \frac{1}{s} \times \frac{\sqrt{2}}{\sigma\sqrt{\pi}} exp\left(-\frac{d^2}{2\sigma^2}\right) \quad d > 0, \tag{5.4}$$

where d represents the estimated distance between the plume and the monitoring well, in the first term s is a scaling parameter that controls the range of tuning parameter values on the y-axis, while the value of σ controls the sharpness of the decrease. The asymptote of the function was set to 1 to avoid generating values below 1, which would decrease the value of the resulting inclusion weights due to multiplication as shown in equation 5.5. Together, the two scaling parameters s and σ control the strength of prioritisation. A larger range in tuning parameter values induces larger differences between target inclusion weights while a sharper decrease reduces the distance at which substantial increase in weights occurs. The values of these two parameters should not be constant since that would result in different levels of prioritisation across different scenarios. Instead, they should be tuned automatically by the data to adapt to different ranges in predicted well-plume distances and the number of monitoring wells.

The value of the scaling parameter, s was set to the value of the equal sample inclusion weight ($e_j = W/n$), which represents the ratio between the number of samples to be taken (n) at a given sampling time and the number of available sampling locations, i.e. monitoring wells (W). The reason for this was that for a lower sample-to-well ratio (low number of samples from a high number of wells) a higher range in the tuning parameter values has to be induced to result in larger differences in inclusion weights, and consequently, stronger well prioritisation. Thus, the samples would be more likely to be drawn from wells that are predicted to be closer to the plume. Conversely, in case of a high sample-to-well ratio, there is less need for strong prioritisation since each well already has a higher probability of being selected for sampling. Hence, the strength of prioritisation, i.e. the range of tuning parameter values, was a function of the sample-to-well ratio.



(a) Linear kernel function

(b) Half-normal kernel function

Figure 5.3: Tested kernel functions establishing the relationship between predicted well-plume distance and the target inclusion weight tuning parameters.

The parameter σ controls the shape of the function, in terms of how quickly the value of the function decreases with increasing predicted distance. This parameter can adjust the distance at which the prioritization of wells becomes more dominant. Therefore, this should reflect the uncertainty about the well-plume distances predicted by the GLMs. If the uncertainty about the predicted distances is high, the sampling design should account for the possibility that the given well has already come into contact with the plume. In this case, it is sensible to have a greater distance to the drop-off point on the curve. On the other hand, if there is less uncertainty about the predictions, the prioritization can be more focused on the shorter predicted distances. Therefore, to scale σ automatically based on the input data, its value was determined by the median estimated standard error of the predicted well-plume distances from the GLMs, which were described by equation 5.2.

After adequate scaling, the kernel function can be used to determine the values of the tuning parameters for each well at each future sampling time, by solving the function for corresponding values of predicted distances (d).

5.1.7 Tuning of sample inclusion weights

The tuning parameters were used to adjust the (initially equal) target sample inclusion weights for the LPM-based sampling design via the equation:

$$u_l = e_l p_l \frac{\sum_{l=1}^{n} e_l}{\sum_{l=1}^{n} e_l p_l}, \tag{5.5}$$

where u_l are the adjusted target sample inclusion weights, e_l are the initial, equal sample inclusion weights, p_l are the tuning parameters and $l = 1, 2, ..., n$ where n is the sample size. In LPM-based sampling designs, the sum of the target inclusion weights has to equal the number of samples. Equation 5.5 ensures that the sum of the adjusted inclusion weights still equal the number of samples, but their values are proportional to the tuning parameters as determined via the kernel function and the novel approach.

5.2 Simulation Study

The proposed approach for adjusting sample inclusion weights was evaluated in a simulation study using synthetic groundwater contamination data. The approach was used to determine target sample inclusion weights in sampling designs created using the LPM algorithm. Spatiotemporal P-splines models (see Section 1.3.1) were fitted to the obtained samples, and the resulting predictions were used to estimate various metrics describing the difference between the estimates and the truth. Results from the approach which will be referred to as proportional LPM or pLPM, were compared to results from SRS and spatially balanced LPM with equal target inclusion weights (eLPM). The flowchart in Figure 5.4 shows the structure of the simulation study in terms of how the pLPM and LPM samples were drawn, and how the resulting estimates were compared. The simulation study was performed on various scenarios to allow for the investigation of trends in the results that were not specific to any one scenario but were common among all of them. 100 simulation runs were completed for each scenario. The scenarios were constructed from a combination of the following parameters:

1. synthetic contamination scenario (see Section 1.4.2 for more details):
   (a) simple plume
   (b) mid plume
   (c) complex plume
2. monitoring well arrangement:
   (a) grid
   (b) random
   (c) clustered random
3. number of monitoring wells:
   (a) 24

(b) 48

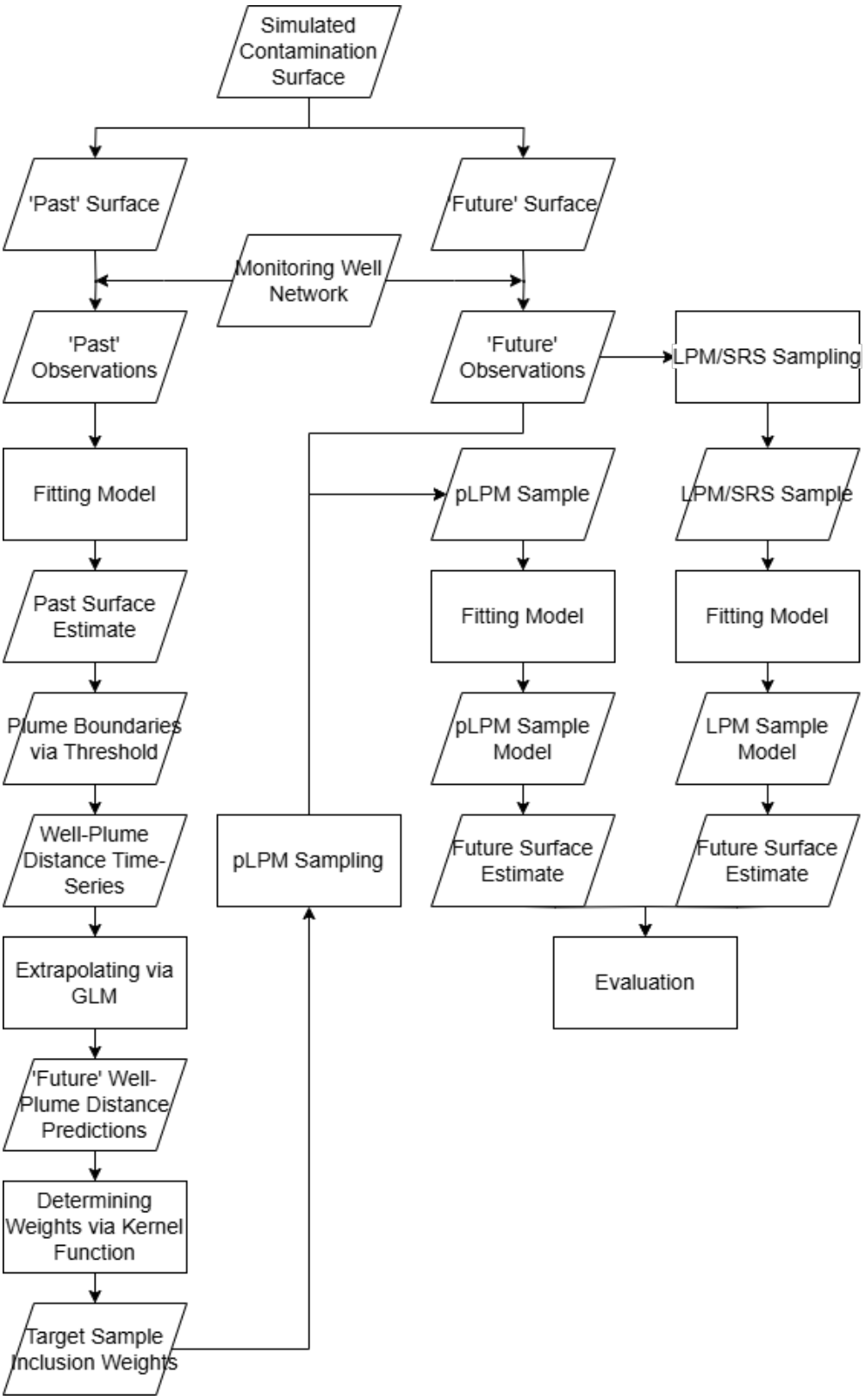


Figure 5.4: Flowchart for the structure of the simulation study including the steps involved in the pLPM sampling approach as described in Section 5.1.

The solute concentration surfaces of the simulated CoPC plumes are shown in Figure 3.1 in Chapter 3. The number of samples drawn (n) was also varied in the simulation study. The range of n depended on the number of available monitoring wells. For 24-well networks the number of samples taken throughout the 10 hypothetical sampling events were 30, 60, 120 and 240. For 48-well networks an additional 480 sample scenario was also included.

5.2.1 Temporal Distribution of Samples

Section 4.1 provided an overview on the available options regarding the distribution of samples in time when designing long-term sampling schemes based on stochastic algorithms. The first option is to consider each sampling event separately and generate independent spatial sampling designs for them. This approach has several practical benefits. For instance, there is no need to specify the dates of sampling events in advance, hence, they can be determined independently based on other relevant factors such as expert judgement or legal requirements. Another benefit of this approach is that it is easy to make alterations to it throughout the monitoring period. The only input parameter required for each individual sample design in this approach is the sample size.

If there is a specific number of samples that need to be collected throughout the monitoring period, and either the exact dates of the sampling events or the beginning and end dates of the monitoring period are predetermined, the samples can also be distributed in time, not just space. In this case, a finite 3-dimensional space can be constructed using the monitoring well coordinates and either the sampling dates or the beginning and end dates of the monitoring period. Using the exact dates of the sampling events will provide a discrete time axis, while using the beginning and end dates will provide a continuous one. In either case, the required number of samples can be distributed spatiotemporally in this 3-dimensional space using a stochastic sampling algorithm. The benefit of this approach is that it can generate a sampling scheme for the entire monitoring period, thus reducing the effort that needs to be put into sampling design later on. However, it requires determining the total number of samples and the length of the monitoring period in advance. This is impractical, since long-term monitoring commonly needs to be performed until CoPC concentrations return to acceptable levels, which is dependant on environmental factors. This difficulty can be circumvented by only generating the sampling scheme for an arbitrary, shorter period nested within the unspecified length monitoring period. Moreover, the sample size at each sampling event can vary significantly, and as mentioned before, due to transportation costs, it is more economical to collect a higher number of samples at a time. Low sample sizes at certain time slices would also produce low spatial resolution samples. It is also more difficult to make alterations to such a sampling scheme throughout the monitoring period.

The third option is to divide the available monitoring wells into two or more groups. Two groups can be created by selecting half of the wells using a sampling design algorithm, and putting the non-selected wells into a second group. The two groups can then be sampled in an alternating manner from sampling event to sampling event. Three groups can be created the same way, by selecting a third of the wells first, and then another third and so on. This approach provides some of the same benefits as the first one, namely that there is no need to specify the number of sampling events or total required number of samples in advance. If it is specified that a well cannot be included in multiple sub-groups, this approach can also ensure that all of the wells will be sampled at some point. Logistically, it is also convenient that the same groups of wells need to be visited during sampling events. However, it can be difficult to choose samples sizes, as it is determined by the distribution of wells among the sampling groups. Moreover, it can be complicated to make alterations to this design throughout the monitoring period. If certain wells become unusable, the group structure needs to be re-designed if the sample size is to be maintained. This design is also not ideal to use with the pLPM sampling design. Due to the unequal inclusion probabilities of the wells, some sample groups could provide less relevant information than others. This approach is also more prone to the weaknesses of stochastic sampling designs. If the initial group selection is inadequate for some reason, all of the samples collected down the line will be inadequate. While the modelling approach can make up for some of this inadequacy, generating new stochastic designs for each sampling event is statistically a more reliable strategy to avoid these situations.

Overall, the first option is the least complex from a practical point of view, and it has the least amount of drawbacks. It is common practice in long-term groundwater contamination monitoring to only prepare for the upcoming sampling event rather than generate a complete, long-term design. Thus, this approach of creating independent spatial sampling designs for each sampling event can help maximise the efficacy of long-term monitoring.

Therefore, this approach was used throughout the simulation study, where the samples were distributed evenly among the sampling events. This means that an equal number of samples were taken at each of the 10 events. In each event, wells were selected for sampling using the investigated design algorithms. Table 5.1 shows an example of such a sample distribution. The next chapter (Chapter 7) will investigate the effects of the trade-off between sampling frequency (number of sampling events) and sample volume (number of samples taken) per event.

Table 5.1: Example of sample distribution among sampling events for collecting a total of 120 samples.

Sampling Event	1	2	3	4	5	6	7	8	9	10	**Total**
Number of Samples Drawn	12	12	12	12	12	12	12	12	12	12	120

5.2.2 Synthetic groundwater contamination data

The generation of the synthetic groundwater contamination data sets used for the simulation study was described in detail in Section 1.4.2. The three hypothetical scenarios simulated geometrically increasingly complex CoPC plumes. They contained values of solute concentration in groundwater for a 1×1 km^2 area with 150×150 point coordinates representing the concentration surface. The concentration values were recorded at $T = 20$ hypothetical sampling events representing 6-month intervals for a 10-year monitoring period and range between $0 - 100$. In the simulation study, 15% multiplicative error was added to the concentration values to mimic the noise from sampling and analytical inaccuracies of real observations. The amount and type of noise added was based on expert recommendation, initially for the work of (McLean et al. [2019]). The noise was added according to the equation:

$$y_i = z_i * \varepsilon_i, \tag{5.6}$$

where y_i is the i-th data point with added noise, z_i are data points without measurement noise and ε_i is a random variable drawn from a normal distribution $N(\mu, \sigma^2)$ with mean $\mu = 1$ and standard deviation, $\sigma = 0.15$. The CoPC concentration values with added noise were also transformed to normalise their distribution as discussed in Section 2.2. The transformation also described in Section 2.2 gives the transformed concentration values y'_i by the expression

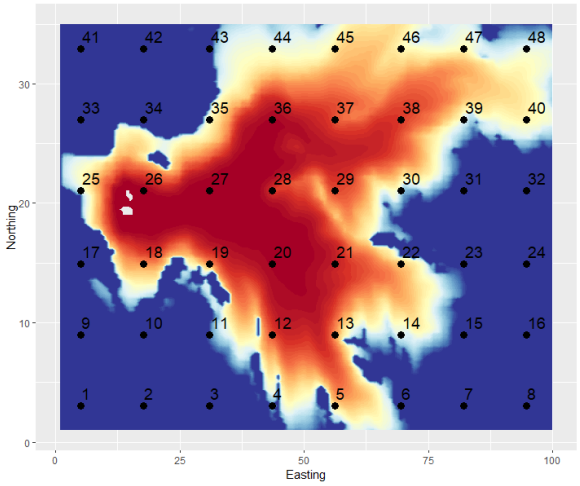
$$y'_i = log_e(1 + y_i).$$

The novel pLPM approach relies on past observations to estimate future well-plume distances, while the aim of the study is to evaluate error metrics derived from the newly generated sample designs. Therefore, the synthetic data sets were split into two parts along the temporal axis. The first ten sampling times ($1 \leq t \leq 10$) were treated as the historic data and the last ten sampling times ($10 < t \leq 20$) were treated as the future data. The historic data were subsequently used in the pLPM sampling design approach to tune the target sample inclusion weights as described in Section 5.1, while the future data were the parts actually being sampled with the three different sampling design approaches, namely pLPM, LPM and SRS (see Figure 5.4).

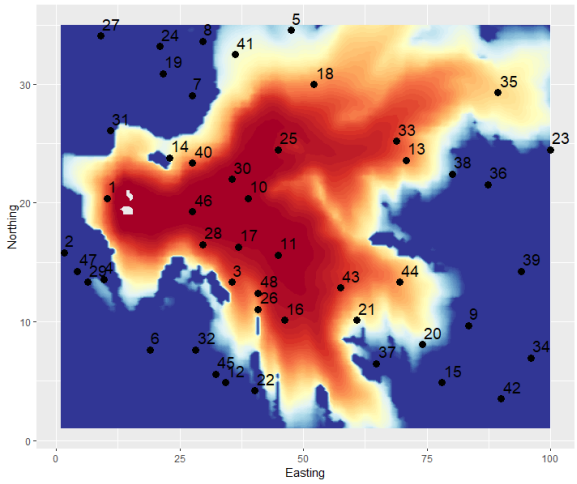
5.2.3 Simulated monitoring well networks

The monitoring wells represented the population that the sampling designs ultimately targeted. As discussed in the previous chapter (4), this meant that the sampled population was discrete

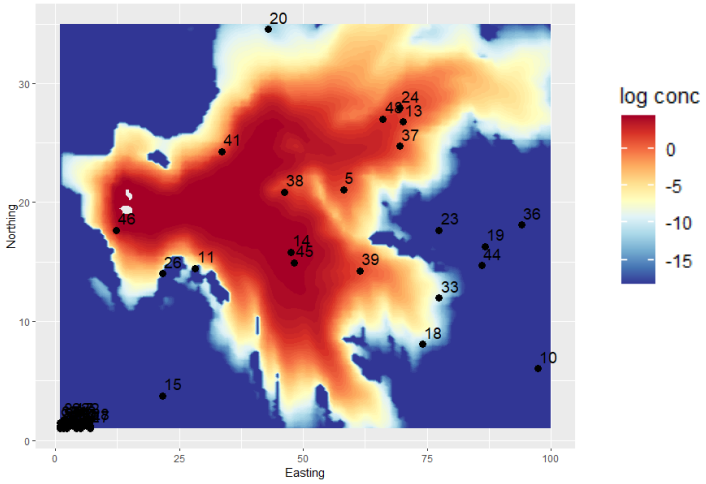
rather than continuous. Therefore, the arrangement of the monitoring wells was the limiting factor in terms of how spatially balanced the samples could be with respect to the contaminated area, and in terms of how well the sample can capture the spatial and temporal trends in the data.



(a) Grid arrangement



(b) Random arrangement



(c) Clustered random arrangement

Figure 5.5: Simulated monitoring networks with 48 wells against the heatmap of the complex contamination plume scenario showing the log-transformed values of the solute concentrations.

The arrangement of the monitoring wells in a grid pattern (see Figure 5.5a) provided the most well-spread population, that maximised the influence of the sample design on the outcome of the sampling process. The randomised arrangement (see Figure 5.5b) mimics a more realistic monitoring network, where the locations of the wells are subject to many different factors such as land availability, hydrogeological parameters and accessibility (see Section 4.1). Due to these factors, the arrangement of monitoring wells can appear somewhat random without additional information about the site. Moreover, such factors do not need to be taken into account for the synthetic contamination scenarios, since they have not been specified. Finally, the clustered

randomised arrangement (see Figure 5.5c) was designed to test the efficacy of the spatially balanced LPM-based sample designs.

It was expected that SRS would draw more samples from the same cluster of wells, whereas LPM would account for the presence of the cluster and would compensate for it. Figure 5.5 shows a comparison of the monitoring network arrangements. The synthetic groundwater contamination data sets were filtered using the locations of the monitoring wells given by the selected arrangement. The resulting data points were considered the observations in the study.

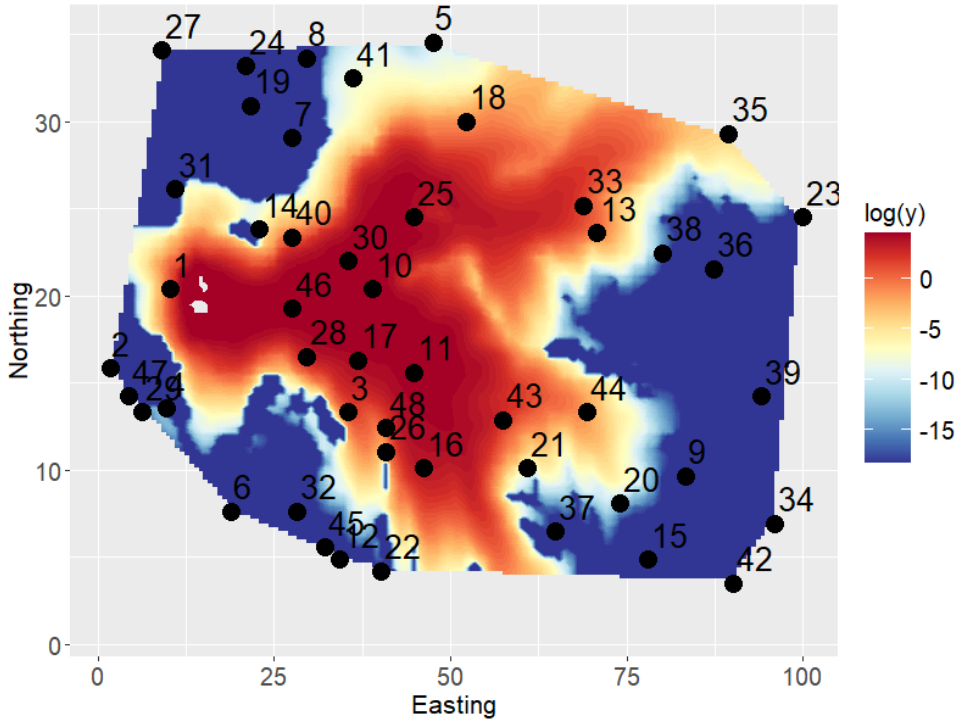


Figure 5.6: Convex hull of the randomly arranged network with 48 monitoring wells against the heatmap of the complex CoPC plume showing the natural logarithm of the solute concentrations.

When modelling real groundwater contamination data sets it is common practice to only predict concentration values within the convex hull, i.e. the area encircled by the monitoring wells, since there are no observations available from outside this area and extrapolation can be highly uncertain. This is especially true for splines-based models, which tend to show extreme behaviour towards the edges of the modelled domain. Thus, additional tests were conducted to investigate if the results obtained for the synthetic data set would still hold true if only the convex hull were considered. Figure 5.6 shows the convex hull of the random network with 48 monitoring wells.

5.2.4 Evaluation metrics

The sampling designs were evaluated using a variety of evaluation metrics that intended to explore the performance of the proposed proportional inclusion weight design. The primary focus of the assessment was how well the samples could be used to estimate the spatiotemporal CoPC concentration surface. Therefore, the main evaluation metric was the RMSPE of the surface estimates, made by the P-splines models. The estimators of population characteristics such as the Horvitz-Thompson estimator, the Yates-Grundy-Sen estimator, the Hajek-Rosen estimator or the local mean variance estimator introduced in Chapter 4 were not utilised in this research. Chapter 4 discussed the available literature on how the different spatially balanced sampling designs affect the variances of these estimators. As Benedetti et al. [2017] discusses, the uncertainty in the inference of population characteristics is introduced by the randomness of the sample selection algorithms. In the presented research, this uncertainty was approximated by the variability in the RMSPE values observed in the simulation study resulting from the estimation of the concentration surfaces using the obtained samples. Other parameters such as the model specifications remained constant throughout the simulations.

Ratio of samples drawn from the CoPC plume

The ratio of samples drawn from within the area delineated as the CoPC plume was used to evaluate the degree of prioritisation towards these samples introduced by the pLPM approach. It was expected that this ratio would be higher for pLPM samples than SRS or LPM samples. The ratio was calculated by counting the number of samples (s_i) within a sample design where the value of the CoPC concentration in groundwater exceeded the concentration threshold defined in Section 5.1, i.e. $s_i \geq 0.01$ mg/l. Let us call the number of such elements n_p. The ratio is then calculated by n_p/N, where N is the total number of samples.

Spatial balance

The degree of spatial balance of the samples was calculated using the Voronoi polygon method described by equation 4.16 in Section 4.5:

$$v_k = \sum_{i \in VP(k)} \pi_i,$$

where v_k was the sum of the inclusion probabilities of the monitoring wells within the k-th Voronoi polygon. For a spatially balanced sample, all of the v_k values should be close to 1.

Thus, the variance of v_k ($Var(v_k)$) is an indicator of the spatial balance of a sample, where a lower value indicates a spatially more balanced sample. The obtained spatial balance values are relative to the locations of the monitoring wells, meaning that a sample design that draws a sample from each well would have a spatial balance of 0.

Root mean squared prediction error

The groundwater CoPC concentration estimates derived for the complete future part of the data set using the various sample design approaches were compared to the truth using root mean squared prediction error (RMSPE) values. The RMSPEs were calculated using the equation:

$$\text{RMSPE} = \sqrt{\frac{\sum_{i=1}^{I}(y'_i - \hat{y}'_i)^2}{I}}, \tag{5.7}$$

where I is the total number of data points in the future part of the data set with i=1,2,...,I, y'_i are the log-transformed (see Section 5.2.2) true CoPC concentration values of the data points and $\hat{y}'_i$ are the estimated values obtained using the P-splines models.

Besides the RMSPE for the complete data set, the RMSPEs for the CoPC plume and the area outside the plume specifically were also isolated, by considering data points in equation 5.7 which exceeded or did not exceed the concentration threshold value (0.01) respectively. This was done to allow for the analysis of potential benefits or downsides to using the pLPM sampling approach, as it was expected to lead to more precise estimation of the CoPC concentration values within the plume and correspondingly less precise estimation outside of it.

Plume mass

The plume mass was also estimated from the predicted CoPC concentrations and compared to the true value calculated using the true CoPC concentrations. The plume mass in this case refers to the total of the concentration values over space at the recorded sampling times. Thus, the plume mass was defined as the sum of concentration values within the plume, divided by the area of the hypothetical site:

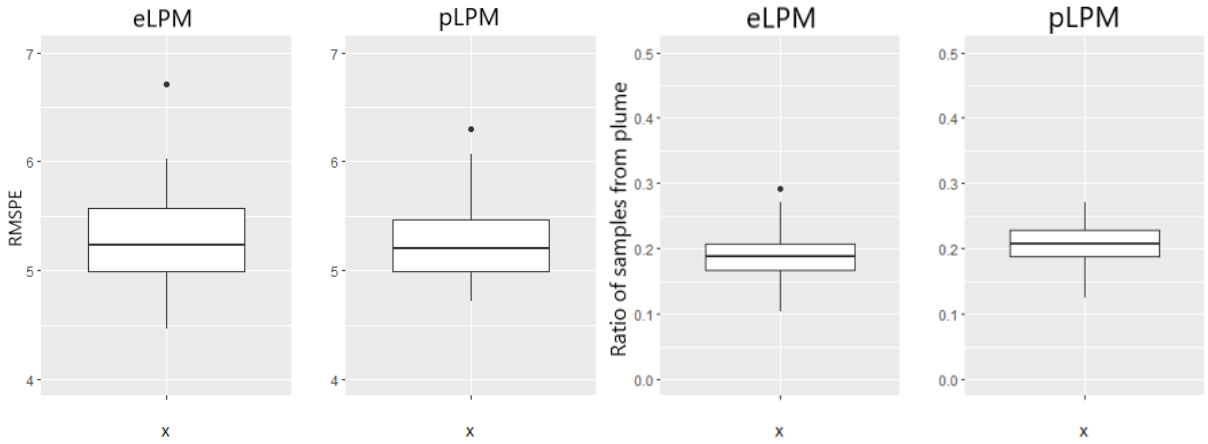
$$\hat{M}_{plume} = \frac{\sum_{j=1}^{J}\hat{y}_j}{A}, \tag{5.8}$$

where J was the number of data points that exceeded the plume threshold value of 0.01 mg/L,

and A was the area of the full simulated site with a value of $35 * 100 = 3500$ as shown on Figure 5.5. The division by the area as a constant was introduced to reduce the order of magnitude of the results. The origin of this constant has no bearing on the comparison of plume masses, as any arbitrary constant would suffice. The true plume mass was calculated using the true simulated values instead of the estimated ones.

5.3 Results

This section will describe the results of the simulation study and any additional investigations carried out to gain a better understanding of the impact of the novel sampling design approach on modelling CoPC concentrations compared to simple random and spatially balanced designs.



(a) RMSPE results for samples drawn via eLPM and pLPM with a Linear Kernel function

(b) Ratio of samples drawn from plume in samples drawn via eLPM and pLPM with a linear kernel function

Figure 5.7: Comparison of the ratio of samples drawn from the plume and RMSPE results of samples drawn via equal probability LPM (eLPM) and proportional probability LPM (pLPM) with a linear kernel function after 100 simulation runs of the 48-well random network and complex plume scenario.

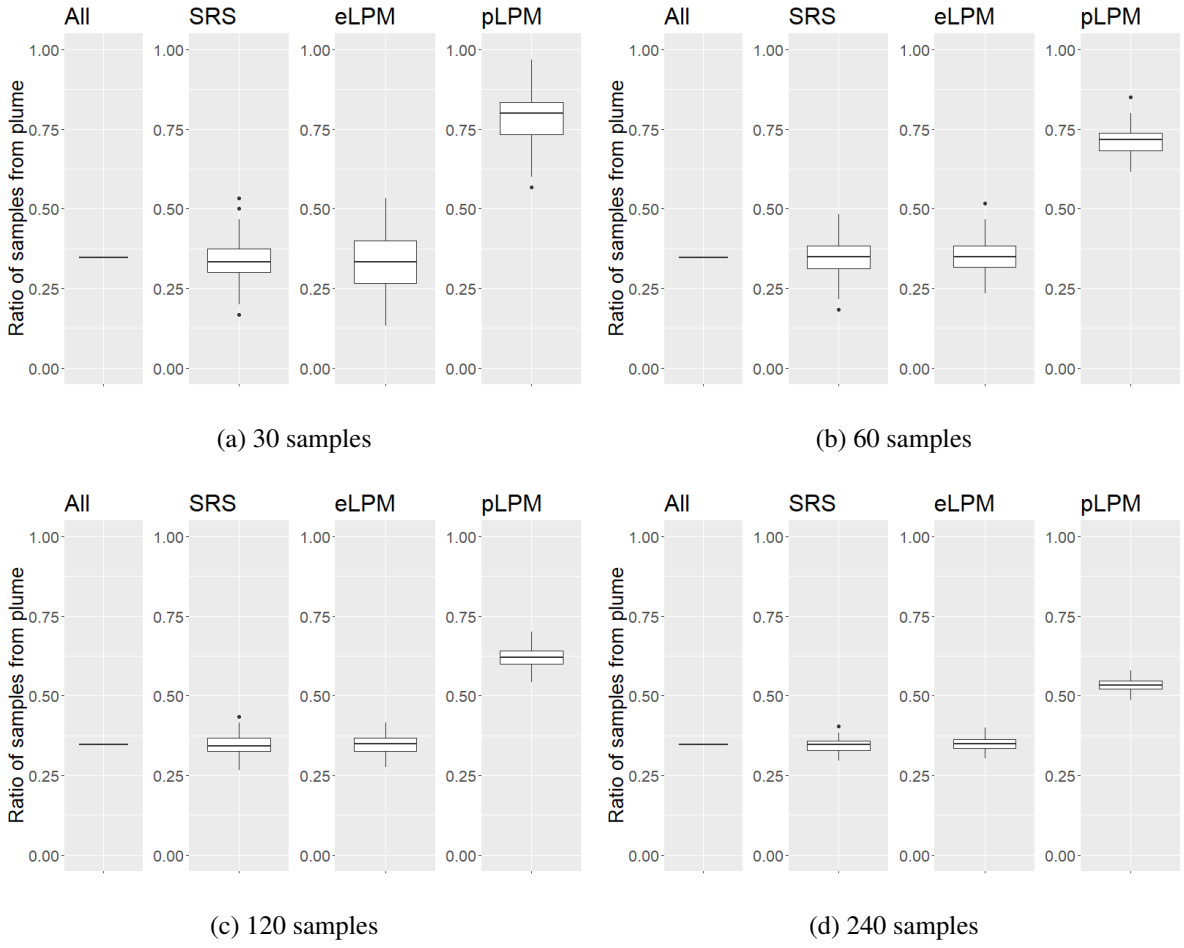
5.3.1 Kernel function selection

As mentioned in Section 5.1.6, using a linear kernel function for determining the tuning parameters for the sample inclusion weights does not induce large enough differences between the weights. This results in sample designs that are nearly indiscernible from equal-weight designs. Figure 5.7a shows a comparison between RMSPEs of equal and linear kernel function-based proportional inclusion weight LPM designs in the 48-well random network and complex plume scenario. The simulations were performed on the moderate and simple plume scenarios as well

with different monitoring networks and number of samples drawn. The results showed no difference between the two designs regardless of scenario.

Figure 5.7b confirmed that the reason why no difference was observed in the RMSPE results is that the linear kernel function-based proportional weight sample designs did not greatly increase the ratio of samples drawn from within the plume compared to the equal weight designs. This indicated that the kernel function was not strict enough to induce appropriate levels of prioritisation towards monitoring wells located closer to the plume. This issue was not present when the half-normal kernel function was implemented as described in Section 5.1.6. The ratio of samples drawn from the plume increased significantly as can be seen on Figure 5.8.

5.3.2 Ratio of samples from the plume



(a) 30 samples

(b) 60 samples

(c) 120 samples

(d) 240 samples

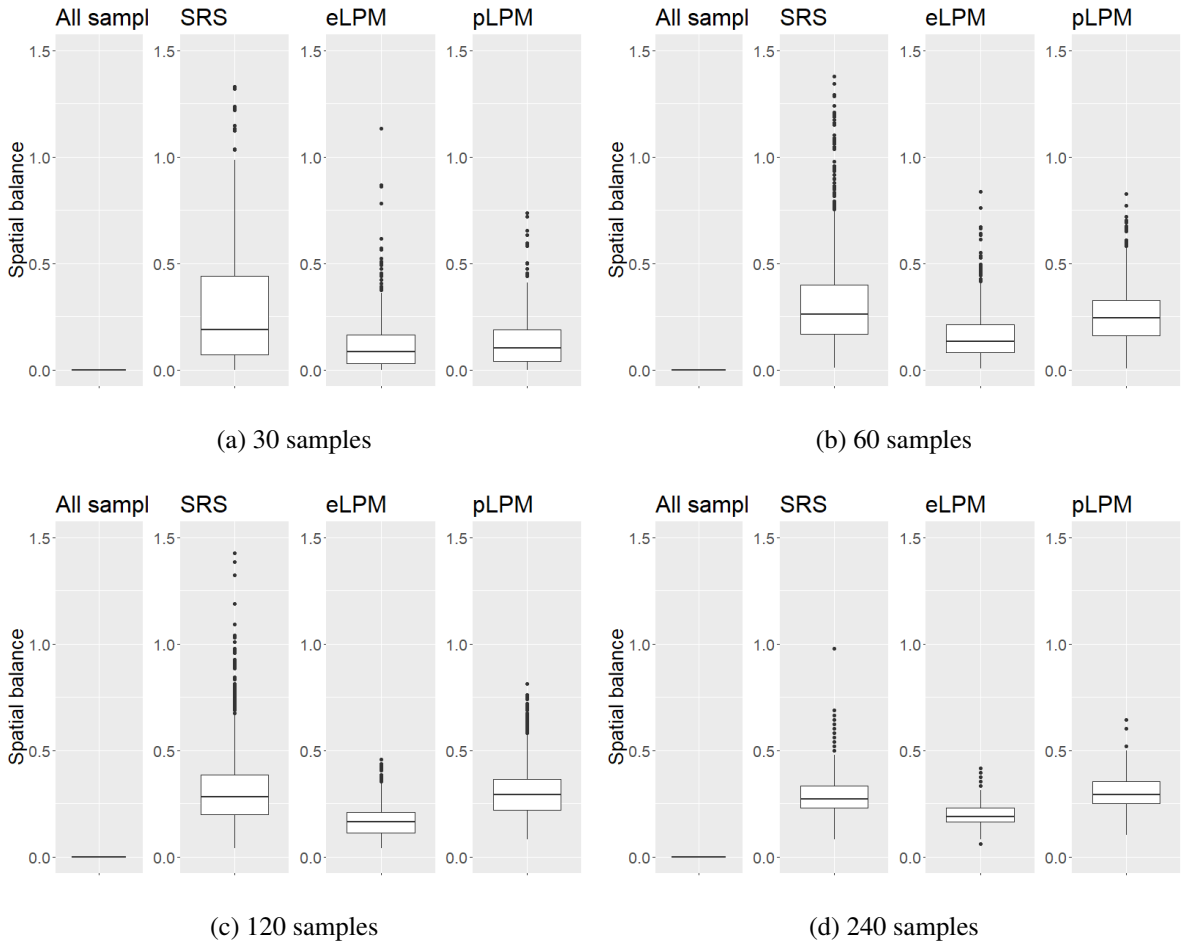
Figure 5.8: Ratio of samples drawn from inside vs outside of the delineated plume area. Comparison of drawing all potential samples and the three investigated sampling approaches, SRS, eLPM and pLPM at 30 (a), 60 (b), 120 (c) and 240 (d) total samples drawn from a potential of 480 (48 wells and 10 sampling times). The results are from the scenario with 48 monitoring wells in random arrangement over the complex contamination plume.

As shown on Figures 5.8a, 5.8b, 5.8c and 5.8d the novel pLPM approach consistently drew a higher percentage of its samples from within the bounds of the CoPC plume (as defined in Section 5.1) than the eLPM with equal inclusion weights and the SRS designs. The figures show the results for a particular scenario: 48 wells in random arrangement over the complex CoPC plume with 30, 60, 120 and 240 total samples drawn from a potential of 480. However, these results were consistent across all other scenarios as well. As the number of samples drawn increases, the variances of the ratios decrease and their value approaches that of the scenario where all potential samples are drawn (see Figure 5.8). This is the case for the pLPM approach as well, albeit requiring even larger sample sizes. These results indicated a strong prioritisation of wells located within the CoPC plume, which meant that: a) the kernel function induced significant enough differences between the sample inclusion weights and b) the GLMs predicted the future well-plume distances sufficiently well.

5.3.3 Spatial balance

Figure 5.9 shows the results of the simulation study performed on the randomised, 48-well network with the complex CoPC plume, for the degree of spatial balance of samples drawn using the investigated methods. On the figure, lower values indicate higher degrees of spatial balance. The degree of spatial balance on average was much higher in equal inclusion weight LPM samples than SRS or pLPM samples regardless of the number of samples drawn. This result was expected, since the objective function of the equal weight LPM algorithm is to reduce this value (see Section 4.4). In 100 simulation runs this method also had the lowest variability and produced relatively few outliers. The variability naturally decreased with increasing number of samples drawn, as the number of possible permutations between selected and unselected monitoring wells also decreased.

The SRS method produced much more variable and on average less spatially balanced results than the LPM designs, but this difference, especially compared to the pLPM approach, decreased significantly at higher sample sizes, where the number of permutations was much lower (see Figure 5.9d). However, the pLPM approach was not expected to produce highly spatially balanced samples, since prioritisation toward certain wells was introduced by tuning the sample inclusion weights. Despite the similar degrees of spatial balance, especially at a sample size of 240, Figure 5.8d shows that the pLPM designs are drawing significantly more samples from within the area of the plume than SRS designs. Moreover, the pLPM design performed similarly to the eLPM design at a sample size of 30 (see Figure 5.9a) despite also drawing more samples from the plume directly (see Figure 5.8a). With larger sample sizes however, the equal weight eLPM produces more spatially balanced samples.



(a) 30 samples

(b) 60 samples

(c) 120 samples

(d) 240 samples

Figure 5.9: Spatial balance of sampling designs created using all potential samples, SRS, eLPM and pLPM at 30 (a), 60 (b), 120 (c) and 240 (d) total samples drawn from a potential of 480 (48 wells and 10 sampling times). A lower value indicates a higher degree of spatial balance. The results are from the scenario with 48 monitoring wells in random arrangement over the complex contamination plume.

It should also be noted that the differences between the three investigated methods would decrease with further increasing samples sizes (beyond 240 samples) until eventually converging on a spatial balance value of zero when all the available 480 samples are selected. Similar results were observed for all other investigated contamination scenarios and monitoring network setups.

5.3.4 RMSPE

Figure 5.10 shows the median RMSPEs of the models corresponding to samples drawn using the investigated methods as a function of the sample size with 95% variability bands observed after 100 simulation runs. These results were obtained using the same scenario as in the previous

sections (randomly arranged network with 48 wells on the complex CoPC plume).

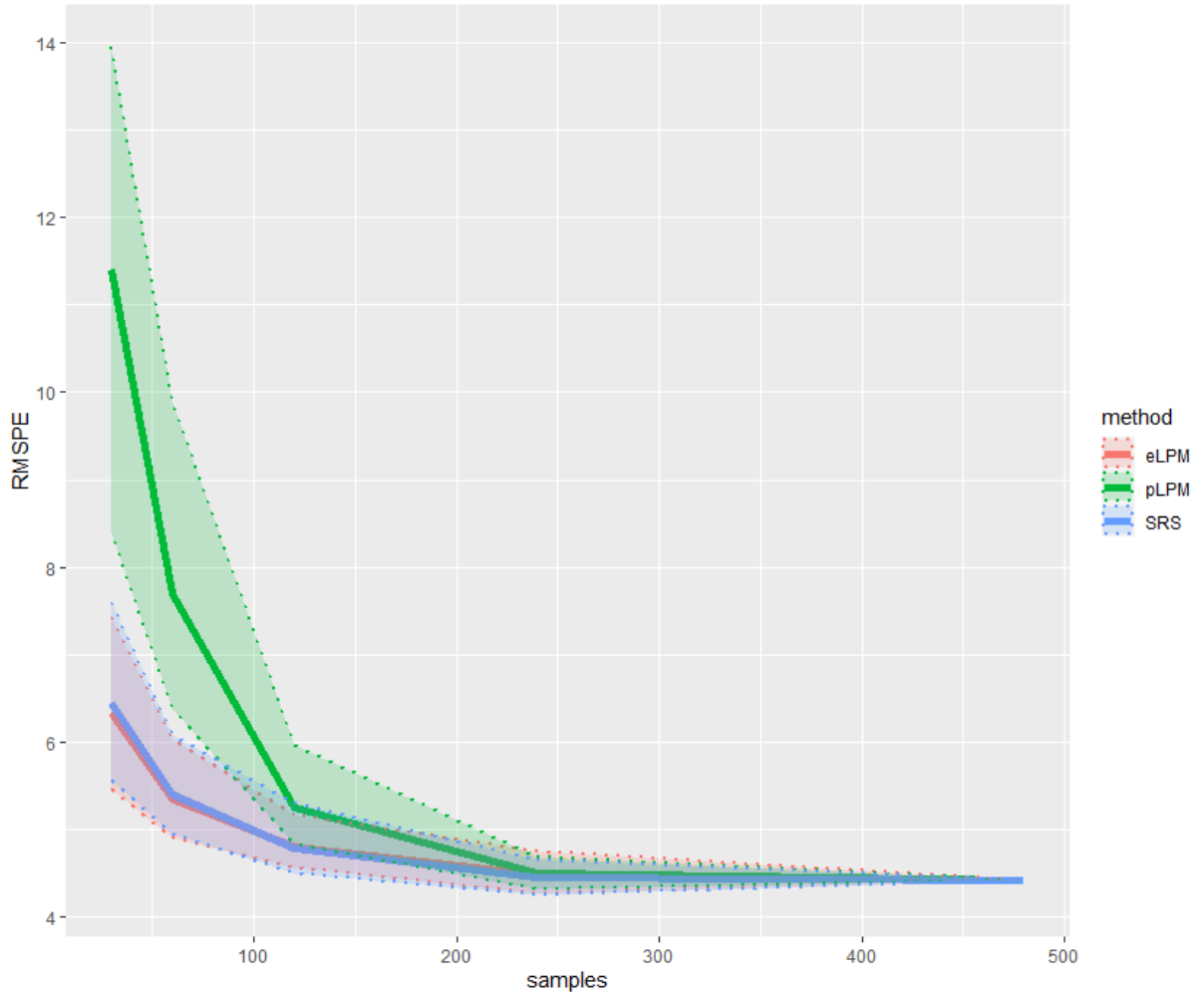


Figure 5.10: RMSPEs of the P-splines models fitted to SRS, eLPM and pLPM-based samples as a function of total sample size from a pool of 480 potential samples (48 wells and 10 sampling times) after 100 simulation runs. The lines represent the median values and the shaded areas represent 95% variability bands. The results are from the scenario with 48 monitoring wells in random arrangement over the complex contamination plume.

The median RMSPE values exhibit an exponential decay pattern, where they decrease with increasing sample size until they converge on the lowest achievable value. This value is achieved when all potential samples are drawn, i.e. each well is sampled at each sampling time. The variability of the results also decreases with increasing sample size.

The results on Figure 5.10 also show that the pLPM-based samples resulted in models with higher RMSPE than the other two methods for sample sizes below 240. As shown on Figure 5.5, the data domain especially at the boundaries is dominated by low concentration values. Since the pLPM approach prioritises sampling wells closer to the plume, depending on the scaling of the kernel function (see Section 5.1.6), it might collect less information on these boundary

regions, which would lead to the ballooning of concentration estimates and could explain the higher RMSPE values in the corresponding models. To confirm this assumption, the RMSPEs in the region delineated by the plume (see Section 5.1.2) were also isolated. Figure 5.11 shows these RMSPEs corresponding only to the region of the domain where the true concentration values exceed the plume threshold value.

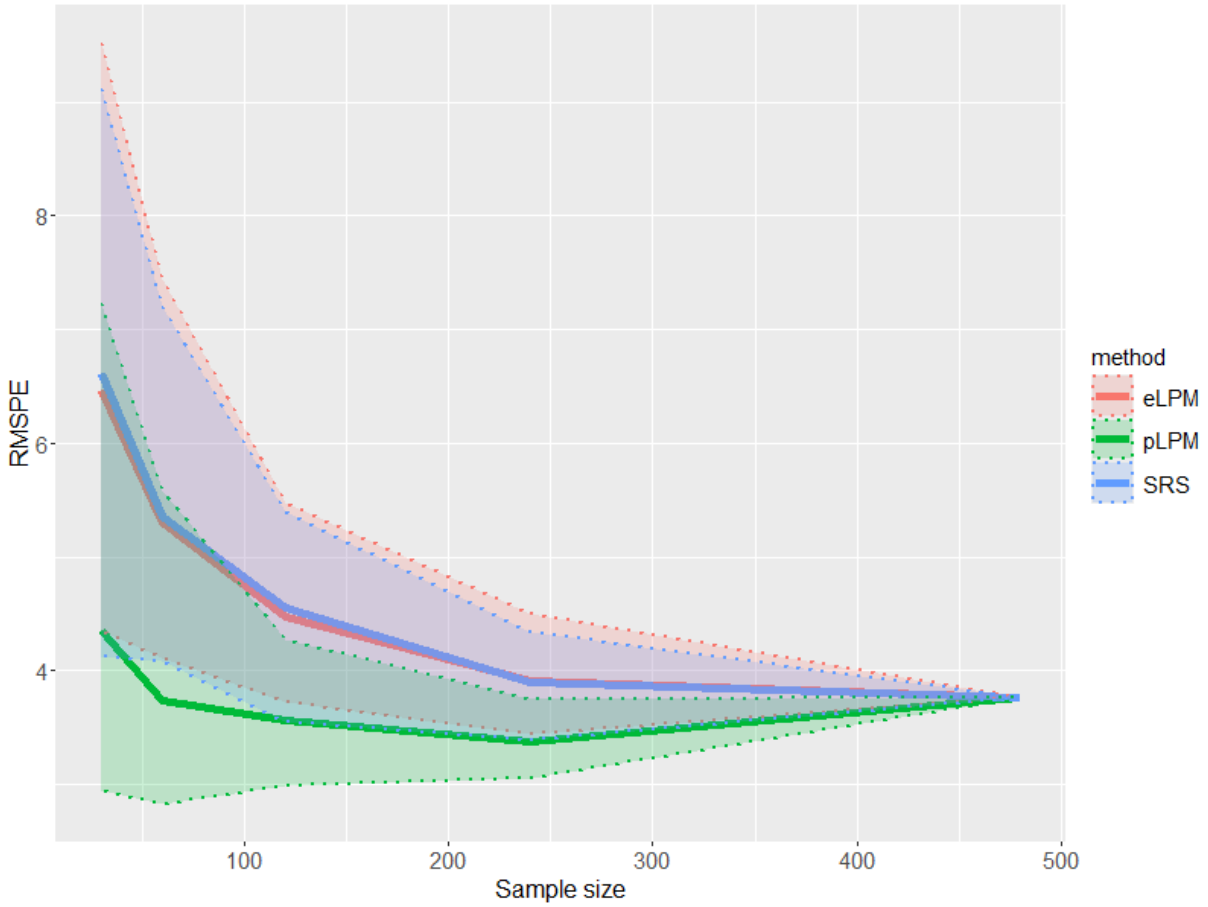
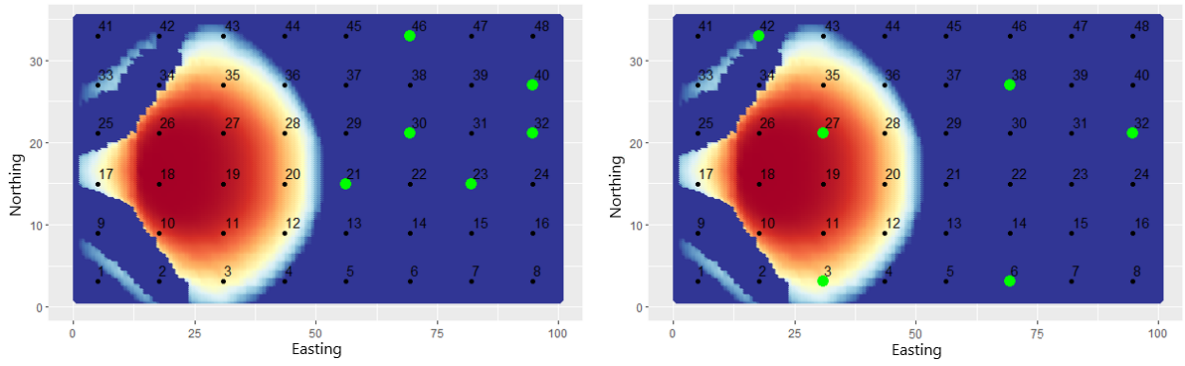


Figure 5.11: RMSPEs from within the plume area as a function of total sample size from a pool of 480 potential samples (48 wells and 10 sampling times) after 100 simulation runs. Based on the P-splines models fitted to SRS, LPM and pLPM-based samples. The lines represent the median values and the shaded areas represent 95% variability bands. The results are from the scenario with 48 monitoring wells in random arrangement over the complex contamination plume.

As Figure 5.11 shows, the pLPM designs performed much better than SRS and LPM with equal inclusion weights (eLPM) when only looking at the area affected by the plume. This seems to confirm the above assumption that the pLPM results considering the whole domain are negatively affected by the lack of information from low-concentration boundary regions (outside the estimated plume), due to the prioritisation of the more central monitoring wells (from within the estimated plume). This aspect of the pLPM design could be an advantage in real scenarios where modelling is performed on the convex hull (see Figure 5.6) where the boundary regions are not

estimated, if the monitoring objective is the more accurate characterisation of the CoPC plume. The slight increase in pLPM-based RMSPE between sample sizes of 240 and 480 could be explained by ballooning effects (McLean et al. [2019]), which arise when neighbouring sample locations exhibiting a steep gradient between estimated solute concentrations are followed by sparsely sampled regions. In these cases, the model continues predicting concentrations along the gradient into the data-sparse region. The results in Figures 5.10 and 5.11 also indicate that the models perform similarly in terms of estimating the solute concentration surface using only about 50% of the observations as using 100%. This suggests that given a sufficient amount of samples, there is no additional gain in useful information as sample size is increased, making additional observations redundant. This assessment is crucial in optimising the sampling effort required for efficient monitoring.

Another key observation on Figure 5.10 is that there is only a small difference in median RMSPEs and variabilities between SRS and eLPM. This result is surprising given the stark difference shown in spatial balance values on Figure 5.9. This indicates that despite this difference in spatial balance, models fitted to SRS samples perform similarly well in terms of estimating the CoPC concentration field. However, a slight divergence can be observed at a sample size of 30, where eLPM-based models performed better. This divergence could potentially widen at even lower sample sizes, as the location of the samples becomes even more influential on the outcome. The lack of difference between the two methods could be explained by the efficacy of the spatiotemporal P-splines modelling approach. Given a sufficient number of random spatial samples over time, the model is able to compensate for the spatial imbalance of the samples. In order to test this assumption, additional simulations were performed, first using spatial models for a single time slice and then incrementally adding more and more time slices and thus extending the model temporally.



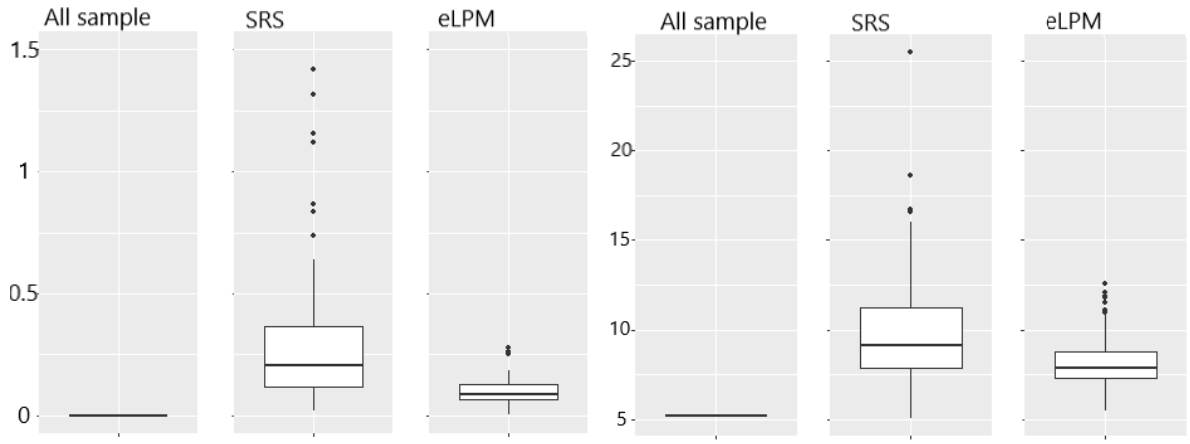
(a) SRS sample

(b) eLPM sample

Figure 5.12: SRS and eLPM samples of size $n = 6$ from the grid-type monitoring network over the log-transformed solute concentration values of the simple CoPC plume.

Figure 5.13 shows the results for the simple spatial models. The results were obtained by modelling the samples from a single event using a small sample size of 6 out of 48 potential samples (given 48 monitoring wells) from the simple plume scenario with a grid-type network arrangement. This was done to minimise the influence of the plume complexity and network type on the outcome, and focus on the correlation between the spatial balance of the samples and the corresponding RMSPE results. Figure 5.12 shows an example of two spatial samples drawn from the network using SRS and eLPM. As the figure shows, using SRS could result in samples that obtain no information on the plume. In contrast, a spatially balanced sample ensures that such information is retrieved.

As the results on Figure 5.13 indicate, in this scenario the difference in spatial balance does translate to the RMSPE results, with the eLPM-based samples providing lower RMSPEs on average with lower variability.



(a) Spatial balance (b) RMSPE

Figure 5.13: Spatial balance (a) and RMSPE (b) results for SRS and eLPM-based spatial sample designs and corresponding spatial models with a sample size of 6. The results were obtained using 48 monitoring wells arranged in a grid pattern over the simple plume scenario.

The positive correlation between spatial balance and RMSPE in this scenario is also shown on Figure 5.14. The results indicate that at this sample size, the SRS design is less likely to select samples that provide enough information to the model than the well spread eLPM samples.

After increasing the sample size from 6 to 12, the difference in spatial balance between SRS and eLPM remains similar, however, the difference in RMSPEs decreases, as shown on Figure 5.15. The median RMSPE of eLPM-based models is still lower than that of the SRS-based ones, but the difference is not as significant as with a sample size of 6.

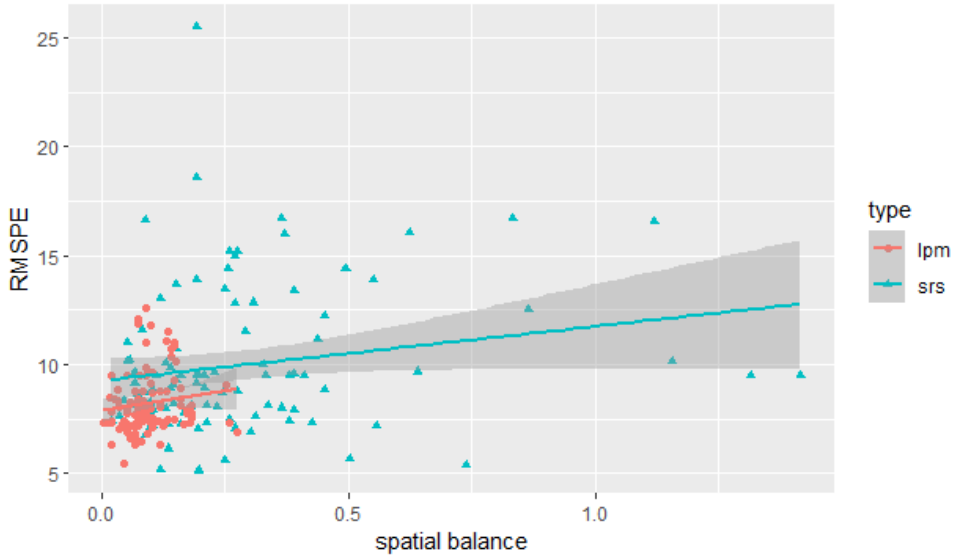
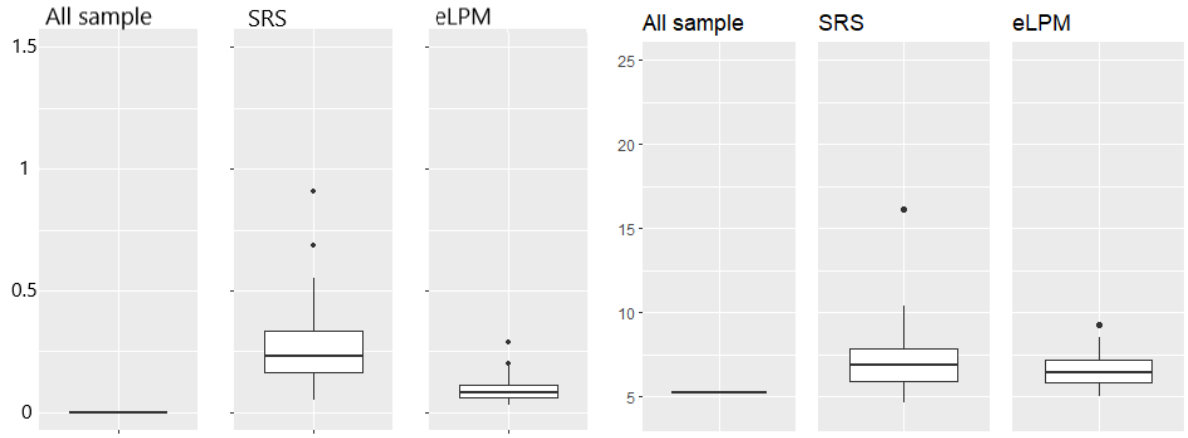


Figure 5.14: Correlation between spatial balance and RMSPE in the simple plume, 48-well grid network scenario with a sample size of 6 using SRS and eLPM for sampling.

The variability in RMSPEs of SRS and eLPM based models also decreased with the increased sample size. However, Figure 5.16 indicates that there is still a positive correlation between spatial balance and RMSPE. These results indicate that at this increased sample size, the SRS method is more likely than before to select samples that provide enough information to the model, resulting in similar predictions as with the eLPM samples.



(a) Spatial balance (b) RMSPE

Figure 5.15: Spatial balance (a) and RMSPE (b) results for SRS and eLPM-based spatial sample designs and corresponding spatial models with a sample size of 12. The results were obtained using 48 monitoring wells arranged in a grid pattern over the simple plume scenario.

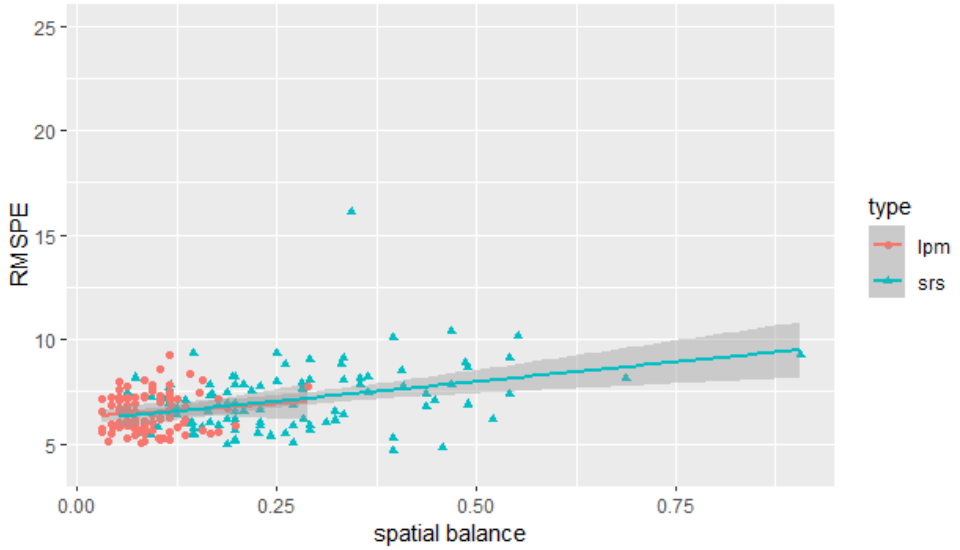


Figure 5.16: Correlation between spatial balance and RMSPE in the simple plume, 48-well grid network scenario with a sample size of 12 using SRS and eLPM for sampling.

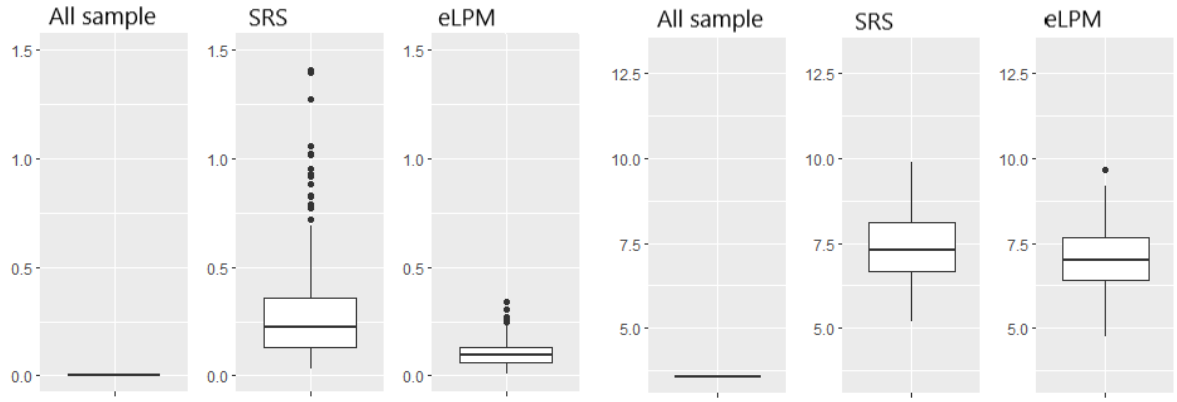
Hence, given a large enough sample size, the higher degree of spatial balance in the eLPM samples does not necessarily result in more accurate models. However, the correlation between RMSPE and spatial balance is still present and SRS produces slightly more designs with high RMSPEs than eLPM. Therefore, collecting well-spread samples using the eLPM design in real cases is more reliable than collecting samples randomly due to the lower variability.

In the next test, rather than increasing the sample size for the spatial model, an additional sampling time was added, drawing 6 samples at each time, and a spatiotemporal model was fitted to the resulting samples. The goal was to assess how the difference between SRS and eLPM changes when the model is extended temporally. The results, shown on Figure 5.17, paint a similar picture as Figure 5.15. The difference in spatial balance does not carry over to the RMSPE values to the same extent, indicating that increasing sample size in time has a similar effect as increasing it in space. Given a small spatial sample size but introducing more sampling events means that the SRS design can draw enough samples over time from different monitoring wells to provide sufficient information to the model.

However, as observed on Figure 5.18, the previously positive correlation between spatial balance and RMSPE does not appear in this scenario for SRS and appears for eLPM to a lesser extent, potentially indicating a difference between spatial and spatiotemporal models.

In order to see if the lack of positive correlation is caused by some aspect of the spatiotemporal model, the test was repeated with the restriction that the same wells had to be sampled at both sampling times. This meant that the same spatial sampling design was used at both times, thus eliminating the possibility of introducing additional information from other sampling locations

over time.



(a) Spatial balance (b) RMSPE

Figure 5.17: Spatial balance (a) and RMSPE (b) results for SRS and eLPM-based sample designs for two sampling events with a sample size of 6 for each, and corresponding spatiotemporal models. The results were obtained using 48 monitoring wells arranged in a grid pattern over the simple plume scenario.

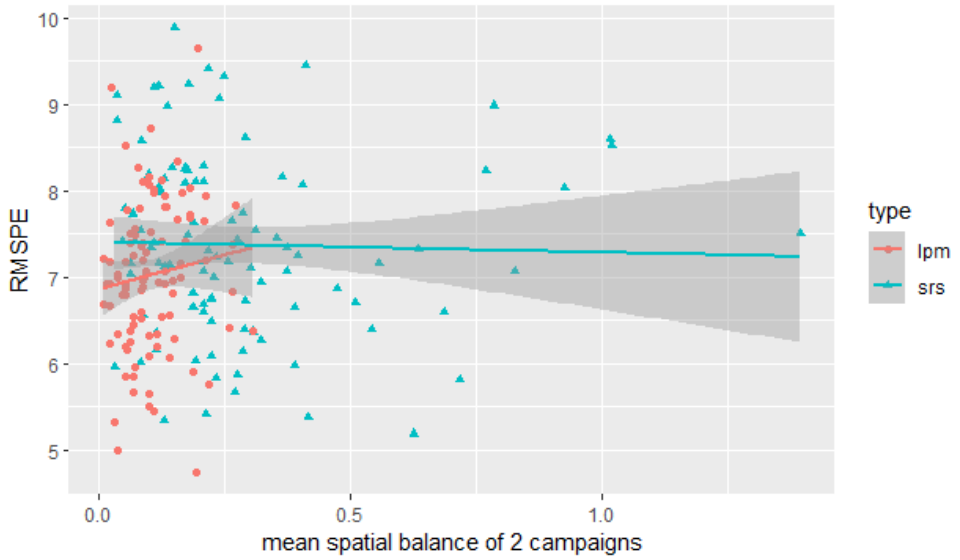


Figure 5.18: Correlation between spatial balance and RMSPE from 2 sampling events in the simple plume, 48-well grid network scenario with a sample size of 6 for each event using SRS and eLPM for sampling.

As Figure 5.19 shows, the results once again exhibit a slightly positive correlation overall for the SRS designs. This indicates that the introduction of additional sampling times with new spatial sample designs could result in the diminishing of the correlation between spatial balance and RMSPE values.

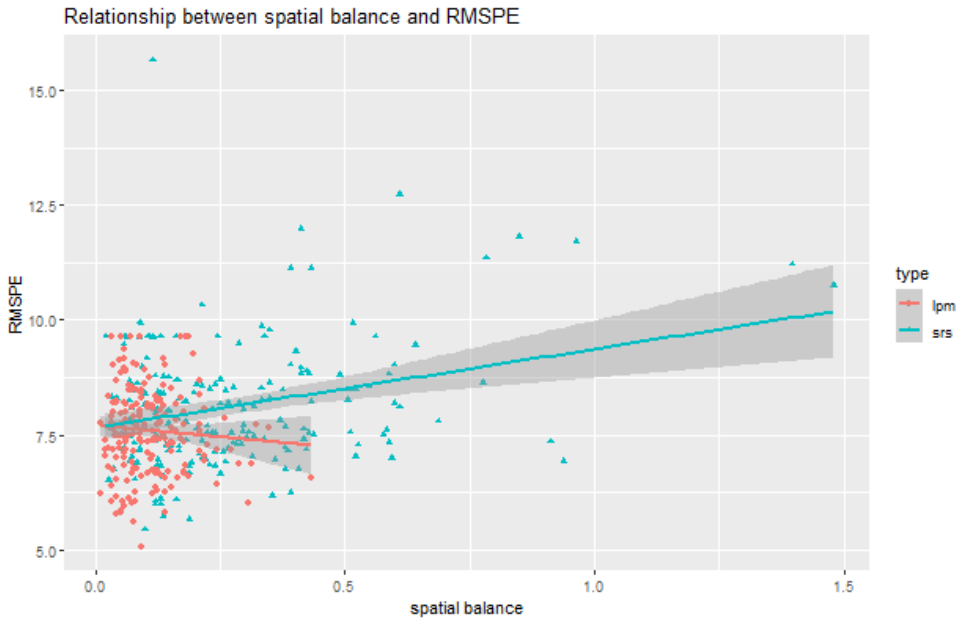


Figure 5.19: Correlation between spatial balance and RMSPE from 2 sampling events in the simple plume, 48-well grid network scenario with a sample size of 6 for each event, with the same sampling locations used in both events. The sampling locations were selected using SRS and eLPM.

5.3.5 Plume Mass

Figure 5.20 shows the ratio between the estimated plume mass and the true plume mass on a logarithmic scale as a function of the sample size on the simulation conducted on the randomly arranged, 48 well network over the complex CoPC plume. The shaded areas represent 95% variability bands. The line at $y = 1$ represents situations where the estimated plume mass equals the true plume mass. Results below the line underestimate the plume mass whereas results above the line overestimate it. The y-axis indicates the degree of over- or underestimation. For example a result at $y = 3$ would represent a model in which the estimated plume mass is three times as large as the true plume mass.

Figure 5.20 indicates large variability at lower sample sizes, especially for eLPM and SRS designs, that decreases with increasing sample size. The median plume mass estimates and variability of the pLPM sample based models stay much closer to the true value than the samples drawn using the other two methods until the maximum sample size is reached.

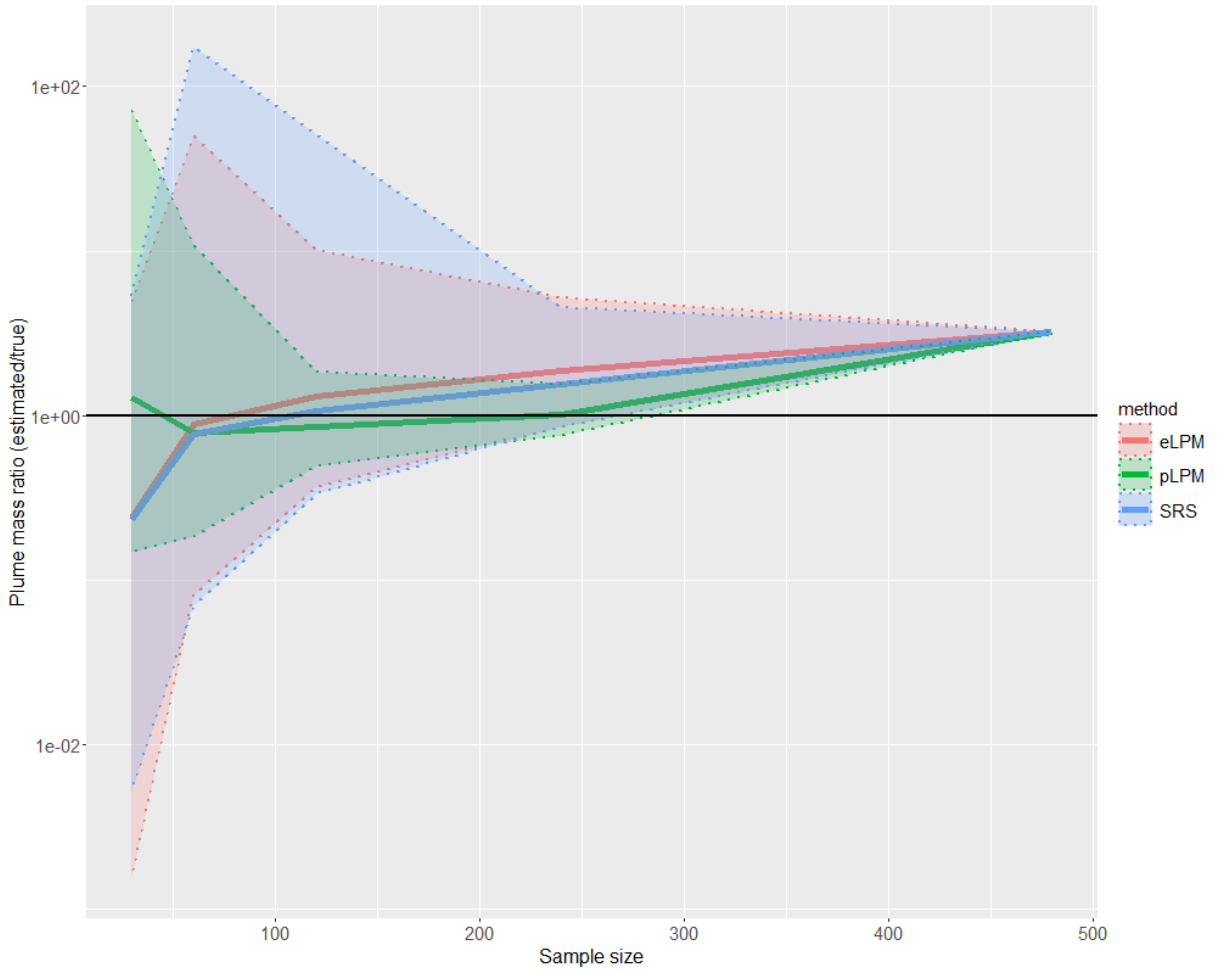


Figure 5.20: Estimated plume mass over true plume mass as a function of sample size in the randomly arranged, 48 well network with complex plume scenario. The y-axis is on a logarithmic scale and the line at $y = 1$ represents equal estimated and true plume mass. The shaded areas represent 95% variability bands.

SRS and eLPM based models exhibit an increasing trend, where on average, they underestimate the plume mass with smaller sample sizes (n=30 and n=60) and overestimate it above that (n=120 and n=240). At a sample size of n=120, the three methods produce similar estimates on average but the variability of the pLPM-based models is much lower. The overestimation at maximum sample size (n=480) compared to lower sample sizes could once again be attributed to the ballooning of concentration estimates in the P-splines models. As the model is fitted to an increasing number of observations, there are bound to be close neighbours among them, which ultimately leads to ballooning in more sparsely sampled regions. At a sample size of 240, the pLPM-based models on average estimate the plume mass to a high degree of precision, with a median plume mass ratio of 1.01, a lower quartile of 0.9 and an upper quartile of 1.1 as shown on Figure 5.21, which shows a cross section of the results at this sample size. The pLPM method in this case has a significantly lower variability compared to eLPM and SRS, which tend to overestimate the plume mass.

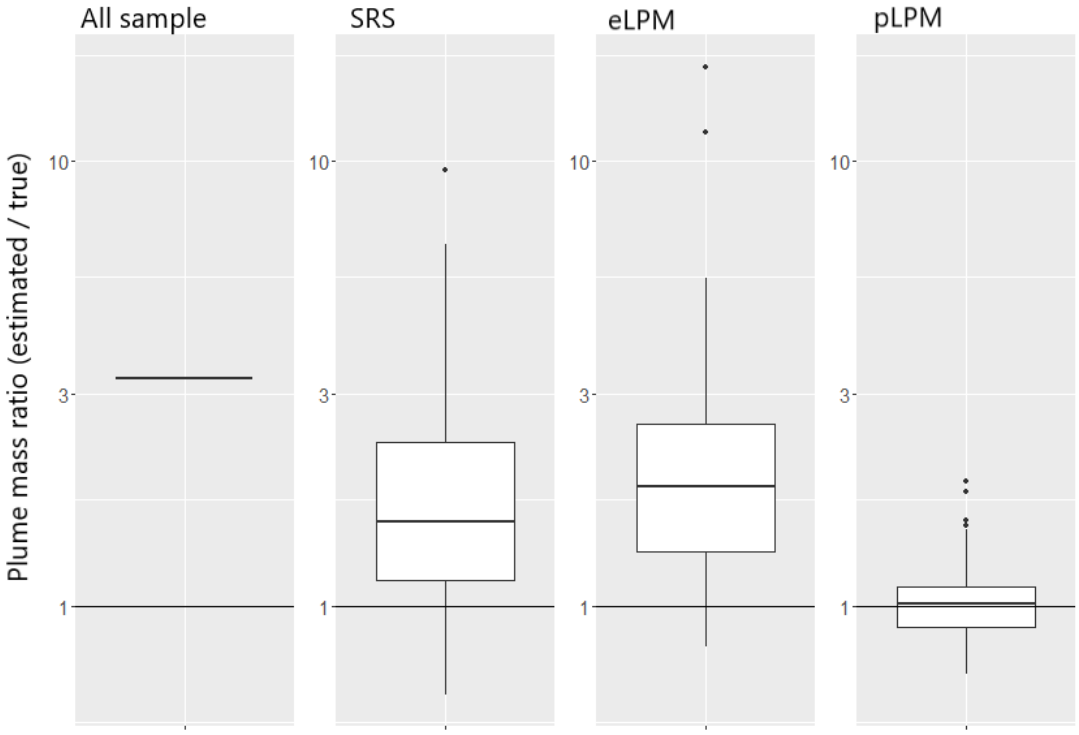


Figure 5.21: Box plots of estimated plume mass over true plume mass in the randomly arranged, 48 well network with complex plume scenario at sample size n=240. The y-axis is on a logarithmic scale and the line at $y = 1$ represents equal estimated and true plume mass.

5.3.6 Convex Hull

As mentioned before, real groundwater quality data is commonly modelled on the convex hull outlined by the monitoring wells (see example on Figure 5.6). Therefore, to ensure that the results obtained by considering the whole domain are representative of the convex hull as well, the RMSPE and plume mass ratio results were compared in these two cases. Figure 5.22 shows the RMSPE results from the same scenario as 5.10 but these were obtained by considering the convex hull area only. The results exhibit a similar, exponentially decaying trend. The pLPM-based median RMSPE values are somewhat lower at a sample sizes of 30 and 60 in the convex hull case, while the SRS and eLPM-based values are higher and the results converge on a slightly higher value here than in the whole domain case. However, the difference is not substantial. Overall, very similar results were obtained in both cases.

Figure 5.23 shows the RMSPE results isolated for the plume area only in the convex hull case. Here, there is no indication of difference compared to the whole domain case (see 5.11).

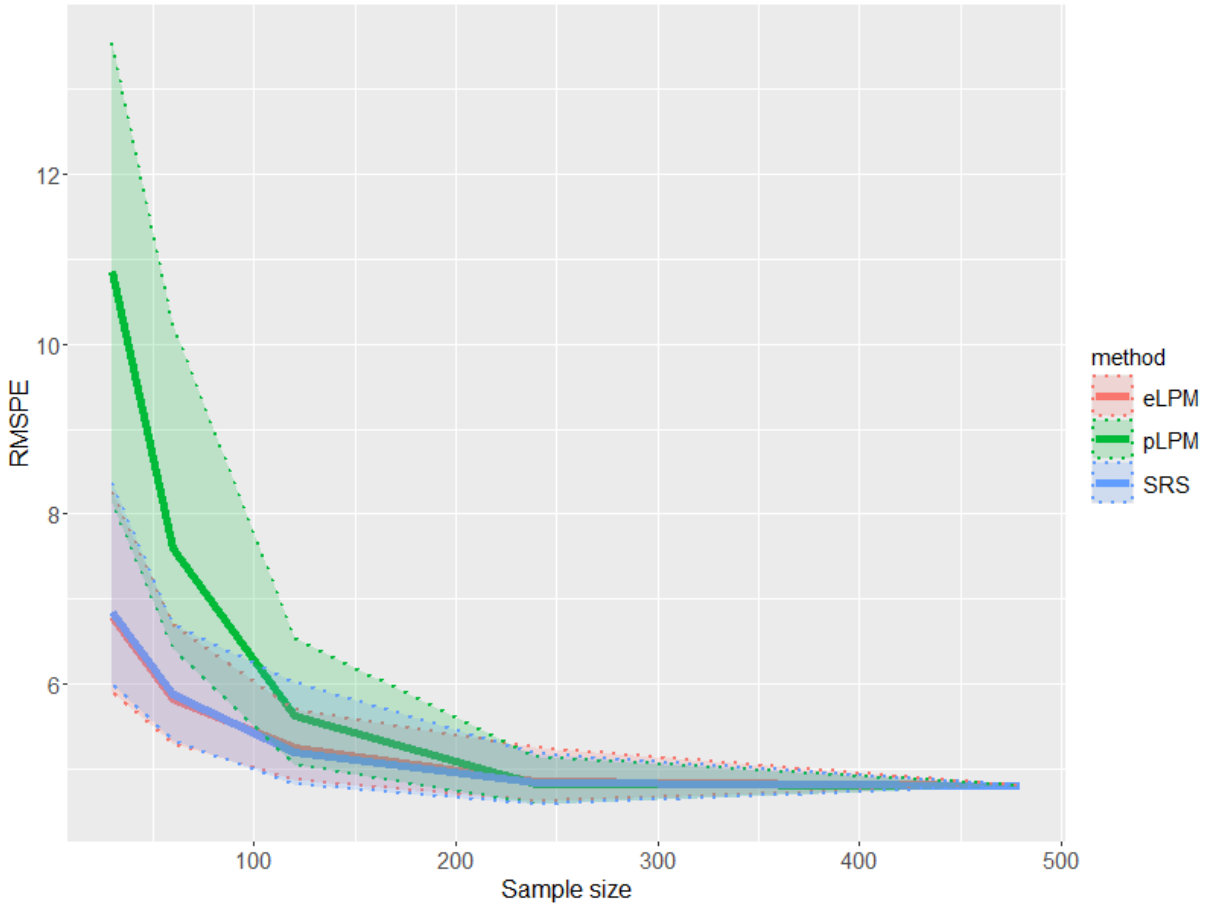


Figure 5.22: RMSPEs of the P-splines models fitted to SRS, LPM and pLPM-based samples as a function of total sample size from a pool of 480 potential samples (48 wells and 10 sampling times) after 100 simulation runs. The lines represent the median values and the shaded areas represent 95% variability bands. The results are from the convex hull of the scenario with 48 monitoring wells in random arrangement over the complex contamination plume.

The plume mass results on Figure 5.24 also show little difference compared to the whole domain case (see 5.20). The median and variability of the SRS, eLPM and pLPM designs show similar patterns and magnitudes indicating that the convex hull encloses the estimated plume well. The convex hull of a less spatially well spread monitoring network that does not capture the full extent of the plume is expected to induce larger differences in plume mass estimation.

Overall, similar results were obtained regardless of whether the whole synthetic data domain or only the convex hull was taken into account. It should be noted that the models are fitted to the observations obtained via the different sampling designs and therefore, they are based on the same amount of information. The difference between considering the whole domain and the convex hull is that the former involves extrapolation as well to obtain predictions outside the convex hull.

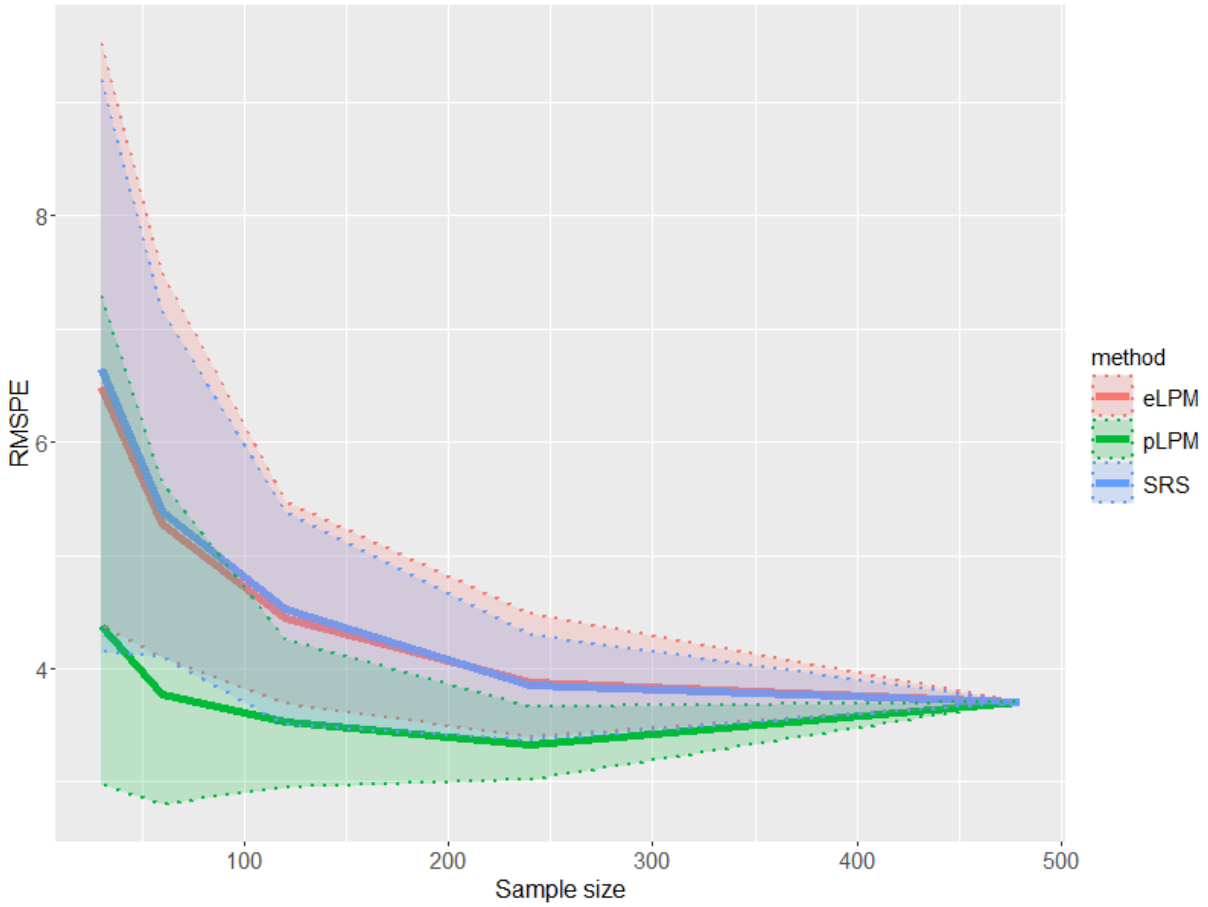


Figure 5.23: RMSPEs from within the plume area as a function of total sample size from a pool of 480 potential samples (48 wells and 10 sampling times) after 100 simulation runs. Based on the P-splines models fitted to SRS, LPM and pLPM-based samples. The lines represent the median values and the shaded areas represent 95% variability bands. The results are from the convex hull of the scenario with 48 monitoring wells in random arrangement over the complex contamination plume.

Hence, for the RMSPE results, it is a matter of which data points are used in the calculation (see Section 5.2.4). This is why no difference can be observed between the plume-specific RMSPEs (see Figures 5.11 and 5.23), since all corresponding data points are located within the convex hull. It can be stated that the difference is mostly driven by the monitoring network arrangement, since it determines how much of the investigated site is omitted, i.e. how much extrapolation is required to obtain predictions for the whole site. In this presented case, the extrapolated predictions were reasonable and thus they did not affect the RMSPE and plume mass results substantially.

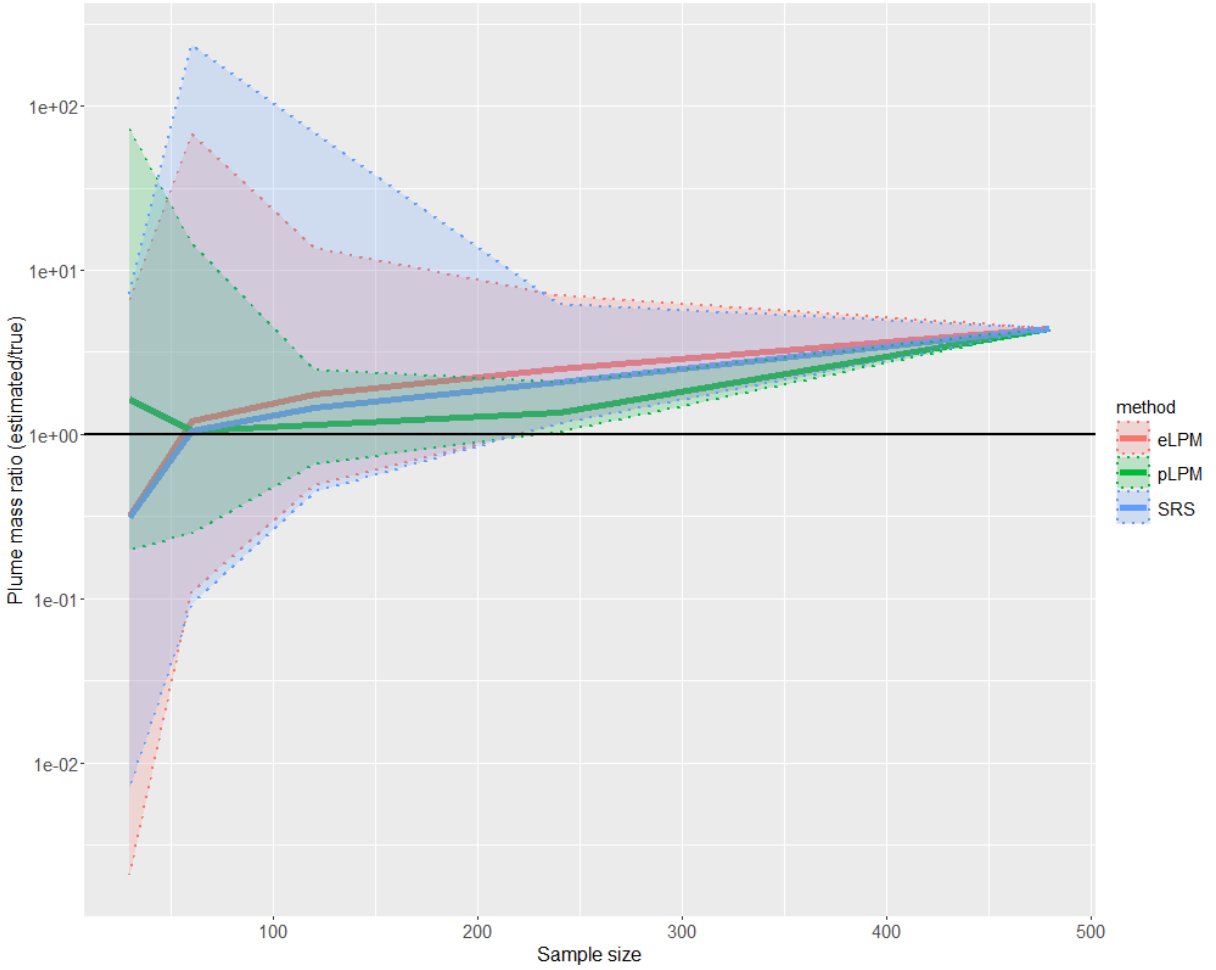


Figure 5.24: Estimated plume mass over true plume mass as a function of sample size in the convex hull of the randomly arranged, 48 well network with complex plume scenario. The y-axis is on a logarithmic scale and the line at $y = 1$ represents equal estimated and true plume mass. The shaded areas represent 95% variability bands.

5.4 Conclusions and Future Work

In conclusion, the pLPM sampling design shows promising results and could potentially provide benefits in groundwater contamination monitoring, especially in the estimation of plume characteristics. By tracking the evolution of the plume in space and time using historic data, pLPM designs were able to estimate CoPC concentrations within the plume and the plume mass with a higher degree of precision than simple random sampling and spatially balanced sampling designs. Using pLPM to draw samples did result in higher RMSPEs on average when the whole site was considered potentially due to ballooning effects, but the difference disappeared with large enough sample sizes. At the same time, the prediction errors of pLPM designs within the plume remained lower than the other two methods regardless of sample size (see Figure 5.11).

Hence, given a sufficiently large sample size, the pLPM approach resulted in the same prediction accuracy over the whole domain and provided additional precision for the estimation of concentrations within the plume. Furthermore, the ability to control well prioritisation strength through the kernel function means that the method can be fine-tuned to strike an appropriate balance between more precision within the plume and less error overall. The freedom to customize or choose different kernel functions is also an attractive, useful property of this approach that makes it generalizable to other environmental sampling targets, where historic observations are available. It allows for the fine-tuning of the sampling design based on various features of the target populations, not limited to the distance from sampling location.

The results also indicated that the applied spatiotemporal P-splines modelling approach is fairly insensitive to the sampling design as long as the collected samples cover a sufficiently large area and time period. This is highlighted by the similarities in results based on the SRS and eLPM designs, despite their differences in the achieved degree of spatial balance. Given a sufficient number of sampling events or sample sizes, the random spatial samples obtained via SRS were adequate enough to produce similar estimates as eLPM designs. Substantial differences in the results from pLPM-based samples were observed compared to SRS and eLPM, because they inherently introduce a certain bias towards sampling locations near the CoPC plume. Also notably, the results suggested that in the investigated scenarios, using about 50% of the observations can provide similar precision in estimating the solute concentration surface as using 100% of the data. This substantiates the results obtained by McLean et al. [2019] and is a testament to the strength of a spatiotemporal modelling framework.

Despite the insensitivity of the modelling approach, it is still recommended to use an LPM-based sampling design rather than SRS when planning for an upcoming sampling event. That is because, as the results have shown, the higher degree of spatial balance does have an impact on the precision of predictions in spatial models. LPM-based designs produce more reliable and spatially balanced samples in general. Therefore, when drawing samples one after another over a certain time period, the chance of drawing multiple non-informative samples is lower if LPM-based designs are used.

Future work on the pLPM sampling design should focus on the scaling of the the kernel function, to identify the most appropriate prioritisation strength and ensure balance between precision in estimating concentration levels within and at the boundaries of the plume. The approach should also be tested on different scenarios and case study data-sets to build a strong empirical base for its application. Notably, as mentioned in Chapter 4.5, Prentius and Grafström [2024] have recently proposed the local spatial balance measure, which is based on capturing the balance between sample units and their neighbours. The new measure can differentiate between sampling designs with similar Voronoi partitions, since it measures balance within the polytopes.

This measure could also be investigated in the research presented in this chapter to substantiate the results obtained regarding the differences in spatial balance between the different sampling methods. The sampling design approach proposed in this chapter is convenient for potential future implementation in groundwater data analysis tools due to its data-driven nature and the computational efficiency of the LPM2 sampling technique (see Chapter 4.4.4). It could be used to propose future sampling designs on the basis of historic observations that could provide optimal data to support subsequent data analyses.

The following chapter will investigate another aspect of sampling designs, which concerns the difference between adjusting sampling frequency or the number of samples collected at a time with a given total number of samples.

Chapter 6

Evaluation of Sampling Intensity, Frequency and Sample Size Using Balanced Designs

In this chapter, consideration will be given to practical aspects related to the implementation of the spatiotemporal, stochastic sampling designs introduced in the previous chapters with respect to long-term groundwater contamination monitoring. The first section will explore an innovative approach on how spatiotemporally balanced sampling designs could be used in groundwater contamination modelling applications such as GWSDAT (1.3), to inform future sampling designs based on the analysis of historic sampling intensity. The second section will explore how different combinations of the number of sampling events and the number of samples collected per event in a given monitoring period affect the subsequent model predictions. Similarly to Chapter 5, the variation induced by the randomness in the balanced sampling designs was characterised via the variability in the RMSE and RMSPE values of the resulting model estimates in the simulation study.

6.1 Evaluating Historic Sampling Intensity Using Spatially Balanced Designs

This section will explore the potential implementation of spatiotemporally balanced sampling designs in groundwater contamination modelling applications such as GWSDAT, with the aim of optimising future sampling designs. The spatiotemporally balanced designs can be used to evaluate the sufficiency of the historic sampling intensity from the perspective of estimating con-

tamination surfaces, and potentially generate new sampling schemes for the future based on the results. The method works by drawing increasingly smaller sub-samples of the complete historic data set using the eLPM spatiotemporally balanced sampling design technique, and comparing the predictions made by the models fitted to these sub-samples. Historic groundwater monitoring data sets contain CoPC concentration values obtained from monitoring wells over a period of time. In this implementation, the samples were drawn from the 3-dimensional space of observations, made up of two spatial dimensions encoding the monitoring well coordinates and a dimension representing the time of the observation. This approach to distributing samples in time was discussed in Section 5.2.1. Ultimately, the aim of the analysis is to check if a reduced number of samples, as if drawn by a spatiotemporally balanced method, would have produced the same predictions as the original, complete data set. For example, if using 75% of the observations results in sufficiently similar predictions as using 100% of them, then the sampling intensity could be reduced in the future by using a spatially balanced sampling design. On the other hand, if sufficiently dissimilar predictions are obtained when using a reduced number of samples, that could suggest that the contaminated site is undersampled. In this case, sampling intensity could be increased in the future.

6.1.1 Simulation Study

In this study, the method was tested on both synthetic and case study groundwater CoPC concentration data. The synthetic data and monitoring networks used were the same as in the previous chapter, described in 5.2.2 and 5.2.3 respectively. The synthetic groundwater contamination data included three hypothetical CoPC plumes with increasing geometric complexity. These were introduced in Chapter 1.4.2, and a visual representation can be seen on Figure 3.1. The monitoring networks differed in the spatial arrangement of the wells and the number of wells. As highlighted in Chapter 5.2.3, the randomly arranged network was the closest representation to real groundwater monitoring networks. Figure 5.5 shows the monitoring network arrangements. The models of the synthetic data were first compared by the RMSPEs (see Section 5.2.4) based on the predicted CoPC concentration surfaces over the convex hull (see Section 5.2.3) at the sampling times. The aim of analysing the synthetic data this way was to get a baseline understanding of the relationship between the size of the sub-sample and the RMSPE of the resulting model. Figure 6.1 shows a flowchart of the comparison between the complete set of observations and the sub-samples using the synthetic contamination data. The complete set of observations was defined as the CoPC concentration levels from the synthetic data at the coordinates of the monitoring wells taken at each sampling time. The P-splines model was fitted to these observations to estimate the concentration surfaces. The true concentration surfaces were used to calculate the RMSPE of the model over the convex hull of the monitoring network.

The eLPM sampling technique was used to draw spatiotemporally balanced sub-samples of the observations. The sample sizes ranged from 5% to 95% of the total number of observations, increasing by 5% each time. This resulted in a total of 19 sub-samples. The P-splines model was fitted to each of the sub-samples to estimate the concentration surfaces and subsequently calculate the corresponding RMSPEs.

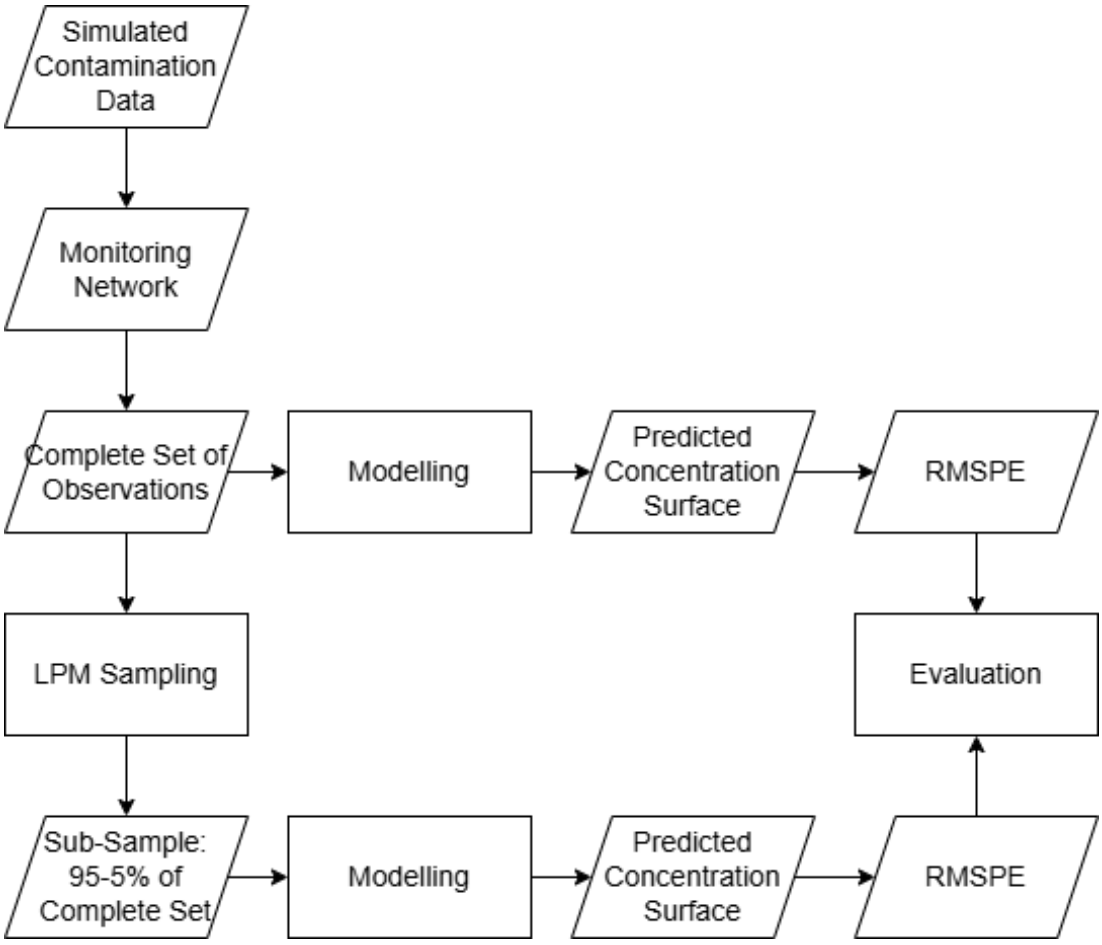


Figure 6.1: Flowchart of the method to evaluate sampling intensity in the scenarios using synthetic groundwater contamination data via RMSPE.

However, in real data applications the RMSPEs as defined in Section 5.2.4 cannot be used as evaluation metrics, because the true concentration surface is unknown. Therefore, a reasonable proxy should be used to evaluate the sub-samples in real-data applications. In this study, the RMSEs between the observed and predicted concentration values of the unselected observations were tested for this purpose. Thus, the RMSEs evaluate how well the models based on the sub-samples can predict the observations that were not selected from the complete set. For the synthetic data set, both the RMSPEs and the RMSEs were calculated to evaluate whether the RMSE is a good proxy metric. Consequently, the RMSE was used to evaluate the method when using case study data. Figure 6.2 shows a flowchart of how the LPM-based sub-samples were analysed using the above described RMSE metric. The sub-samples were drawn from

the complete set of observations using the eLPM approach. The sub-sample sizes again ranged from 5% to 95% of the observations increasing by 5%, resulting in a total of 19 sub-samples. Spatiotemporal P-splines models were used to predict the concentration levels of the data points not selected by the eLPM sampling. Using the actual observations and the predicted values, the RMSE was calculated. For every investigated synthetic or case study scenario, 100 simulation runs were performed. The stochastic nature of the eLPM sampling approach ensured variation in the selected sub-samples among the simulation runs.

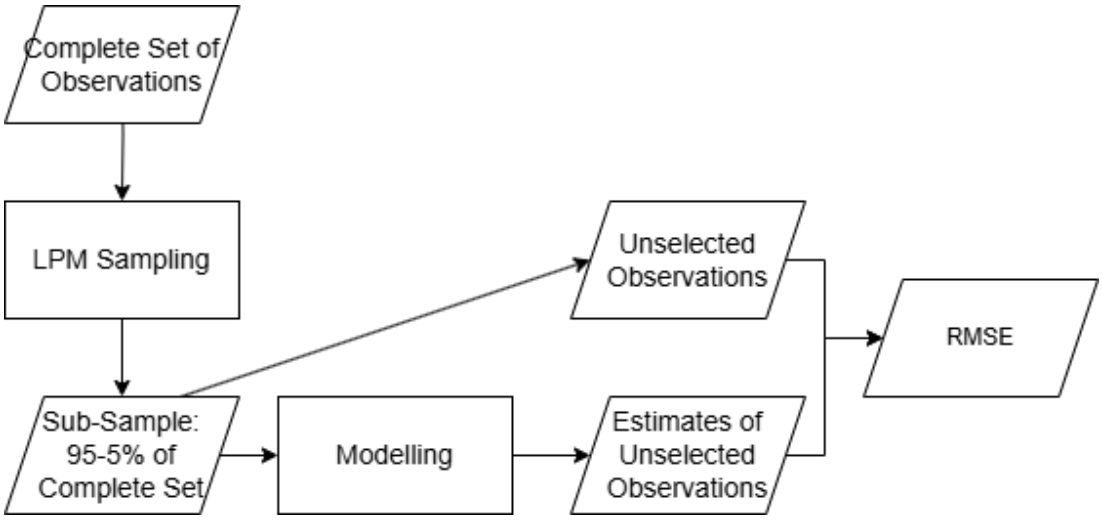


Figure 6.2: Flowchart of the method to evaluate past sampling intensity in the scenarios using case study groundwater contamination data via the RMSEs of unselected observations.

The case study groundwater contamination data set is publicly available as an example in the software GWSDAT (Jones et al. [2014]), and was introduced in Chapter 1.4.4. The data come from the long-term monitoring of a decommissioned petrol station. It contains the concentration measurements of five different CoPCs in groundwater samples from 32 existing monitoring wells collected over a 4-year period. The five solutes monitored were ethylbenzene, toluene, total petroleum hydrocarbons (TPH), nitrate and sulphate. There were a total of 382 observations of ethylbenzene concentrations, which were used in the simulation study. Figure 2.2 shows the observations of concentrations collected from the monitoring wells, while Figure 6.3 shows estimated ethylbenzene concentrations and groundwater levels at the final time point of the monitoring period via the P-splines model introduced in Chapter 1.3.1, around the decommissioned petrol station as well as the layout of the monitoring wells.

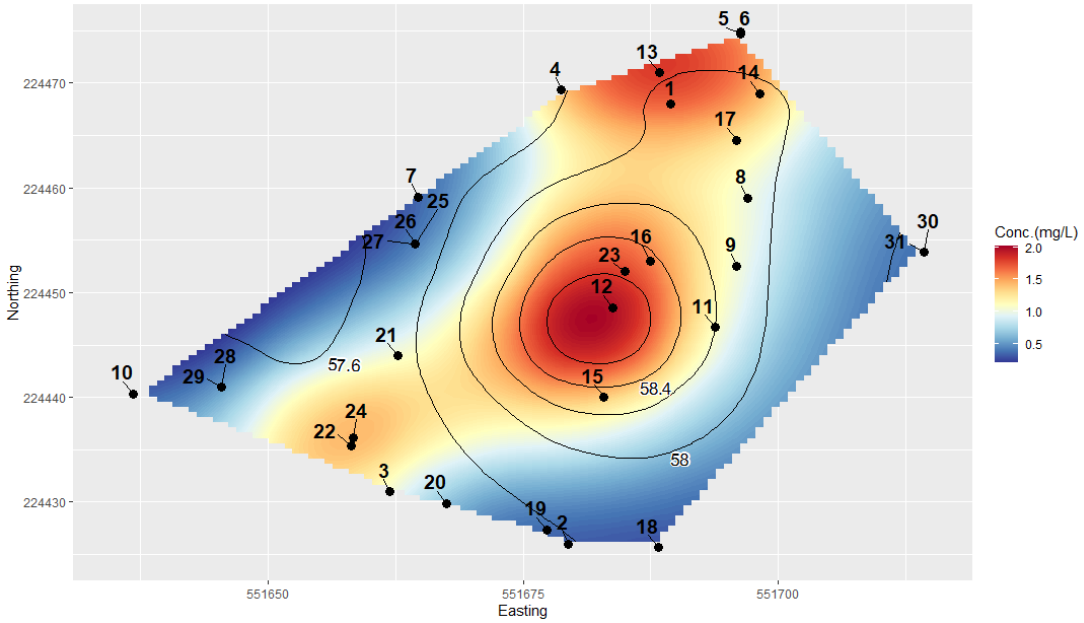
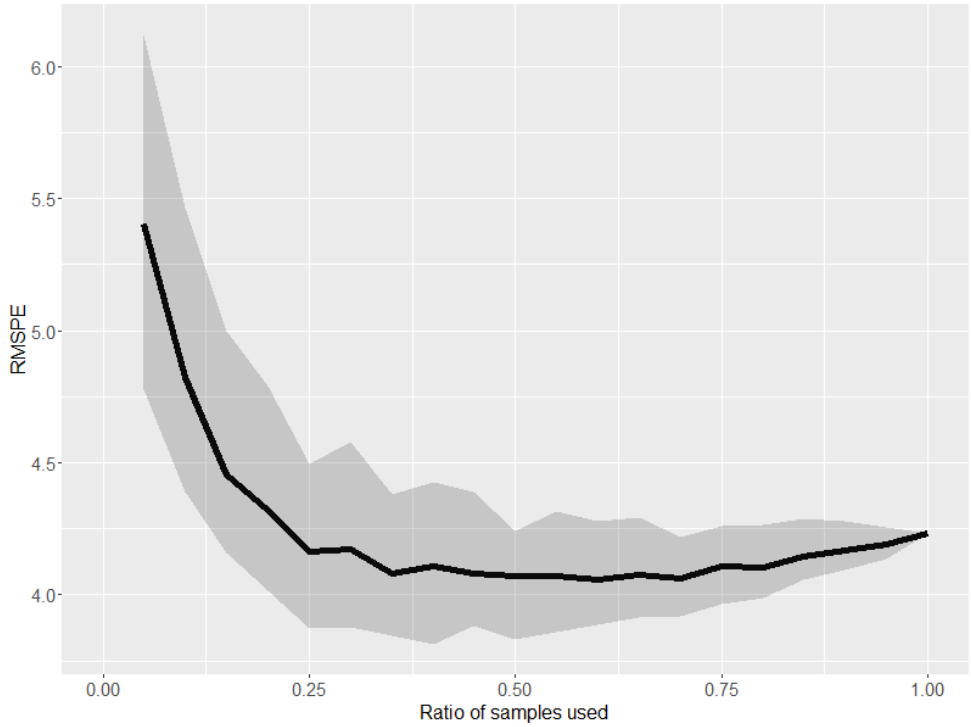


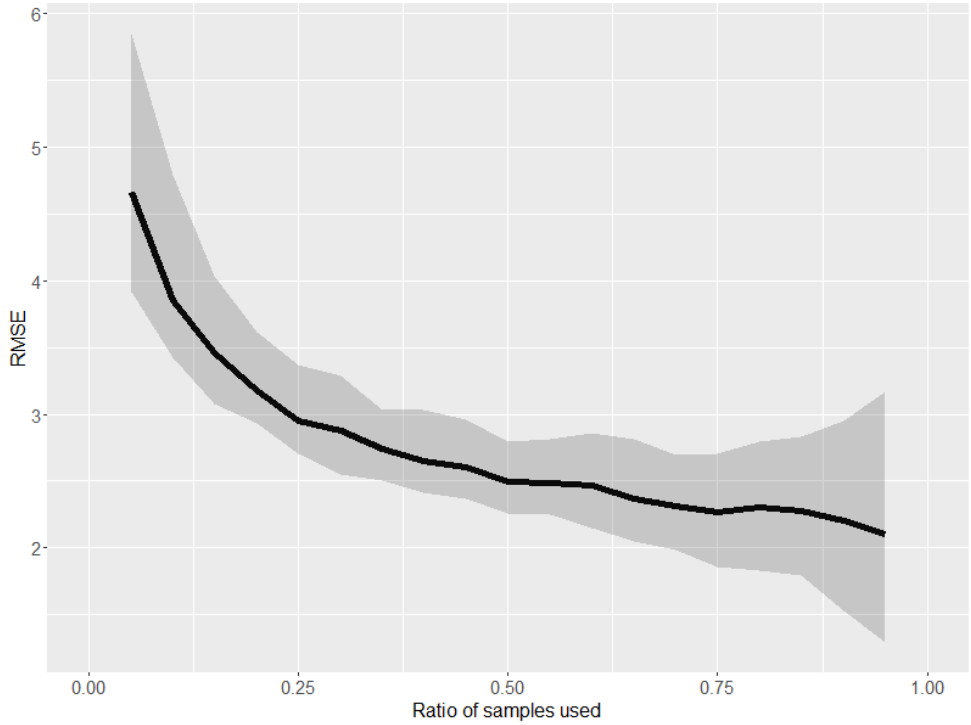
Figure 6.3: Ethylbenzene concentration surface and groundwater elevation contour lines at the end of the monitored period predicted via spatiotemporal P-splines.

6.1.2 Results

The RMSPE and RMSE results for the synthetic scenario with the complex plume data and the 48-well random network are shown on Figure 6.4.



(a) RMSPE



(b) RMSE

Figure 6.4: RMSPE (a) and RMSE (b) as a function of eLPM sub-sample size compared to the total number of available observations in the synthetic scenario with 48 randomly arranged wells over the complex CoPC plume. The line represents the median value and the shaded area represents the 95% variability band.

The RMSPE results on Figure 6.4a show exponential decrease with increasing sub-sample size. These results are similar to the results in Section 5.3.4, where the temporal distribution of the samples followed the approach of repeated spatial samples with equal sample size (see Section 5.2.1). In both cases, the trend indicates that using a higher ratio of observations leads to over-sampling. In this case, the median RMSPE approaches its minimum at around a sample size of 50% of the available observations, while variability decreases towards 100%. Thus, using a sub-sample with at least $\sim$50% of the total number of observations leads to the same CoPC concentration surface estimates on average as using $\sim$100%. Figure 6.4b shows the RMSE results, which also exhibit a similar exponential decay. This indicates that the RMSE could be an adequate RMSPE proxy metric to use in real data applications. Although the RMSE values keep decreasing until all observations are used, the rate of decrease is substantially lower beyond the 50% point, going from an RMSE of $\sim$2.5 to $\sim$2.1 at 95%. On the other hand, the variability increases between 75% and 95%, which could indicate that the location of the unselected observations has a large impact on RMSEs, especially as fewer unselected observations remain. Given this trend in the RMSE curve, one could reasonably conclude that a lower sampling intensity could have been sufficient. All other tested synthetic scenarios produced similar RMSPE and RMSE results, adding further empirical evidence supporting the use of RMSE to evaluate

the sufficiency of historic sampling intensity.

Figure 6.5 shows the RMSE results for the case study groundwater contamination data set. The results of this case study also exhibit a similar, exponentially decreasing trend as the synthetic scenarios. Based on the RMSE curve, the models predicted similar concentration levels using only ∼80% of the observations and with substantially lower variability than using 95%. Overall, the case study data application substantiates the results obtained with the synthetic data sets.

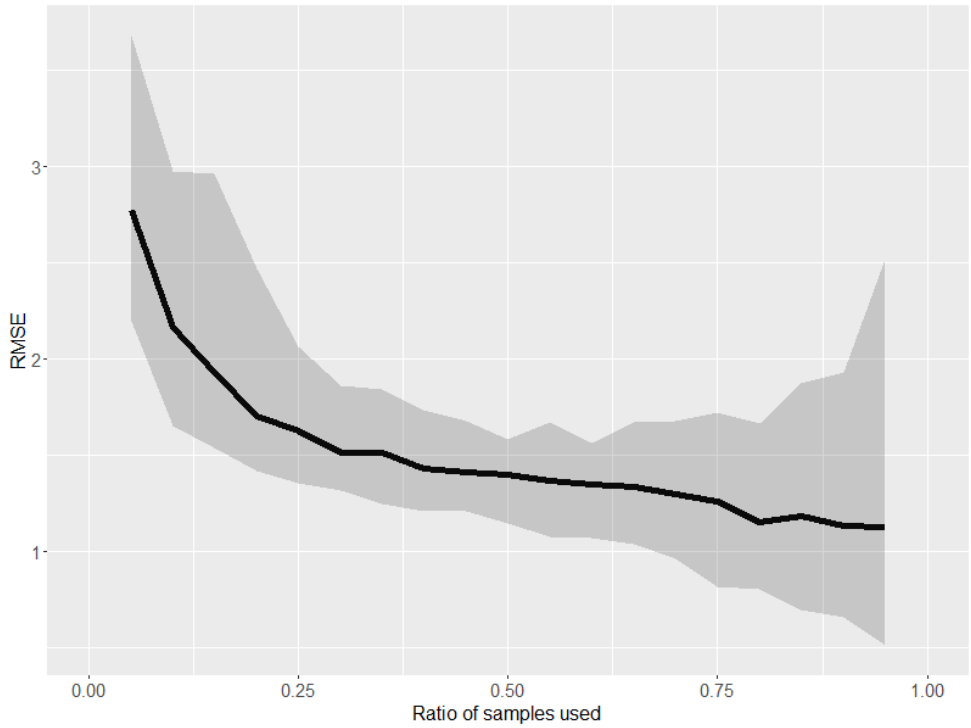


Figure 6.5: RMSE of unselected observations as a function of eLPM sub-sample size in the case study. The line represents the median RMSE value and the shaded area represents the 95% variability band.

To see if the exponentially decreasing pattern is not a general feature arising from the way RMSE is calculated but is actually due to the oversampling of the site, further tests were performed on modified versions of the case study data set. The complete set of 382 observations was reduced in size by randomly removing a varying number of observations to generate increasingly undersampled scenarios. The relationship between the amount of observations used and the corresponding RMSE were then analysed in these undersampled cases. The results showed that the exponentially decreasing trend persisted until the total number of observations was reduced to around 70, from the original 382. Below this, a more linear trend characterized the relationship. Figure 6.6 shows the results of this case. The linear decrease indicates that the relationship does not approach an asymptote over the given sub-sample size range. Thus, each additional observation added to the sub-sample contributes to the decrease in RMSE, substantiating that from

the perspective of the given modelling approach, the site is undersampled.

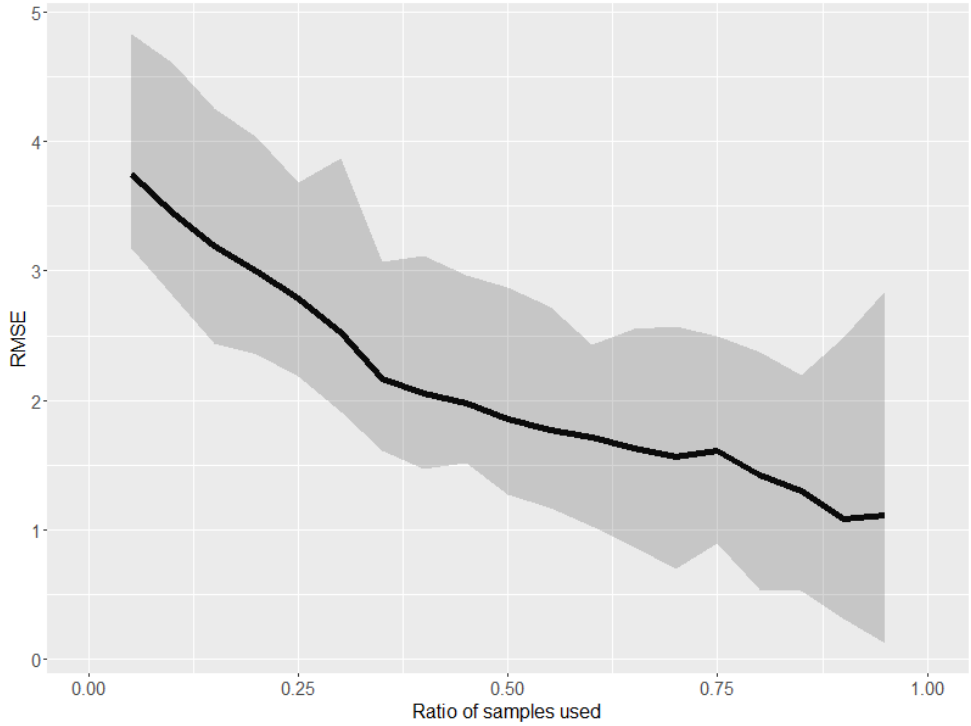


Figure 6.6: RMSE as a function of eLPM sub-sample size in the case study, with the total number of ethylbenze concentration observations reduced to 70 from 382. The line represents the median RMSE value and the shaded area represents the 95% variability band.

6.2 Sampling Frequency and Sample Size in Long-Term Designs

The previous chapter, Chapter 5, was focused on investigating how the total sample size affected groundwater contamination models that were based on observations from probability sampling approaches SRS, eLPM and pLPM. This section examines how the number of sampling events and sample size per sampling event affects these models. The goal was to identify any differences in predictive efficacy between using sampling designs with low and high sampling frequency, whilst assuming a constant total sample size. The outcome of the analysis would also indicate whether the P-splines modelling approach provides more precise estimates given spatially or temporally higher resolution data.

6.2.1 Background

The optimisation of logistics in long-term groundwater monitoring is an important aspect to consider. Logistics in this case refers to the collection of groundwater samples from the monitoring site, and their delivery to laboratories for subsequent analysis. Sampling campaigns require resources, transportation, personnel and they are associated with high costs and safety hazards. Therefore, investigating the impact of sampling frequency on the predictive efficacy of monitoring operations could be an invaluable insight for increasing sustainability. The total number of samples collected over a given period is a product of the number of sampling events that occurred within that period, and the sample size per event. In practice, reducing the number of sampling events and simultaneously increasing per-event sample size to maintain the total number of samples collected is commonly more economical than the other way around, due to the costs associated with transportation. It requires less logistical effort to collect more samples whilst already at a monitoring site, than to perform multiple round-trips for fewer samples each time. Hence, this chapter aims to explore the effects of varying the number of sampling events (sampling frequency) and per-event sample size in spatiotemporal sampling designs generated using the SRS, eLPM and pLPM approaches.

Assuming a fixed total sample size over the monitoring period, sampling with a higher frequency, i.e. having more sampling events, allows for the lowering of the number of samples collected per event. This consequently results in a sampling design with high temporal, but low spatial resolution at the sampling times. Conversely, increasing the number of samples collected per event means a possible reduction in the frequency of sampling events. The resulting sampling design will have high spatial resolution at the sampling times but lower temporal resolution overall. Therefore, the outcome of the analysis is highly dependent on whether the spatiotemporal P-splines models give more precise predictions using spatially or temporally higher resolution data.

In summary, the main questions this chapter attempts to answer are a) given a monitoring period with a constant total sample size, does the combination of the number of sampling events and samples per event make a difference when estimating CoPC concentrations from the samples, or only the total sample size matters and b) is there a difference in the outcome of question a) when using SRS, eLPM or pLPM to design the sampling schedule. The answer to question a) would also indicate whether the P-splines modelling approach provides more precise predictions given spatially or temporally higher resolution data.

6.2.2 Simulation Study

The results were obtained using the simulation study described in Section 5.2 of Chapter 5, where the synthetic contamination data of three hypothetical plumes were used. The flowchart of this simulation study can be seen on Figure 5.4. The hypothetical plume data sets were introduced in Chapter 1.4.2, and a visual representation can be seen on Figure 3.1. The synthetic observations were obtained by considering the data at sampling locations described by hypothetical monitoring networks. The monitoring networks differed in the spatial arrangement of the wells and the number of wells. As highlighted in Chapter 5.2.3, the randomly arranged network was the closest representation to real groundwater monitoring networks. Figure 5.5 shows the monitoring network arrangements. To allow for the evaluation of the trade-off, additional parameters had to be specified in the simulation study, which were the number of sampling events and the sample size per sampling event. In the original simulation setup, only the total sample size varied and all 10 of the sampling events were utilised, whereas in this case, the total sample size determined the potential values the number of sampling events and the sample size per event parameters could take. Table 6.1 shows the possible combinations of these parameter values at different total sample sizes. The total sample size is the product of the number of sampling events and the sample size per event. The maximum number of sampling events is 10, and the maximum sample size per event is determined by the number of available monitoring wells, i.e. the potential sampling locations.

The flowchart in Figure 6.7 shows the sequence of steps in this modified version of the simulation study. After choosing a combination of number of sampling events and sample size per event based on the total required sample size, the sampling design method (SRS, eLPM or pLPM) is applied to the synthetic groundwater monitoring network (see Sections 5.2.2 and 5.2.3) to draw the spatiotemporal samples. These samples are subsequently used to estimate the CoPC concentration surface via the P-splines models.

In the original simulation setup all 10 available sampling events throughout the monitoring period were utilised. In this case however, the sampling events had to be selected out of the 10 potential ones, based on the required number of events. Due to the evolution of the CoPC plume through time, the selection of the sampling event determines spatial structure of the plume at the time of sampling. The P-splines model would then interpolate the concentration levels between the selected events. A fixed scheme was created for the selection of events at different values of the number of sampling events parameter. This scheme is presented in Table 6.2. The scheme aims to achieve an even distribution of sampling events in time while also ensuring that the beginning and end of the monitoring period is sampled.

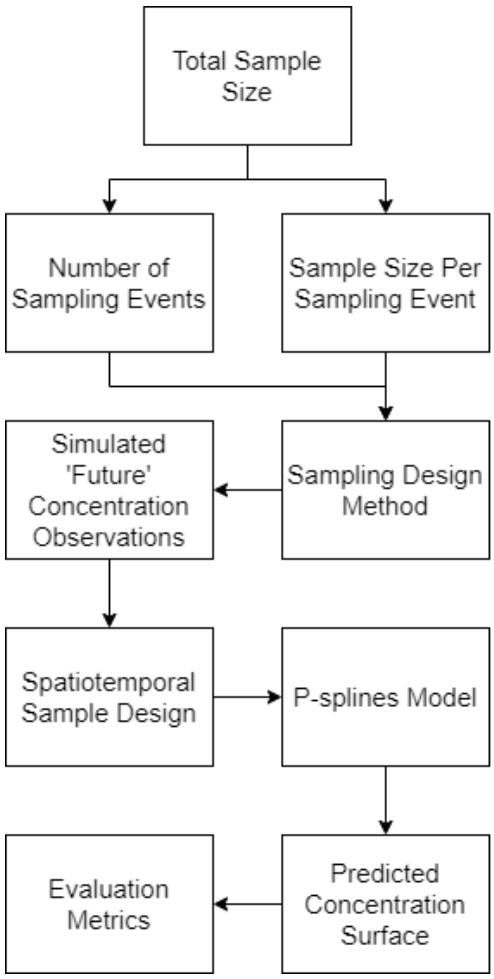


Figure 6.7: Flowchart showing the generation of sampling designs by varying the number of sampling events and the sample sizes per event in the simulation study.

Table 6.2: The distribution of utilised sampling events with 10 potential events in total throughout the monitoring period. The shaded cells represent the sampling events utilised at different levels of sampling frequency (number of sampling events).

	Distribution of Sampling Events									
ID of Sampling Event	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Number of Sampling Events Used										
10	■	■	■	■	■	■	■	■	■	■
8	■		■	■	■	■	■	■		■
6	■		■		■		■		■	■
5	■		■		■		■			■
4	■			■			■			■
3	■				■					■
2	■									■

Table 6.1: Combinations of sample sizes and number of sampling events resulting in certain total sample sizes in a monitoring network consisting of 48 wells throughout a period of 10 sampling events.

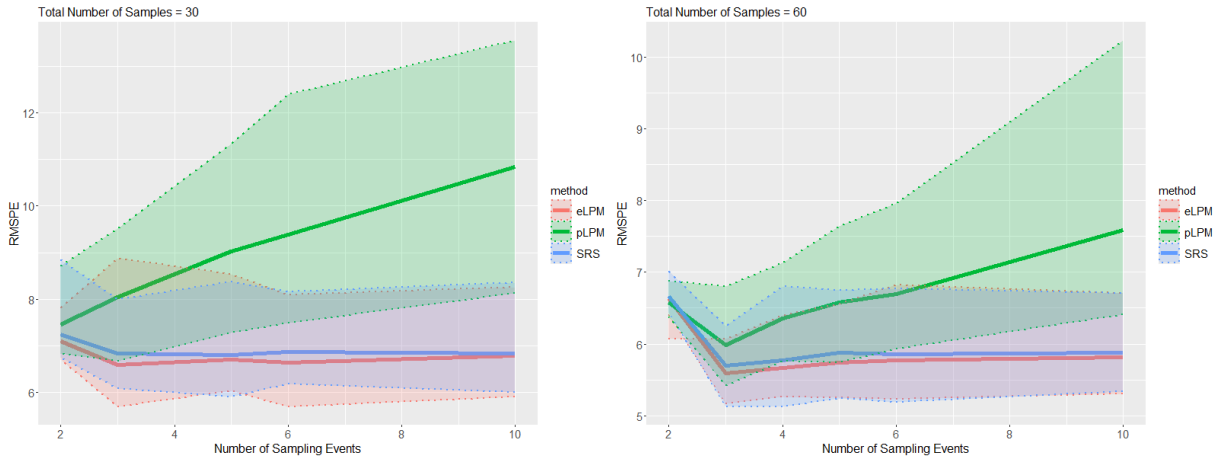
Total Sample Size	Number of Sampling Events	Sample Size Per Sampling Event
480	10	48
240	10	24
	8	30
	6	40
	5	48
120	10	12
	8	15
	6	20
	5	24
	4	30
	3	40
60	10	6
	6	10
	5	12
	4	15
	3	20
	2	30
30	10	3
	6	5
	5	6
	3	10
	2	15

The same evaluation metrics were used to analyse the results of the simulation study as in Section 5.2.4. These were the overall and the plume-specific RMSPE values and the plume mass estimates. The simulation was performed on the randomly arranged 48-well monitoring network scenario, with the complex CoPC plume data (see Figure 5.5b). The evaluation metrics were calculated over the convex hull of the monitoring network (see Figure 5.6).

6.2.3 Results

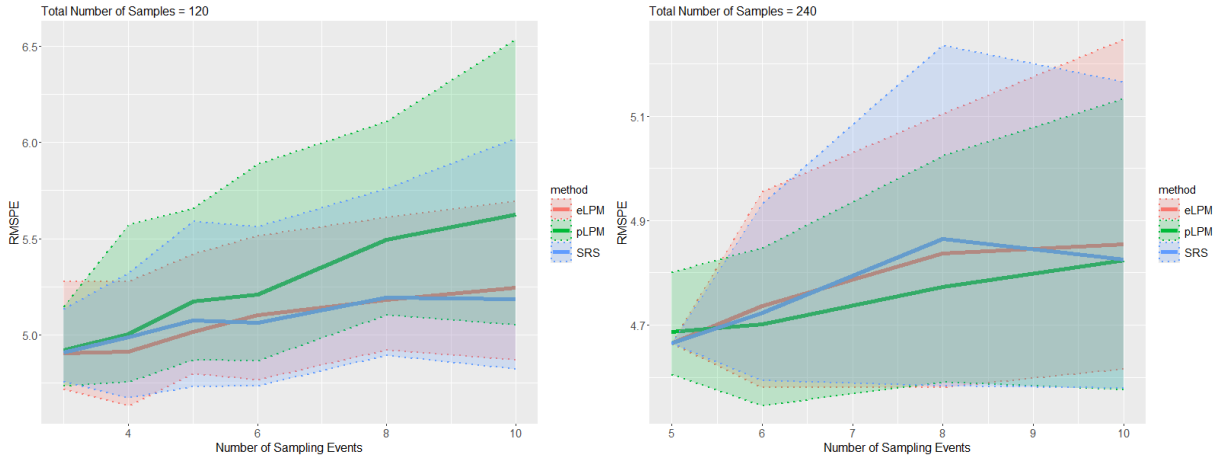
Figure 6.8 shows the RMSPE results. At total sample sizes of 30 (Figure 6.8a) and 60 (Figure 6.8b) both SRS and eLPM appear to be insensitive to sampling frequency. They result in models with similar RMSPEs, regardless of the temporal and spatial resolution of the data, with the exception of sampling frequency of 2. At this frequency, the two samples collected at the first and last sampling events (see Table 6.2), have a high spatial resolution, but the time between the two events appears to be too long for the model to interpolate precisely. If a third sampling event is introduced between these two (3 sampling events), the RMSPE decreases sharply, indicating a

more adequate model. The pLPM-based designs on the other hand show a well defined, increasing trend towards higher sampling frequencies. This is indication that the pLPM designs benefit from drawing higher spatial resolution samples. Because of the prioritisation of wells close to the plume, a pLPM sample with low spatial resolution might overestimate CoPC concentrations, hence the higher RMSPE results. The increasing trend for pLPM designs is still present at total sample sizes of 120 (Figure 6.8c) and 240 (Figure 6.8d), albeit with a smaller difference in RMSPE results. This indicates that higher spatial and lower temporal resolution is still beneficial even with substantially larger total sample sizes. In contrast to the lower sample sizes, SRS and eLPM exhibit the same trend as pLPM here. At a sample size of 240, pLPM-based samples provide lower median RMSPE and variation.



(a) Total Sample Size = 30

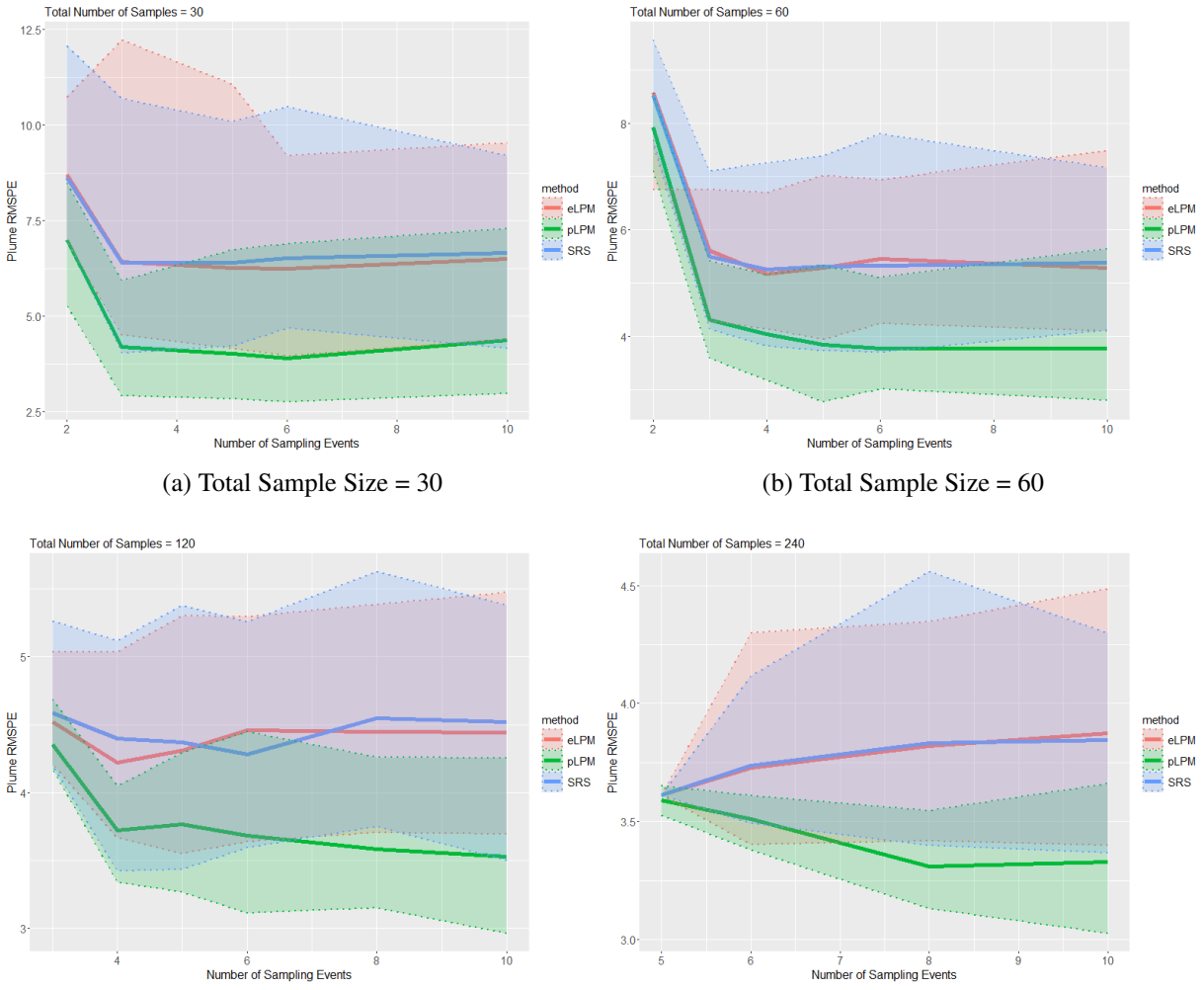
(b) Total Sample Size = 60



(c) Total Sample Size = 120

(d) Total Sample Size = 240

Figure 6.8: RMSPE of sampling designs via SRS, eLPM and pLPM as a function of the number of sampling events within the monitoring period (starting at the lowest possible number needed to maintain sample size), with constant total sample size. The median line is shown with 95% variability bands (shaded area). The simulation was performed on the randomly arranged 48-well monitoring network, with the complex CoPC plume.



(a) Total Sample Size = 30

(b) Total Sample Size = 60

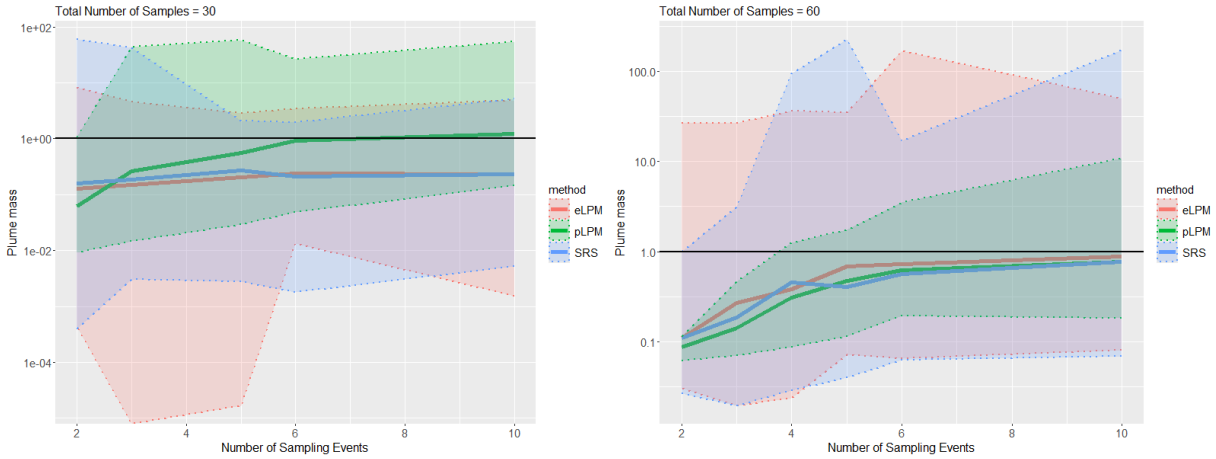
(c) Total Sample Size = 120

(d) Total Sample Size = 240

Figure 6.9: Plume-based RMSPE of sampling designs via SRS, eLPM and pLPM as a function of the number of sampling events within the monitoring period (starting at the lowest possible number needed to maintain sample size), with constant total sample size. The median line is shown with 95% variability bands (shaded area).

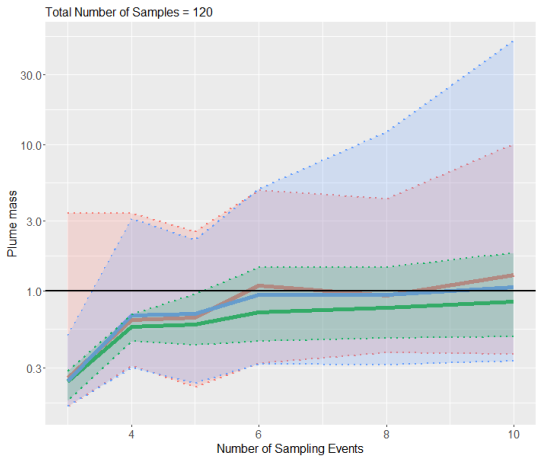
Figure 6.9 shows the RMSPE results isolated for the concentration estimates within the plume. With the exception of the lowest frequency, SRS and eLPM again exhibit invariance to the number of sampling events. At a total sample size of 240, they do show a slightly increasing tendency towards higher sampling frequencies, indicating a preference for higher spatial resolution as opposed to temporal resolution. However, the gain in RMSPE by lowering the number of sampling events is not substantial. The pLPM-based models also exhibit invariance at lower total sample sizes (Figures 6.9a and 6.9b), except at the lowest sampling frequency (2 sampling events). At higher total sample sizes however (Figures 5.8c and 6.9d), they do exhibit a decrease in RMSPE with an increasing number of sampling events. This indicates that for achieving a more precise model of CoPC concentration estimates within the plume, higher temporal resolution data can be beneficial. This effect only seems to occur when per-event sample sizes reach a sufficient

level to provide adequate spatial resolution even at higher sampling frequencies. This is why the trend appears stronger at total sample sizes 120 and 240. The RMSPE and plume-RMSPE results indicate the presence of a trade-off when using the pLPM sampling design. To achieve a more balanced estimate of CoPC concentration levels over the convex hull, higher spatial resolution seems to be more beneficial, while to achieve a more precise estimate within the plume, higher temporal resolution produces more promising results.

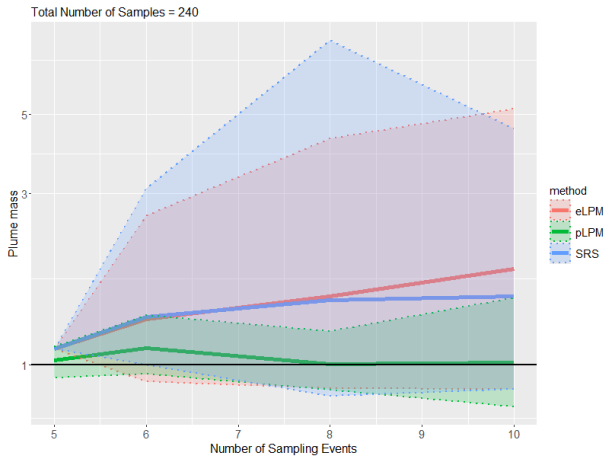


(a) Total Sample Size = 30

(b) Total Sample Size = 60



(c) Total Sample Size = 120



(d) Total Sample Size = 240

Figure 6.10: Plume mass estimated by sampling designs via SRS, eLPM and pLPM over the actual plume mass, as a function of the number of sampling events within the monitoring period (starting at the lowest possible number needed to maintain sample size), with constant total sample size. The y-axis is shown in logarithmic scale. The black line at 1 represents estimated plume mass equal to actual plume mass. The median line is shown with 95% variability bands (shaded area).

Figure 6.10 shows the plume mass estimation results. The results show a general trend that higher sampling frequency leads to more precise estimation of the plume mass, regardless of which sampling design algorithm is applied. However, the pLPM design provides the lowest

variation and the most precise estimates on average. A low number of sampling events generally leads to the underestimation of the plume. At a total sample size of 240 (Figure 6.10d), increasing the number of sampling events leads to an increasing overestimation of the plume mass by SRS and eLPM designs, whereas pLPM designs appear unaffected. The trade-off of pLPM designs extends to estimating the plume mass as well. Collecting data with higher temporal resolution could help produce more precise estimates of the plume, but at the same time, could lead to reducing precision in characterising the concentration levels at the boundaries of the monitoring site.

6.3 Conclusions and Future Work

In conclusion, in the first study, empirical evidence was collected to support the application of the presented approach for analysing the sufficiency of historic sampling designs compared to less intense, spatially balanced sampling designs. The approach can be used to indicate if collecting fewer, spatially balanced samples could have resulted in sufficiently similar predictions. This information could be used to adjust the sampling design and intensity on the monitored site going forward.

The evaluation method works by drawing increasingly smaller samples from the available historic observations using the eLPM algorithm. A spatiotemporal statistical model is then fitted to the samples and the predicted concentration values of the unselected observations are compared to the actual, observed values via RMSE. The relationship between RMSEs and the number of observations considered in the model can then be analysed. If the function of this relationship approaches an asymptote before all available historic observations are considered in the model, then that indicates that fewer observations would have been sufficient to obtain similar predictions. On the other hand, if the relationship exhibits a more linear trend, that could indicate that the historic sampling intensity has been insufficient, since each additional observation contributes to reducing the RMSE.

The results of this study also substantiated that RMSE could be used as the evaluation metric for the proposed method, since it exhibited a similar, exponentially decreasing behaviour as the RMSPE calculated using the real and predicted concentration surfaces of the convex hull in synthetic groundwater contamination data.

The results of the second study indicated that the predictive potential of spatiotemporal groundwater sampling designs does not only depend on the total number of samples collected, but also the frequency of sampling events and the number of samples collected per sampling event. There also appears to be a difference in the effects of sampling frequency and sample size when

it comes to different sampling design approaches. SRS and eLPM in general provide improved CoPC concentration estimates when the sample size per sampling event is maximised, as opposed to the frequency of sampling events. This is also the case for pLPM designs when looking at the overall concentration estimates in the convex hull. However, when estimating concentrations within the plume or the plume mass, pLPM provides more informative observations when a higher sampling frequency is employed. Due to this apparent trade-off, it is important to consider the main objectives of the monitoring operation and select an appropriate sampling frequency and sample size per sampling event when using the pLPM design approach. With adequate planning, the pLPM approach could however provide benefits in estimating CoPC concentrations and the characteristics of the plume. Regarding the P-splines models, the RMSPE results indicate that in general they provide better predictions overall for observations with higher spatial resolution. However, there is some indication that if the objective is the estimation of the plume mass, higher temporal resolution can improve the outcome.

Future work should further substantiate the conclusions drawn for the sampling intensity evaluation approach by testing the method in various synthetic and case study data scenarios, particularly focusing on reinforcing that RMSE is an adequate evaluation metric. Another area of future development should be on using the information obtained via the presented method, to adjust future sampling designs. A reasonable suggestion would be to determine the sample size and design at which the asymptote or the RMSE minimum is approached, and repeat that same sampling pattern in the future. Other options should also be considered, and a systematic approach should be devised and tested.

Regarding the second study on the trade-off between sampling frequency and sample size, more simulation studies with various contamination and monitoring network scenarios could substantiate the presented results. Additionally, testing a wider range of sampling frequencies and sample sizes could provide further insight. In terms of implementation, eventually the aim should be developing an optimisation system for suggesting sampling frequencies and sample sizes in long-term groundwater monitoring designs.

Chapter 7

Discussion & Future Work

The primary aim of this thesis was to develop novel statistical approaches for optimising the sampling strategies of existing long-term groundwater quality monitoring networks, and evaluate them via subsequent spatiotemporal modelling. First, an original comparison of different approaches to spatiotemporal modelling was carried out using the GAM framework. Then, a novel application of influential observation statistics was proposed for ranking groundwater sampling wells in monitoring networks with respect to their contribution to estimating CoPC concentrations over time and space. Well influence analysis aimed to provide a computationally efficient alternative to a more demanding CV-based approach for supporting decision makers in identifying wells that provide redundant information. These wells could potentially be omitted from the sampling design, thus reducing sampling costs whilst maintaining estimation quality. The proposed ranking approach has been successfully implemented in the groundwater quality modelling software GWSDAT (Jones et al. [2014]). Furthermore, an innovative, data-driven approach was developed and evaluated for tuning inclusion weights in spatiotemporally balanced probability sampling designs, with the aim of better capturing the spatial trends arising through the evolution of CoPC plumes in groundwater over time. Finally, an original application of spatiotemporally balanced probability sampling designs was proposed for determining an adequate sampling intensity for future monitoring periods based on available historic data. Additionally, an investigation was carried out on balancing the trade-off between increasing spatial or temporal sampling resolution, given a fixed sample size, with the aim of finding optimal spatiotemporal sample distributions for supporting the estimation of the CoPC concentration surface. The following sections will provide a summary of the work undertaken in this thesis, highlight the obtained results and their significance and discuss limitations and potential future developments. The limitations and recommendations for future research are discussed for each chapter separately at the ends of the corresponding summaries. Additionally, an overview discussion of these topics is presented at the end of this chapter.

7.1 Spatiotemporal Modelling Approaches

As discussed in Chapter 2, McLean et al. [2019] showed that a spatiotemporal modelling framework is more effective for groundwater quality monitoring data than repeated spatial modelling through a comparison using Kriging and P-splines. The P-splines modelling framework used in the study and throughout this thesis was introduced in Chapter 1.3.1, and is also implemented in GWSDAT (Jones et al. [2014]) to estimate the CoPC surface. As highlighted in Chapter 2, this model is a particular case of the broader GAM framework, which provides different ways of approaching spatiotemporal modelling. In long-term groundwater CoPC concentration data, the explanatory variables consist of the coordinates (easting and northing) of the monitoring wells and sampling times (see Chapter 1.4). The GAM framework allows for modelling the spatial and temporal variables as separate smooth terms, with a bivariate term representing easting and northing and a univariate term representing time, or as a single trivariate smooth term representing both space and time. The latter approach is used in the GWSDAT P-splines models (see Chapter 1.3.1). Additionally, the smooth terms can be constructed using different functions, including thin plate regression splines, tensor product splines or P-splines. The aim of Chapter 2 was to conduct a comparative analysis of these different GAM approaches and qualitatively assess the estimated concentration surfaces compared to the P-splines approach from GWSDAT. These investigations also served as a way to explore the common characteristics of groundwater quality monitoring data and highlight challenges in terms of designing sampling strategies.

Firstly, six different approaches to constructing groundwater quality GAMs were explored (see Chapter 2.2). The approaches were based on two factors, the number of smooth terms, and the type of splines used to construct these terms. Three approaches were investigated for representing the covariates in the model. These were separate smooth terms for space and time, a joint spatiotemporal smooth term and a univariate smooth term for time repeated at each well individually. For the type of smoothing spline used, there were two choices. These were penalized thin plate regression splines and penalized tensor product splines (see Chapter 2.1). The resulting six GAM approaches were used to model two case study data sets introduced in Chapter 1 Sections 1.4.3 and 1.4.4 (see also Section 2.2 in Chapter 2). The models were evaluated using three different information criteria: AIC, BIC and R^2 (see Section 2.3).

Overall, the trivariate, spatiotemporal smooth term models constructed using tensor product splines appeared to represent the groundwater quality data better. The results also indicated that for separate spatial and temporal smooth term GAMs, the choice of smoothing spline function was not critical. Both thin plate regression and tensor product splines resulted in similar models according to the evaluation statistics. The typically higher computational cost of thin plate regression splines suggested that in these cases, using tensor product splines could be a more economical approach (S. Wood [2017]). The separate smooth terms GAMs also produced simi-

lar estimates of the concentration surface as GWSDAT. On the other hand, for trivariate smooth term GAMs, using tensor product splines produces better model fits. The three model evaluation criteria indicated that a spatiotemporal models worked better on the case study 1 data set (introduced in Chapter 1.4.3). However, the BIC indicated that the separate smooth term model produced a better fit for case study data set 2 (introduced in Chapter 1.4.4). Using the approach that constructed separate temporal smooths for each individual monitoring well produced adequate models when looking at the time-series estimates. However, this approach could not be used to produce spatial estimates of the CoPC concentration surface, which is often the main goal of groundwater quality monitoring.

Secondly, the two best fitting approaches (separate thin plate spline and trivariate tensor product spline models) were used to model simulated groundwater contamination data (introduced in Chapter 1.4.1), in order to allow for the evaluation of their prediction accuracy using RMSPE (see Chapter 2.5.6). Two distinct hypothetical sampling scenarios were constructed from the simulated CoPC concentration data (see Chapter 2.5.1). The first one represented a realistic scenario where the sampling of the monitoring wells took place sporadically. The second one represented an idealistic scenario where each of the monitoring wells were sampled continuously with high frequency.

The results showed a discrepancy between relying on criteria such as AIC and BIC versus RMSPE to identify the most adequate model. AIC and BIC indicated a better fit for the trivariate, spatiotemporal model in both scenarios, but looking at prediction errors for the concentration surface via RMSPEs suggested using the separate smooth terms model. The investigation of the spatial concentration estimates showed that the trivariate tensor product spline model produced a better fit for the observations, but suffered from ballooning (see Chapter 2.5.4) issues when interpolating the CoPC concentration surfaces over the entire site. The separate smooth term approach in contrast appears to provide more accurate estimates of CoPC concentration surfaces overall, but this is likely due to the model resulting in a smoother fit. Since the data is mostly dominated by low concentrations, the smoother fit results in a lower RMSPE value. However, the residual autocorrelation plots highlighted issues with model specification for this approach, which should be addressed when using this model for inference. The ballooning of the trivariate smooth model could be mitigated by adjusting the number of basis functions used in the three dimensions or using different smoothing parameter estimation methods.

In summary, the results obtained in Chapter 2 support the use of spatiotemporal modelling frameworks in groundwater quality monitoring, but highlight the need for adequate model specification and adjustment of model smoothness to obtain reliable spatiotemporal CoPC concentration estimates. The investigations also suggested that it is important to reduce the presence of information sparse regions and periods in the data to mitigate the ballooning of concentration

estimates. One way to achieve this within the constraints of pre-existing monitoring wells is spreading the samples as evenly as possible in space and time. This concept was investigated in Chapters 4, 5 and 6 through the use of spatiotemporally balanced probability sampling designs.

An additional GAM structure could also be explored in which a spatial smooth term is constructed separately for each sampling time. This produces an approach similar to repeated spatial models. While this approach produces a model that incorporates time and space simultaneously, it does not consider time as a continuous variable, and thus does not borrow strength from its correlation to changes in CoPC concentrations (McLean et al. [2019]). Therefore, this approach could be less useful in estimating spatiotemporal CoPC concentration surfaces. However, it could still provide benefits in certain circumstances. For example, such repeated spatial models could be useful for data with low sampling frequency as the lack of temporal resolution could result in the ballooning of a trivariate smooth term model. Evidence for this was found in Chapter 6, Figure 6.9, where prediction errors tended to increase with a decreasing number of sampling events, given a fixed sample size.

7.2 Well Influence Analysis

As highlighted throughout this thesis (see Chapter 1.2.2), the sampling of wells in existing groundwater monitoring networks can be costly, and can also contribute to detrimental effects in the surrounding environment (Meray et al. [2022]). Therefore, the selection of an optimal sampling intensity that minimises detrimental effects whilst ensures the quality of statistical estimation is crucial. The analysis of historic observations, if available, provides several opportunities for adjusting future sampling strategies. One way to reduce intensity is to reduce the size of the monitoring network by omitting wells that provide redundant information. The approach that was investigated in Chapter 3, is the ranking of monitoring wells based on how much useful information they provide. Since the primary goal of groundwater quality data analysis is commonly the estimation of the concentration surface of solutes, the redundancy or influence of monitoring wells can be assessed on how much they contribute to the accuracy of this estimation. As discussed in Chapter 3, through feedback from users of GWSDAT (Jones et al. [2014]), it was identified that a feature that facilitates such an analysis was highly sought after. In GWSDAT version 3.1, released in 2022 (Jones et al. [2022]), a new functionality called well redundancy analysis was added, which allowed users to manually remove one or multiple wells from the data set and re-compute the concentration surface to provide supporting evidence that the conclusions of the analysis would not be substantially different with the omission of certain wells. However, this feature relied on a manual trial and error approach, which is not feasible for monitoring networks with a large number of wells. Since the omission of a well, affects

the importance of neighbouring wells too, if the goal is to omit multiple wells, the number of combinations of candidate wells can become prohibitively high to exhaustively investigate this way. Therefore, the need to compliment the well redundancy analysis feature with functionality for automatically ranking wells was identified.

The manual trial and error approach could be substituted by an automatic one that iterates through the list of wells, leaving one out in each iteration, and estimates the prediction error via RMSE. This approach was referred to as well-based cross-validation (WBCV) after Evers et al. [2015], and was introduced in Chapter 3.4. The drawback of this approach was that the P-splines modelling framework had to be applied to the training data in each iteration, making it computationally taxing. The primary aim of the work presented in Chapter 3 was to develop a computationally efficient novel approach as an alternative for ranking monitoring wells by influence. Since the P-splines framework is analogous to regression analysis (Eilers and Marx [1996]), the use of influential observation identifying metrics was proposed. These metrics are commonly used to identify outliers that have a disproportional influence on the estimates in regression models (Belsley et al. [2005]). The innovative application of these metrics in this thesis was ranking the monitoring wells based on the average influence of their observations. Wells with the least average influence are the most redundant for estimating the concentration surface, therefore they would be suggested first for omission from the sampling design. The advantage of this approach compared to WBCV is that the initial P-splines model fit is enough to produce a ranking, since the influence metrics are calculated using residuals and leverages (see Chapter 3.5).

The simulation study carried out in Chapter 3 aimed to identify the influence metric that produced the most similar rankings to WBCV (see Chapter 3.8). Several different hypothetical scenarios were constructed for the simulation study using different CoPC plumes (introduced in Chapter 1.4.2) and monitoring networks (see Section 3.1.1). Assessing the similarity between the two ranking approaches in these various scenarios also provided information on the effects of the different features that make up these scenarios. These features were the plume's geometric complexity and the number and spatial arrangement of the monitoring wells. There were also two additional lines of inquiry. The first was related to how the type of measurement error associated with the observations affected the comparison (see Chapter 3.1.2). The laboratory analyses performed on groundwater samples to determine the concentrations of solutes produce results within certain error margins. It is commonly assumed that these measurement errors are multiplicative (McLean [2018]), meaning that smaller concentrations are associated with smaller errors and larger concentrations with larger errors. With additive errors on the other hand, the magnitude of the noise is the same, regardless of the concentration. Both of these measurement error types were tested to assess whether the type of noise present in the data affects the performance of the influence analysis metrics. The second additional inquiry investigated

the effects of increasing the number of basis functions used in the P-splines models (see Chapter 1.3.1). To allow the comparison of the rankings, an evaluation metric was developed to measure the absolute difference between WBCV and influence metric based rankings, by aggregating the differences in the rank positions of individual wells (see Chapter 3.7). The comparison was also performed on case study 2 (introduced in Chapter 1.4.4), to support the the results of the simulation study.

Ultimately, the closest and most consistent approximation of the WBCV rankings was achieved using Cook's distance as the influence metric for influence analysis. The rankings were adequately close as highlighted by a comparison of ranked lists achieved in the case study (see Chapter 3.10). This was the case across all scenarios in the simulation study as well. The computation time for producing a ranking using the Cook's distance based approach was also substantially lower than using WBCV (see Chapter 3.10). Since the influence ranking is intended to provide a starting point for more exhaustive, manual investigation of the impacts of removing a well, the obtained results were adequate to support the application of the proposed approach.

One of the indications of the results was that the performance of the influence metric based approach in estimating WBCV rankings was substantially worse when the data had additive measurement errors. It was identified that large parts of the simulated data set as well as the case study data set had low solute concentrations, thus applying the additive error structure resulted in much higher overall noise levels in the data. This negatively affected the calculation of influence metrics, resulting in larger deviations from WBCV results. However, as mentioned before, it is commonly assumed that groundwater quality monitoring data has a multiplicative error structure (McLean [2018]). It could still be beneficial to assess the error structure of the data when applying the influence metric based well ranking approach.

Another observation from the simulation study was that varying the number of basis functions in the P-splines model had a limited effect on the comparison of rankings, although only two different parameter values were investigated. The number of basis functions changes the flexibility of the model and thus has an impact on the residuals, which could have a substantial impact on the rankings. This is because, as discussed in Chapter 3.5, the influence statistics are commonly based on the values of residuals and leverages.

The simulation study also suggested that the monitoring network design and the CoPC plume's geometric complexity also have an effect on the similarity of the rankings. One of the more identifiable trends was that on average, more complex plumes result in the influence metric based approach being less accurate at estimating the WBCV rankings. This effect could also be attributed to the higher residuals produced by the model when estimating the concentration surface of the more complex plume. It was also identified that monitoring well arrangements

that follow the spatial structure of the CoPC plume lead to smaller differences between the two ranking approaches, possibly due to similar reasons. In terms of the number of monitoring wells, there was some indication that fewer wells result in better approximations by the influence metric based approach. However, this effect is likely the result of the fewer possible permutations in the rankings.

It should be noted that the ranking produced by the proposed influence metric based approach is an approximation to WBCV and it aims to provide a starting point for omitting wells from the sampling design. In this proposed implementation, the main design objective was computational and time efficiency. Hence, the computed ranking alone is not sufficient justification for omitting wells from the sampling design and other factors such as hydrogeological features, well positioning and accessibility should also be considered, along with an exhaustive investigation of the changes produced in the model predictions. Another point to consider is that influence diagnostics are normally used to identify influential observations such as outliers by checking whether they cross a certain threshold value. Data points that cross the threshold warrant a second look to determine whether they should be removed from the model to improve the fit. In the proposed implementation, it was assumed that the influence of an observation is proportional to its influence metric value, and thus the well-wise averages of these metrics could be used to order the monitoring wells. The impact of using different averaging functions or potentially considering the sum of influences instead, could be investigated as well in the future. In summary, the simulation and case studies in Chapter 3 provide empirical evidence supporting the use of influence metrics to assess the redundancy of wells in groundwater monitoring networks.

A potential area of improvement in well redundancy ranking is the implementation of computationally more efficient cross validation approaches in WBCV. Reducing high computation times would allow the integration of WBCV in groundwater quality modelling software such as GWSDAT. In the R package *mgcv* (S. Wood [2021]) for example, neighbourhood cross validation (NCV) can be used to optimize smoothing parameters in GAMs. In order to reduce the computational cost of this optimisation procedure, S. Wood [2024] developed a quadratic approximation to the NCV criterion (QNCV). This approach makes the computational cost of NCV comparable to a single model fit, which could represent an $O(n)$ saving when modelling n data. If this approach could be generalised and implemented in the context of well redundancy ranking, it could provide a computationally comparable alternative to influence metrics. The efficiency, accuracy and explainability of such an approach would have to be investigated.

The work presented in Chapter 3 of this thesis has been peer reviewed and published in the proceedings of the 37th International Workshop on Statistical Modelling (Radvanyi et al. [2023]), and has been submitted to the journal of Environmental and Ecological Statistics [1] for review.

[1] https://link.springer.com/journal/10651

Additionally, the influence metric based well influence analysis is currently implemented in GWSDAT to compliment the well redundancy analysis feature by providing a ranked list of the monitoring wells (Jones et al. [2023]). The code used to perform the simulation study and the case studies was written in the open-source statistical programming language R (R Core Team [2020]), and is available on GitHub [2]. A shiny (Chang et al. [2021]) web application has also been developed to allow for the reproduction of the simulation study results [3].

In summary, the well influence analysis approach proposed in this thesis was shown via empirical testing to be able to approximate the importance of individual wells within a monitoring network to an adequate degree and computationally very efficiently. The achieved saving in time consumption allowed the approach to be implemented in a widely used groundwater quality data analysis software, GWSDAT, where it can aid groundwater quality monitoring practitioners in optimising sampling designs and thus contribute to increasing the sustainability of long-term sampling practices.

7.3 Spatiotemporal Groundwater Sampling Designs

Throughout this thesis, the importance of an optimal sampling strategy for long-term groundwater quality monitoring has been highlighted. It can reduce the costs, risks and environmental impacts associated with transportation, equipment, analyses and personnel, whilst ensuring that high quality data are collected. However, most of the literature on this topic is focused on identifying optimal locations for new monitoring wells, rather than selecting the most optimal wells to sample in an existing network (Farlin et al. [2019]). In addition, most of the methods proposed to address this gap commonly focus only on spatial optimality and provide no suggestions on sampling frequency, despite groundwater contamination being a spatiotemporally evolving system. As mentioned in Chapter 1.2.3, McLean [2018] proposed two objective functions that could be used to optimise sample selection in a spatiotemporal framework, but they result in non-probability designs and have not been evaluated in terms of how precisely they can estimate CoPC concentration surfaces. Chapter 1.2.4 also discussed that current groundwater contamination monitoring practices are often inadequate for statistical analysis, and are only carried out to comply with regulations (Meray et al. [2022]). The sampling frequency is often prescribed by these regulations, while sampling locations are commonly selected by experts or by a systematic approach, such as alternating between subregions within the monitored site or individual wells. The data collected using these practices can be difficult to analyse statistically. Hence, this thesis argued that the long-term monitoring of existing well networks can benefit from the application

[2] https://github.com/peterradv/Well-Influence-Analysis
[3] https://peterradv.shinyapps.io/well-influence-analysis/

of probability sampling designs, specifically spatiotemporally balanced sampling approaches. Chapter 5 provided an overview of the concepts behind spatially balanced sampling and compared different designs that have been proposed in the literature. In summary, without explicit knowledge of underlying spatial patterns in the CoPC concentration, more accurate estimates can be achieved, if the collected samples are spread evenly over space and time. Therefore, spatially balanced designs aim to avoid sampling neighbouring locations that are close to each other, whilst utilising probabilistic processes for sample selection. Through the presented comparisons, the chapter explained why the Local Pivotal Methods (LPM) proposed by Grafström et al. [2012] were the most suitable choice for application and further development in long-term groundwater monitoring. At the time of writing this thesis, there was no instance found of the LPM design being applied to groundwater quality monitoring in the literature. Therefore, the comparison in Chapter 4 is an original contribution that uses the spatiotemporal version of the LPM design in a groundwater quality monitoring problem.

An attractive property of probability sampling approaches is that each potential sampling location is assigned a certain probability of being included in the sampling design, and these inclusion probabilities can be adjusted by an inclusion density function. This function can be based on any relevant, supplementary information about the population itself or the potential sampling locations. In the context of groundwater quality monitoring, the aim is usually to track the spatial evolution of the CoPC plume over time, in order to assess its stability and identify anomalies. Given that historic observations are available for a groundwater monitoring system, this data could be used to create an inclusion density function and subsequently generate a spatiotemporal sampling design that is optimised for tracking projected changes in the plume. Chapter 5 addressed this by developing and evaluation an innovative method for tuning sample inclusion probabilities based on predicted distances between monitoring wells and the CoPC plume. In this thesis, probability sampling approaches were also proposed for analysing historic monitoring data to determine whether past sampling practices led to under- or oversampling compared to balanced designs. In Chapter 6, this approach was evaluated by modelling increasingly smaller subsets of the historic data selected via spatiotemporally balanced sampling and comparing the resulting concentration surface prediction errors. Another aspect related to the implementation of probabilistic sampling designs for groundwater monitoring that was addressed in Chapter 6 was the trade-off between spatial and temporal resolution. Given a total sample size and a monitoring period of a certain length, sampling less frequently but visiting more wells each time produces a spatially higher, but temporally lower resolution data set. This approach is often preferred by experts, due to the cost and inconvenience of frequent transportation. This thesis investigated the impact of varying the balance of this trade-off on estimating the CoPC concentration surface. The following sections will summarise the results of the above described three lines of research, and discuss their implications, limitations and potential future developments.

7.3.1 Proportional Weight LPM Sampling Design

The aim of Chapter 5 was to develop an innovative technique for tuning sample inclusion probabilities in LPM sampling designs based on historic groundwater monitoring data. The resulting sampling designs carry over the benefits of spatially balanced designs whilst concentrating sampling effort on monitoring wells most affected by the CoPC plume. This allows the sampling design to capture the spatiotemporal trends of the CoPC concentrations more effectively.

The proposed approach works by modelling the CoPC concentration surface from historic monitoring data, delineating the plume based on a concentration threshold value, which is commonly determined by regulations, and predicting distances between its boundaries and the monitoring wells at future sampling times. The sample inclusion probabilities are then tuned so that they are proportional to the predicted distances, with smaller distances corresponding to higher inclusion probabilities and vice versa. After testing, it was determined that the relationship between predicted distances and inclusion weights should be non-linear to ensure an adequate level of strictness in well prioritisation. Ultimately, the best results were achieved under a relationship characterised by a half-normal kernel function (see Chapter 5.1.6), that decreased with increasing distance from the well (see Chapter 5.3.1). The result is that the LPM design at any given sampling time will be more likely to select sampling locations that are closer to the plume, thus tracking its evolution over time. The kernel function can be adjusted to be less or more strict at prioritising wells.

The proposed proportional weight LPM sampling design (pLPM) was evaluated in a simulation study that compared its effectiveness at supporting the prediction of the CoPC concentration surface to simple random sampling (SRS) and LPM with equal inclusion probabilities (eLPM). The simulation study used the simulated CoPC plume scenarios introduced in Chapter 1.4.2 and also used for the well influence analysis in Chapter 3. The three hypothetical cases represented plumes with geometries of increasing complexity. As also described in Chapter 5, the data sets contained concentration values for twenty time slices. The data were split, with the first ten slices representing the historic states of the system and the last ten representing future states. This allowed the use of the historic states to train generalised linear models (GLMs) to predict distances between the plume and the monitoring wells during future system states (see Chapter 5.1.5). Figure 5.4 in Chapter 5 shows a flowchart summary of the proposed approach for tuning the inclusion weights. In short, observations obtained from the historic part of the data were used to estimate the spatiotemporal concentration surface. This enabled the delineation of the CoPC plume and consequently the recording of the time-series of estimated distances between the plume boundaries and each individual well. The time-series data was used to predict future distances using GLMs. Finally, the half-normal kernel function was used to determine inclusion weights based on the predicted distances. The plume-well distance predictions could not be

obtained directly from the P-splines models (as described in Chapter 1.3.1) due to their limitations in predicting values outside of the data domain. The proposed approach circumvents these limitations by modelling the time-series of distances directly in individual wells, using a model framework capable of future extrapolation. The resulting spatiotemporal sampling designs were then applied to the future part of the data to obtain an optimised set of random samples, which were subsequently used to estimate the future concentration surface via the P-splines modelling approach. The estimated surface could then be compared to the true simulated surface and thus the prediction errors achieved by pLPM, eLPM and SRS could be evaluated. Additional design parameters in the simulation study included the size and arrangement of monitoring well networks (see Figure 5.5). It should be noted that the spatial arrangement of monitoring wells in the network is always the limiting factor in any groundwater quality sampling design. If the wells are located in non-optimal locations with respect to the CoPC plume, collecting more samples will not improve the model predictions. In the simulation study, using different plume scenarios and monitoring networks ensured that the results of the study are generalisable and are not only valid for specific cases. Initially, the CoPC concentration surfaces were estimated over the whole spatial domain contained in the simulated data sets (see Figure 3.1). However, to better align the results with real-world practices, additional simulations were also carried out where the concentration surfaces were only estimated over the convex hull, which is the space enclosed by the monitoring wells (see Figure 5.6).

The results indicated that the pLPM designs resulted in a more precise estimation of CoPC concentrations within the area affected by the plume. Given a certain sample size, the pLPM designs selected a higher percentage of their samples from within the boundaries of the plume, which evidently lead to better estimates. However, since less information was collected from outside the plume, the pLPM samples tended to result in higher prediction errors when considering the entire spatial data domain. This was a result of the ballooning of concentration estimates in these information-sparse regions. The eLPM and SRS samples were more effective at reducing these overall prediction errors at lower sample sizes, but the gap to the pLPM designs disappeared given large enough sample sizes (see Figure 5.10). At the same time, the prediction errors of pLPM designs within the plume remained ahead of the other two methods regardless of sample size (see Figure 5.11). Hence, given a sufficiently large sample size, the pLPM approach could deliver the same prediction accuracy over the whole domain and provide additional precision for the estimation of concentrations within the plume. Consequently, the pLPM approach also performed better at estimating the total plume mass than eLPM and SRS designs. The same conclusions were obtained for the convex hull as for the whole domain (see Section 5.3.6), supporting the real-world application of pLPM designs.

It was also identified in the results, that the difference between eLPM and SRS designs in terms of estimating the concentration surface was negligible despite the higher degree of spatial bal-

ance in eLPM samples (see Figure 5.9). Further investigation revealed that given a sufficient number of sampling events or sample sizes, the random spatial samples obtained via SRS were adequate enough to produce similar estimates as eLPM designs despite the lower degree of spatial balance. The difference between the two approaches is more relevant when considering individual sampling events separately. This result also highlights the strengths of a spatiotemporal modelling approach, which considers all information across time, thus mitigating the detrimental effects of spatially unbalanced samples. The use of spatially balanced sampling designs is still recommended over SRS due to the resulting higher estimation precision in spatial models. When drawing repeated probability samples over a certain monitoring period, the chances of obtaining multiple non-representative spatial samples is lower when spatially balanced designs are applied.

In the proposed implementation, the kernel function is automatically adjusted by the number of monitoring wells in the network and the estimated standard deviation of well-plume distance predictions (see Section 5.1.6). The adjustments reflect the amount of well combinations and the uncertainty of the predictions respectively. Given a network with only a few wells, there is no need to apply strong prioritisation, unlike in larger networks. Similarly, if a distance prediction is associated with high uncertainty, a stronger prioritisation could be applied to reduce the chance of missing a well that has come into contact with the plume. In addition to this proposed implementation, the developed approach still offers the ability to fine-tune the strictness of well prioritisation through the manual adjustment of the kernel function.

Based on the work presented in this thesis, two main areas of potential development were identified for pLPM designs. Both of these are related to the strictness of well prioritisation. The estimation of the historic concentration surface via the P-splines modelling framework is not considered here, however, it should be noted that a more accurate delineation of the CoPC plume and modelling of spatial and temporal trends is crucial for an effective sampling design. The first potential improvement is the prediction of future distances between monitoring wells and plume. Trends in the time-series of CoPC concentrations in individual wells can be very weak if present at all as well as very strong. Therefore, it is important to have a robust and adaptable modelling framework that is able to reliably model these trends and allow for the extrapolation of results to future time points. The current implementation uses GLMs with a log link function, assuming that an exponential decreasing trend will be characteristic of most well-plume distance time-series. This is due to the way the CoPC plume spreads. However, this is not always the case, and this can result in unreliable distance predictions, which subsequently affect the inclusion probabilities. The second potential improvement could be the adjustment of the kernel function. The proposed approach using the number of monitoring wells in the network and the estimated standard deviation of predictions is sensible but additional, more exhaustive analysis of this aspect of the pLPM design could further improve its reliability.

In summary, strong empirical evidence has been collected through the simulation study and additional analyses that indicates that the proposed pLPM sampling design can provide benefits for supporting the modelling of long-term groundwater contamination monitoring data. pLPM designs were shown to be more precise at estimating the CoPC concentration surface of the plume compared to traditional practices and other probability sampling methods such as SRS and LPM with equal inclusion probabilities.

7.3.2 Evaluating Historic Sampling Intensity

The aim of Chapter 6 was to assess practical aspects related to the implementation of spatiotemporally balanced sampling designs for groundwater quality monitoring. Section 6.1 proposed an innovative application of these designs for evaluating the sampling intensity of historic groundwater monitoring data sets. The proposed approach aims to evaluate whether past sampling practices resulted in under- or oversampling of the site with respect to the estimation of the CoPC concentration surface. More optimal future sampling designs can then be suggested that either reduce or increase sampling intensity based on the assessment. The proposed application works by taking increasingly smaller spatiotemporally balanced subsamples of the available historic observations and evaluating the resulting concentration surface estimates (see Figure 6.1). The approach was explored in a simulation study using the hypothetical plume data sets introduced in Chapter 1.4.2. If similar concentration surfaces could be estimated using a smaller subsample as using the complete data set, that would indicate oversampling (see Figure 6.4a). On the other hand, if the reduced subsample resulted in higher prediction errors, that would be indicative of undersampling (see Figure 6.6). To allow the application of the proposed approach on actual groundwater quality data, where the real concentration surface is unknown, a proxy to the prediction error of the concentration surface was proposed, which was the prediction error of the data points not included in the subsamples. The flowchart summarising the application of the proposed analysis for case study data sets is shown on Figure 6.2.

The simulation study results showed a clear exponentially decreasing trend in prediction error with increasing sample size (see Figure 6.4a), where the prediction errors reached an asymptote already at much smaller sample sizes. This indicated that similar prediction precision could have been achieved using a much smaller, spatially and temporally balanced sampling design. The prediction errors of the omitted observations reflected the exponentially decreasing trend (see Figure 6.4b), supporting the use of this metric as a proxy to concentration surface estimates. The results of the analysis on the case study data (introduced in Chapter 1.4.4) also exhibited the same trend, indicating that a smaller, spatially and temporally balanced sampling design would have been sufficient to reach similar estimates. To ensure that the decreasing trend is present due to oversampling and is not a feature of the prediction error calculation, the case

study data was modified to represent a clearly undersampled scenario. The results showed that in an undersampled network, the relationship between prediction error and sample size exhibits a linearly decreasing trend. This means that any additional sample included in the design will help decrease prediction error.

In conclusion, the proposed analysis could be used to identify over- and undersampling in historic groundwater quality monitoring data by looking at the relationship between the prediction errors of excluded points and the ratio of observations used in the subsample models. If the relationship exhibits an exponentially decreasing trend, sampling intensity could be reduced in the future by using spatiotemporally balanced sampling designs. If the relationship exhibits a linearly decreasing trend, sampling intensity should be increased in the future, until the exponential trend is achieved. Thus, there is strong empirical evidence that the proposed sampling intensity analysis could contribute to optimising groundwater monitoring practices. However, future work should focus on further reinforcing that the proposed evaluation metric is adequate. Another area of potential development would be devising an approach that uses the information obtained in the sampling intensity analysis to generate future sampling designs automatically. A reasonable approach would be for example identifying the optimal sample size based on the observed trend and generating a spatially and temporally balanced sample for a monitoring period of the same length in the future.

7.3.3 Spatial & Temporal Resolution Trade-Off

The aim of Section 6.2 in Chapter 6 was to assess the trade-off between high spatial versus high temporal resolution in groundwater quality monitoring data sets in terms of how well they support the precision of estimating concentration surfaces. Given a certain total sample size for a monitoring period, a sampling design can prescribe more frequent sampling events with fewer samples collected each time, or less frequent sampling events with more samples collected each time. The former results in temporally higher but spatially lower resolution data, while the latter results in spatially higher but temporally lower resolution. The question was whether this trade-off has any impact on how well the pLPM, eLPM and SRS designs can support the spatiotemporal modelling framework. It should be noted that in practice, lower sampling frequencies and higher per-event sample sizes are often preferred due to the cost and inconvenience of transportation.

The simulation study presented in Chapter 5 was modified to facilitate the investigation of the spatial versus temporal resolution trade-off. The total sample size determined the possible combinations of number of sampling events and sample sizes per event (see Table 6.1). A scheme was also devised that determined the distribution of sampling events based on the number of

events used in each scenario (see Table 6.2). Figure 6.7 shows the flowchart summary of the modified simulation study. The sampling designs were evaluated using the prediction error of concentration surfaces (over the whole domain and over the plume only) and plume mass estimates.

The results showed that for estimating the concentration surface over the network, spatially higher resolution data were preferred, as they achieved lower prediction errors than temporally high resolution data given the same sample size. This trend was stronger at higher total sample sizes and with the pLPM designs (see Figure 6.8). At the most extreme low sampling frequency however, an increase in prediction errors was observed, indicating that the sampling events were too far in time for the model to interpolate accurately. On the other hand, for estimating the concentration surface over the plume only, the pLPM approach achieved lower prediction errors with temporally higher resolution data. The plume mass estimation was also more precise given temporally high resolution data.

In conclusion, it has been shown that the modelling precision of long-term groundwater probability sampling data does not only depend on total sample size, but also sampling frequency and the spatial resolution of the data at each time point. Different probability sampling designs also perform differently based on the spatial and temporal resolution of data sets. An optimal long-term groundwater sampling design should strike a good balance between spatial and temporal resolution to maximise the amount of information obtained on plume concentrations, whilst not sacrificing the precision of estimating its spatial characteristics over the monitoring network. Testing a wider range of sampling frequencies and sample sizes in various case studies could provide further insights into the trade-off between spatial and temporal resolution in long-term groundwater quality monitoring data. Potential future development should aim to incorporate this trade-off analysis in long-term sampling designs to automatically optimise sampling frequencies and sample sizes.

7.4 Summary

To summarise, the most important limitation of the presented methods is relying on the spatiotemporal P-splines models to assess the effectiveness of the sampling designs. As shown in Chapter 2, these models can sometimes lead to ballooning when estimating solute concentration surfaces, which is an issue for groundwater monitoring data due to their common feature of low spatial and temporal resolution. Despite robust, automatic smoothing parameter selection, manual fine-tuning is sometimes necessary. Ballooning adversely affects prediction errors, which in turn affect the assessment of the sampling design. Other non-spline-based spatiotemporal modelling approaches could be used to perform the presented analyses and compare results.

As highlighted in Section 7.3.1, future research should also focus on robust forecasting of distances between monitoring wells and CoPC plumes to improve the effectiveness of proportional weight sampling designs. Additionally, the fine-tuning of the kernel function used to determine appropriate sample inclusion weights should be further investigated.

In conclusion, the main aim of this thesis was to develop innovative, statistical approaches that can contribute to the optimisation of long-term groundwater quality monitoring designs. The benefits of using a spatiotemporal modelling framework to analyse groundwater contamination systems has been well established in the literature (McLean et al. [2019]), but the predictive power of any model is dependent on the quality of the data. A chief contribution of the work undertaken in this thesis is the investigation of how much more precision can be gained form spatiotemporal modelling, if the data are collected specifically to maximise the effectiveness of the model.

Throughout this thesis, different approaches have been explored for analysing groundwater quality monitoring data using spatiotemporal modelling frameworks, the characteristics of long-term groundwater monitoring data sets have been assessed, and methodologies have been put forward that can aid groundwater monitoring practitioners in designing more sustainable sampling strategies whilst maximising the amount of information that can be gained from monitoring well networks. Some of the proposed methodologies already have real-world impact as they have been implemented in open-source groundwater data analysis software. Moreover, the implementation of the contributions presented in Chapters 5 and 6 into a single, data-driven spatiotemporal sampling design framework presents an opportunity to develop a tool to facilitate and promote probability sampling in groundwater quality monitoring. This tool could analyse historic data to determine an adequate total sample size, optimise sampling frequency and volume per sampling event and generate a long-term sampling design that maximises the effectiveness of future data analyses whilst minimising costs. This could help move groundwater monitoring practices towards a more statistically motivated direction, which would improve the protection of this important resource.

Bibliography

Abi, Naeimeh, Moradi, Mohammad, Salehi, Mohammad, Brown, Jennifer, Al-Khayat, Jassim A., and Moltchanova, Elena (2017). "Application of Balanced Acceptance Sampling to an Intertidal Survey". In: *Journal of Landscape Ecology(Czech Republic)* 10.1. ISSN: 18054196. DOI: 10.1515/jlecol-2017-0012.

Akaike, Hirotogu (1998). "Information Theory and an Extension of the Maximum Likelihood Principle". In: *Selected Papers of Hirotugu Akaike*. Ed. by Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa. New York, NY: Springer New York, pp. 199–213. ISBN: 978-1-4612-1694-0. DOI: 10.1007/978-1-4612-1694-0_15. URL: https://doi.org/10.1007/978-1-4612-1694-0_15.

Andersson, Martin and Gråsjö, Urban (Mar. 1, 2009). "Spatial dependence and the representation of space in empirical models". In: *The Annals of Regional Science* 43.1, pp. 159–180. ISSN: 1432-0592. DOI: 10.1007/s00168-008-0211-5. URL: https://doi.org/10.1007/s00168-008-0211-5 (visited on 11/05/2024).

Arbia, G. (1993). "The Use of GIS in Spatial Statistical Surveys". In: *International Statistical Review / Revue Internationale de Statistique* 61.2, pp. 339–359. ISSN: 03067734, 17515823. URL: http://www.jstor.org/stable/1403632 (visited on 10/24/2023).

Bear, Jacob and Cheng, Alexander H.-D. (2010). *Modeling Groundwater Flow and Contaminant Transport*. Dordrecht: Springer Netherlands. DOI: 10.1007/978-1-4020-6682-5. URL: https://link.springer.com/10.1007/978-1-4020-6682-5 (visited on 11/05/2024).

Belsley, David A, Kuh, Edwin, and Welsch, Roy E (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.

Benedetti, Roberto, Piersimoni, Federica, and Postiglione, Paolo (2017). "Spatially Balanced Sampling: A Review and A Reappraisal". In: *International Statistical Review* 85.3, pp. 439–454. DOI: https://doi.org/10.1111/insr.12216. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12216. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12216.

Bondesson, Lennart and Thorburn, Daniel (2008). "A List Sequential Sampling Method Suitable for Real-Time Sampling". In: *Scandinavian Journal of Statistics* 35.3, pp. 466–483. DOI: https://doi.org/10.1111/j.1467-9469.2008.00596.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9469.2008.00596.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9469.2008.00596.x.

Box, G. E. P. and Cox, D. R. (1964). "An analysis of transformations". English. In: *J. R. Stat. Soc., Ser. B* 26, pp. 211–243. ISSN: 0035-9246.

Chakrabarti, Arijit and Ghosh, Jayanta K. (2011a). "AIC, BIC and Recent Advances in Model Selection". In: *Philosophy of Statistics*. Ed. by Prasanta S. Bandyopadhyay and Malcolm R. Forster. Vol. 7. Handbook of the Philosophy of Science. Amsterdam: North-Holland, pp. 583–605. DOI: https://doi.org/10.1016/B978-0-444-51862-0.50018-6. URL: https://www.sciencedirect.com/science/article/pii/B9780444518620500186.

– (2011b). "AIC, BIC and Recent Advances in Model Selection". In: *Philosophy of Statistics*. Ed. by Prasanta S. Bandyopadhyay and Malcolm R. Forster. Vol. 7. Handbook of the Philosophy of Science. Amsterdam: North-Holland, pp. 583–605. DOI: https://doi.org/10.1016/B978-0-444-51862-0.50018-6. URL: https://www.sciencedirect.com/science/article/pii/B9780444518620500186.

Chang, Winston, Cheng, Joe, Allaire, JJ, Sievert, Carson, Schloerke, Barret, Xie, Yihui, Allen, Jeff, McPherson, Jonathan, Dipert, Alan, and Borges, Barbara (2021). *shiny: Web Application Framework for R*. R package version 1.6.0. URL: https://CRAN.R-project.org/package=shiny.

Claeskens, Gerda, KRIVOBOKOVA, TATYANA, and OPSOMER, JEAN D. (2009). "Asymptotic properties of penalized spline estimators". In: *Biometrika* 96.3, pp. 529–544. ISSN: 00063444, 14643510. URL: http://www.jstor.org/stable/27798846 (visited on 02/14/2025).

Contamination of Groundwater | U.S. Geological Survey (2024). URL: https://www.usgs.gov/special-topics/water-science-school/science/contamination-groundwater (visited on 10/17/2024).

Cook, R. Dennis (1977). "Detection of Influential Observation in Linear Regression". In: *Technometrics* 19.1, pp. 15–18. ISSN: 00401706. URL: http://www.jstor.org/stable/1268249 (visited on 11/11/2022).

Corbeil, R. R. and Searle, S. R. (Feb. 1, 1976). "Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model". In: *Technometrics* 18.1, pp. 31–38. ISSN: 0040-1706. DOI: 10.1080/00401706.1976.10489397. URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1976.10489397 (visited on 11/23/2024).

Cressie, Noel A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons, Ltd. ISBN: 9781119115151. DOI: https://doi.org/10.1002/9781119115151.

Dam-Bates, Paul van, Gansell, Oliver, and Robertson, Blair (2018). "Using balanced acceptance sampling as a master sample for environmental surveys". In: *Methods in Ecology and Evolution* 9.7, pp. 1718–1726. DOI: https://doi.org/10.1111/2041-210X.13003. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13003. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13003.

Daughney, Christopher J., Raiber, Matthias, Moreau-Fournier, Magali, Morgenstern, Uwe, and Raaij, Rob van der (Feb. 1, 2012). "Use of hierarchical cluster analysis to assess the representativeness of a baseline groundwater quality monitoring network: comparison of New Zealand's national and regional groundwater monitoring programs". In: *Hydrogeology Journal* 20.1, pp. 185–200. ISSN: 1435-0157. DOI: 10.1007/s10040-011-0786-2. URL: https://doi.org/10.1007/s10040-011-0786-2 (visited on 02/13/2025).

Deville, Jean-Claude and Tillé, Yves (1998). "Unequal Probability Sampling Without Replacement Through a Splitting Method". In: *Biometrika* 85.1. Publisher: [Oxford University Press, Biometrika Trust], pp. 89–101. ISSN: 0006-3444. URL: https://www.jstor.org/stable/2337311 (visited on 11/17/2024).

– (Dec. 2004). "Efficient balanced sampling: The cube method". In: *Biometrika* 91.4, pp. 893–912. ISSN: 0006-3444. DOI: 10.1093/biomet/91.4.893. URL: https://doi.org/10.1093/biomet/91.4.893.

Eilers, Paul H. C. and Marx, Brian D. (1996). "Flexible smoothing with B-splines and penalties". In: *Statistical Science* 11.2, pp. 89–121. DOI: 10.1214/ss/1038425655. URL: https://doi.org/10.1214/ss/1038425655.

Evers, L., Molinari, D. A., Bowman, A. W., Jones, W. R., and Spence, M. J. (2015). "Efficient and automatic methods for flexible regression on spatiotemporal data, with applications to groundwater monitoring". In: *Environmetrics* 26.6, pp. 431–441. DOI: https://doi.org/10.1002/env.2347.

Fahrmeir, Ludwig, Kneib, Thomas, Lang, Stefan, and Marx, Brian (2013). *Regression*. Berlin: Springer. DOI: 10.1007/978-3-642-34333-9.

Falah, Fatemeh, Nejad, Samira Ghorbani, Rahmati, Omid, Daneshfar, Mania, and Zeinivand, Hossein (2017). "Applicability of generalized additive model in groundwater potential modelling and comparison its performance by bivariate statistical methods". In: *Geocarto International* 32.10, pp. 1069–1089. DOI: 10.1080/10106049.2016.1188166.

Fang, X. and Chan, K.-S. (2015). "Generalized Additive Models with Spatio-temporal Data". In: *Environmental and Ecological Statistics* 22.1, pp. 61–86. DOI: https://doi.org/10.1007/s10651-014-0283-6.

Farlin, J., Gallé, T., Pittois, D., Bayerle, M., and Schaul, T. (June 1, 2019). "Groundwater quality monitoring network design and optimisation based on measured contaminant concentration and taking solute transit time into account". In: *Journal of Hydrology* 573, pp. 516–523. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2019.01.067. URL: https://www.sciencedirect.com/science/article/pii/S0022169419301489 (visited on 10/24/2023).

Gibbons, Robert D., Bhaumik, Dulal K., and Aryal, Subhash (Oct. 2009). *Statistical Methods for Groundwater Monitoring: Second Edition*. English. Publisher Copyright: © 2009 John Wiley & Sons, Inc. All rights reserved. John Wiley and Sons Inc. ISBN: 9780470164969. DOI: 10.1002/9780470549933.

Golub, Gene H., Heath, Michael, and Wahba, Grace (May 1, 1979). "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter". In: *Technometrics* 21.2, pp. 215–223. ISSN: 0040-1706. DOI: 10.1080/00401706.1979.10489751. URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1979.10489751 (visited on 11/23/2024).

Grafström, Anton (2012). "Spatially correlated Poisson sampling". In: *Journal of Statistical Planning and Inference* 142.1, pp. 139–147. ISSN: 0378-3758. DOI: https://doi.org/10.1016/j.jspi.2011.07.003. URL: https://www.sciencedirect.com/science/article/pii/S0378375811002734.

Grafström, Anton and Lisic, Jonathan (2016). *BalancedSampling: Balanced and Spatially Balanced Sampling*. Version R package version 1.5.2. URL: 10.32614/CRAN.package.BalancedSampling.

– (2018). *SamplingBigData: Sampling Methods for Big Data*. Version R package version 1.0.0. URL: https://doi.org/10.32614/CRAN.package.SamplingBigData.

Grafström, Anton, Lundström, Niklas, and Schelin, Lina (2012). "Spatially Balanced Sampling through the Pivotal Method". In: *Biometrics* 68.2, pp. 514–520. DOI: https://doi.org/10.1111/j.1541-0420.2011.01699.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2011.01699.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2011.01699.x.

Grafström, Anton and Schelin, Lina (2014). "How to Select Representative Samples". In: *Scandinavian Journal of Statistics* 41.2, pp. 277–290. ISSN: 1467-9469. DOI: 10.1111/sjos.12016. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12016 (visited on 06/26/2024).

Grafström, Anton and Tillé, Yves (2013). "Doubly balanced spatial sampling with spreading and restitution of auxiliary totals". In: *Environmetrics* 24.2, pp. 120–131. DOI: https://doi.org/

10.1002/env.2194. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/env.2194. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2194.

Groundwater - European Commission (Oct. 9, 2024). URL: https://environment.ec.europa.eu/topics/water/groundwater_en (visited on 10/16/2024).

Guisan, A., Edwards, T.C., and Hastie, T. (2002). "Generalized Linear and Additive Models in Studies of Species Distributions: Setting the Scene". In: *Ecological Modelling* 157.2, pp. 89–100. DOI: https://doi.org/10.1016/S0304-3800(02)00204-1.

Hadi, Ali S. (1992). "A new measure of overall potential influence in linear regression". In: *Computational Statistics & Data Analysis* 14.1, pp. 1–27. ISSN: 0167-9473. DOI: https://doi.org/10.1016/0167-9473(92)90078-T. URL: https://www.sciencedirect.com/science/article/pii/016794739290078T.

Hájek, J. and Dupač, V. (1981). *Sampling from a Finite Population*. Statistics Series. M. Dekker. ISBN: 978-0-8247-1291-4. URL: https://books.google.hu/books?id=DxHvAAAAMAAJ.

Halton, J. H. (1960). "On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals". In: *Numerische Mathematik* 2 (1), pp. 84–90. ISSN: 0945-3245. DOI: 10.1007/BF01386213. URL: https://doi.org/10.1007/BF01386213 (visited on 12/01/1960).

Harbaugh, A.W., Banta, Edward, Hill, Mary, and McDonald, M.G. (2000). "MODFLOW-2000, the U.S. Geological Survey modular ground-water model — User guide to modularization concepts and the Ground-Water Flow Process". In: *Open-File Report 00-92. U.S. Geological Survey*.

Hastie, T. and Tibshirani, R. (1986). "Generalized Additive Models". In: *Statistical Science* 1.3, pp. 297–310. DOI: https://doi.org/10.1214/ss/1177013604.

Hornberger, George M. and Perrone, Debra (Sept. 3, 2019). *Water Resources: Science and Society*. Google-Books-ID: ONmmDwAAQBAJ. JHU Press. 281 pp. ISBN: 978-1-4214-3295-3.

Horvitz, D. G. and Thompson, D. J. (1952). "A Generalization of Sampling Without Replacement From a Finite Universe". In: *Journal of the American Statistical Association* 47.260, pp. 663–685. ISSN: 01621459, 1537274X. URL: http://www.jstor.org/stable/2280784 (visited on 02/26/2025).

Hosseini, Marjan and Kerachian, Reza (Aug. 4, 2017a). "A Bayesian maximum entropy-based methodology for optimal spatiotemporal design of groundwater monitoring networks". In: *Environmental Monitoring and Assessment* 189.9, p. 433. ISSN: 1573-2959. DOI: 10.1007/

s10661 - 017 - 6129 - 6. URL: https : / / doi . org / 10 . 1007 / s10661 - 017 - 6129 - 6 (visited on 03/07/2024).

Hosseini, Marjan and Kerachian, Reza (Sept. 1, 2017b). "A data fusion-based methodology for optimal redesign of groundwater monitoring networks". In: *Journal of Hydrology* 552, pp. 267–282. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2017.06.046. URL: https://www.sciencedirect.com/science/article/pii/S0022169417304493 (visited on 03/07/2024).

– (May 1, 2023). "Optimal redesign of coastal groundwater quality monitoring networks under uncertainty: application of the theory of belief functions". In: *Environmental Science and Pollution Research* 30.21, pp. 59701–59718. ISSN: 1614-7499. DOI: 10.1007/s11356-023-26764-1. URL: https://doi.org/10.1007/s11356-023-26764-1 (visited on 11/13/2024).

Ito, Yasushi (2015). "Voronoi Tessellation". In: *Encyclopedia of Applied and Computational Mathematics*. Ed. by Björn Engquist. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1546–1547. ISBN: 978-3-540-70529-1. DOI: 10 . 1007 / 978 - 3 - 540 - 70529 - 1 _ 315. URL: https : //doi.org/10.1007/978-3-540-70529-1_315.

Jones, W. R., Rock, L., Miller, C., McLean, M.I., Alexander, C., Bowman, A.W., and Radvanyi, P. (Nov. 2023). "GWSDAT User Manual". In: URL: http://gwsdat.net/gwsdat_manual/.

Jones, W. R., Rock, Luc, Wesch, Alexandra, Marzusch, Emanuel, and Low, Marnie (2022). "Groundwater Spatiotemporal Data Analysis Tool: Case Studies, New Features and Future Developments". In: *Groundwater Monitoring & Remediation* 42.3, pp. 14–22. ISSN: 1745-6592. DOI: 10.1111/gwmr.12522. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/gwmr.12522 (visited on 10/30/2024).

Jones, W. R., Spence, Michael J., Bowman, Adrian W., Evers, Ludger, and Molinari, Daniel A. (2014). "A software tool for the spatiotemporal analysis and reporting of groundwater monitoring data". In: *Environmental Modelling & Software* 55, pp. 242–249. ISSN: 1364-8152. DOI: https://doi.org/10.1016/j.envsoft.2014.01.020. URL: https://www.sciencedirect.com/science/article/pii/S1364815214000309.

Kendall, M. G. (1938). "A New Measure of Rank Correlation". In: *Biometrika* 30.1/2, pp. 81–93. ISSN: 00063444. URL: http://www.jstor.org/stable/2332226 (visited on 01/18/2023).

Kermorvant, Claire, Caill Milly, Nathalie, Bru, Noëlle, and D'Amico, Frank (2019a). "Optimizing cost-efficiency of long term monitoring programs by using spatially balanced sampling designs: The case of manila clams in Arcachon bay". In: *Ecological Informatics* 49. ISSN: 15749541. DOI: 10.1016/j.ecoinf.2018.11.005.

Kermorvant, Claire, D'Amico, Frank, Bru, Noëlle, Caill-Milly, Nathalie, and Robertson, Blair (2019b). "Spatially balanced designs for environmental surveys". In: *Environmental Monitoring and Assessment* 191.524. DOI: https://doi.org/10.1007/s10661-019-7666-y.

Kincaid, T.M. and Olsen, Anthony R. (2016). *spsurvey: Spatial Sampling Design and Analysis*. Version R package version 3.3. URL: 10.32614/CRAN.package.spsurvey.

Koski, Vilja and Eidsvik, Jo (2024). "Sampling design methods for making improved lake management decisions". In: *Environmetrics* n/a (n/a), e2842. ISSN: 1099-095X. DOI: 10.1002/env.2842. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2842 (visited on 02/15/2024).

Larsen, K. (2015). "GAM: The predictive Modeling Silver Bullet". In: *MultiThreaded, Technology at Stitch Fix*. URL: https://multithreaded.stitchfix.com/blog/2015/07/30/gam/.

Li, Jianzhu and Valliant, Richard (2009). "Survey weighted hat matrix and leverages". In: *Survey Methodology* 35.1, pp. 15–24.

Li, Peiyue, Karunanidhi, D., Subramani, T., and Srinivasamoorthy, K. (Jan. 1, 2021). "Sources and Consequences of Groundwater Contamination". In: *Archives of Environmental Contamination and Toxicology* 80.1, pp. 1–10. ISSN: 1432-0703. DOI: 10.1007/s00244-020-00805-z. URL: https://doi.org/10.1007/s00244-020-00805-z (visited on 11/05/2024).

Lisic, Jonathan and Cruze, Nathan B. (2016). "Local Pivotal Methods for Large Surveys". In: URL: https://api.semanticscholar.org/CorpusID:250263302.

Lisic, Jonathan and Grafström, Anton (Sept. 3, 2018). *SamplingBigData: Sampling Methods for Big Data*. Institution: Comprehensive R Archive Network Pages: 1.0.0. DOI: 10.32614/CRAN.package.SamplingBigData. URL: https://CRAN.R-project.org/package=SamplingBigData (visited on 11/18/2024).

Liu, Lili, Dong, Yongcheng, Kong, Ming, Zhou, Jian, Zhao, Hanbin, Wang, Yupeng, Zhang, Meng, and Wang, Zhiping (2020). "Towards the comprehensive water quality control in Lake Taihu: Correlating chlorphyll a and water quality parameters with generalized additive model". In: *Science of The Total Environment* 705, p. 135993. ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2019.135993.

Loaiciga, Hugo A. (1989). "An optimization approach for groundwater quality monitoring network design". In: *Water Resources Research* 25.8, pp. 1771–1782. DOI: https://doi.org/10.1029/WR025i008p01771. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/WR025i008p01771. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR025i008p01771.

Lumley, T. (2014). *survey: analysis of complex survey samples*. Version R package version 3.30. URL: 10.32614/CRAN.package.survey.

Lyons, Kevin J., Ikonen, Jenni, Hokajärvi, Anna-Maria, Räsänen, Teemu, Pitkänen, Tarja, Kauppinen, Ari, Kujala, Katharina, Rossi, Pekka M., and Miettinen, Ilkka T. (Mar. 15, 2023). "Monitoring groundwater quality with real-time data, stable water isotopes, and microbial community analysis: A comparison with conventional methods". In: *Science of The Total Environment* 864, p. 161199. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2022.161199. URL: https://www.sciencedirect.com/science/article/pii/S0048969722083036 (visited on 11/05/2024).

Marra, G. and Wood, S.N. (2011). "Practical Variable Selection for Generalized Additive Models". In: *Computational Statistics and Data Analysis* 55.7, pp. 2372–2387. DOI: https://doi.org/10.1016/j.csda.2011.02.004.

MARRA, GIAMPIERO and WOOD, SIMON N. (2012). "Coverage Properties of Confidence Intervals for Generalized Additive Model Components". In: *Scandinavian Journal of Statistics* 39.1, pp. 53–74. DOI: https://doi.org/10.1111/j.1467-9469.2011.00760.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9469.2011.00760.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9469.2011.00760.x.

McDonald, T., McDonald, A., Kleinsausser, M., and Emmons, S. (2016). *SDraw: Spatially Balanced Samples of Spatial Objects*. Version R package version 2.1.13. URL: https://rdrr.io/cran/SDraw/.

McLean, M.I. (Dec. 2018). "Spatio-temporal Models for the Analysis and Optimisation of Groundwater Quality Monitoring Networks". PhD thesis. University of Glasgow.

McLean, M.I., Evers, L., Bowman, A.W., Bonte, M., and Jones, W. R. (2019). "Statistical modelling of groundwater contamination monitoring data: A comparison of spatial and spatiotemporal methods". In: *Science of The Total Environment* 652, pp. 1339–1346. DOI: https://doi.org/10.1016/j.scitotenv.2018.10.231.

Meray, Aurelien O., Sturla, Savannah, Siddiquee, Masudur R., Serata, Rebecca, Uhlemann, Sebastian, Gonzalez-Raymat, Hansell, Denham, Miles, Upadhyay, Himanshu, Lagos, Leonel E., Eddy-Dilek, Carol, and Wainwright, Haruko M. (May 3, 2022). "PyLEnM: A Machine Learning Framework for Long-Term Groundwater Contamination Monitoring Strategies". In: *Environmental Science & Technology* 56.9. Publisher: American Chemical Society, pp. 5973–5983. ISSN: 0013-936X. DOI: 10.1021/acs.est.1c07440. URL: https://doi.org/10.1021/acs.est.1c07440 (visited on 02/13/2023).

Merli, Marcello (July 2005a). "Outlier recognition in crystal-structure least-squares modelling by diagnostic techniques based on leverage analysis". In: *Acta Crystallographica Section A* 61.4, pp. 471–477. DOI: 10.1107/S010876730501809X. URL: https://doi.org/10.1107/S010876730501809X.

– (July 2005b). "Outlier recognition in crystal-structure least-squares modelling by diagnostic techniques based on leverage analysis". In: *Acta Crystallographica Section A* 61.4, pp. 471–477. DOI: 10.1107/S010876730501809X. URL: https://doi.org/10.1107/S010876730501809X.

Morton, Richard and Henderson, Brent L. (2008). "Estimation of nonlinear trends in water quality: An improved approach using generalized additive models". In: *Water Resources Research* 44.7. DOI: https://doi.org/10.1029/2007WR006191.

Mosavi, A., Sajedi Hosseini, F., Choubin, B., Goodarzi, M., Dineva, A.A., and Rafiei Sardooi, E. (2021). "Ensemble Boosting and Bagging Based Machine Learning Models for Groundwater Potential Prediction". In: *Water Resources Management* 35, pp. 23–37. DOI: https://doi.org/10.1007/s11269-020-02704-3.

Motevalli, Alireza, Pourghasemi, Hamid Reza, Hashemi, Hossein, and Gholami, Vahid (2019). "25 - Assessing the Vulnerability of Groundwater to Salinization Using GIS-Based Data-Mining Techniques in a Coastal Aquifer". In: *Spatial Modeling in GIS and R for Earth and Environmental Sciences*. Ed. by Hamid Reza Pourghasemi and Candan Gokceoglu. Elsevier, pp. 547–571. ISBN: 978-0-12-815226-3. DOI: https://doi.org/10.1016/B978-0-12-815226-3.00025-9.

Naghibi, Seyed Amir, Moghaddam, Davood Davoodi, Kalantar, Bahareh, Pradhan, Biswajeet, and Kisi, Ozgur (2017). "A comparative assessment of GIS-based data mining models and a novel ensemble model in groundwater well potential mapping". In: *Journal of Hydrology* 548, pp. 471–483. DOI: https://doi.org/10.1016/j.jhydrol.2017.03.020.

Nowak, Wolfgang, Rubin, Yoram, and Barros, Felipe P. J. de (2012). "A hypothesis-driven approach to optimize field campaigns". In: *Water Resources Research* 48.6. _eprint: https://onlinelibrary.w ISSN: 1944-7973. DOI: 10.1029/2011WR011016. URL: https://onlinelibrary.wiley.com/doi/abs/10.1029/2011WR011016 (visited on 02/13/2025).

Ohmer, Marc, Liesch, Tanja, and Wunsch, Andreas (Aug. 5, 2022). "Spatiotemporal optimization of groundwater monitoring networks using data-driven sparse sensing methods". In: *Hydrology and Earth System Sciences* 26.15. Publisher: Copernicus GmbH, pp. 4033–4053. ISSN: 1027-5606. DOI: 10.5194/hess-26-4033-2022. URL: https://hess.copernicus.org/articles/26/4033/2022/ (visited on 11/13/2024).

Olsen, Anthony R. (2004). "Spatially-Balanced Survey Designs for Aquatic Resources". In: URL: https://archive.epa.gov/nheerl/arm/web/pdf/grts_ss.pdf.

– (2023). *Spatially balanced survey design for groundwater using existing wells*. URL: https://cfpub.epa.gov/si/si_public_record_Report.cfm?Lab=NHEERL&dirEntryID=59587 (visited on 10/23/2023).

Prentius, Wilmer and Grafström, Anton (2024). "How to find the best sampling design: A new measure of spatial balance". In: *Environmetrics* 35.7, e2878. ISSN: 1099-095X. DOI: 10.1002/env.2878. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2878 (visited on 10/25/2024).

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Radvanyi, Peter, Miller, Claire, Alexander, Craig, Low, Marnie, Jones, W. R., and Rock, Luc (July 21, 2023). *Computationally Efficient Ranking of Groundwater Monitoring Locations*. Conference Name: 37th International Workshop on Statistical Modelling (IWSM) ISBN: 9783947323425 Pages: 332-338 Place: Dortmund, Germany. URL: https://eprints.gla.ac.uk/303347/.

Ranganai, Edmore (2016). "On studentized residuals in the quantile regression framework". In: *Springerplus* 5.1, pp. 1–11.

Räty, Minna, Heikkinen, Juha, Korhonen, Kari T., Peräsaari, Jouni, Ihalainen, Antti, Pitkänen, Juho, and Susanna Kangas, Annika (Oct. 3, 2019). "Effect of cluster configuration and auxiliary variables on the efficiency of local pivotal method for national forest inventory". In: *Scandinavian Journal of Forest Research* 34.7, pp. 607–616. ISSN: 0282-7581. DOI: 10.1080/02827581.2019.1662938. URL: https://doi.org/10.1080/02827581.2019.1662938 (visited on 06/14/2024).

Räty, Minna, Kuronen, Mikko, Myllymäki, Mari, Kangas, Annika, Mäkisara, Kai, and Heikkinen, Juha (Sept. 24, 2020). "Comparison of the local pivotal method and systematic sampling for national forest inventories". In: *Forest Ecosystems* 7.1, p. 54. ISSN: 2197-5620. DOI: 10.1186/s40663-020-00266-9. URL: https://doi.org/10.1186/s40663-020-00266-9 (visited on 06/14/2024).

Ravindra, K., Rattan, P., Mor, S., and Aggarwal, A.N. (2019). "Generalized Additive Models: Building Evidence of Air Pollution, Climate Change and Human Health". In: *Environment International* 132, p. 104987. DOI: https://doi.org/10.1016/j.envint.2019.104987.

Roberge, Cornelia, Grafström, Anton, and Ståhl, Göran (Mar. 2017). "Forest damage inventory using the local pivotal sampling method". In: *Canadian Journal of Forest Research* 47.3. Pub-

lisher: NRC Research Press, pp. 357–365. ISSN: 0045-5067. DOI: 10.1139/cjfr-2016-0411. URL: https://cdnsciencepub.com/doi/abs/10.1139/cjfr-2016-0411 (visited on 06/14/2024).

Robertson, Blair, Brown, J. A., Mcdonald, T., and Jaksons, P. (2013). "BAS: Balanced Acceptance Sampling of Natural Resources". In: *Biometrics* 69.3. ISSN: 15410420. DOI: 10.1111/biom.12059.

Robertson, Blair, McDonald, Trent, Price, Chris, and Brown, Jennifer (2018). "Halton iterative partitioning: spatially balanced sampling via partitioning". In: *Environmental and Ecological Statistics* 25.3, pp. 305–323. DOI: https://doi.org/10.1007/s10651-018-0406-6.

Rügner, Hermann, Finkel, Michael, Kaschl, Arno, and Bittens, Martin (Oct. 1, 2006). "Application of monitored natural attenuation in contaminated land management—A review and recommended approach for Europe". In: *Environmental Science & Policy* 9.6, pp. 568–576. ISSN: 1462-9011. DOI: 10.1016/j.envsci.2006.06.001. URL: https://www.sciencedirect.com/science/article/pii/S1462901106000876 (visited on 11/05/2024).

Saad, Rami, Wallerman, Jörgen, Holmgren, Johan, and Lämås, Tomas (Jan. 26, 2016). "Local pivotal method sampling design combined with micro stands utilizing airborne laser scanning data in a long term forest management planning setting". In: *Silva Fennica* 50.2. URL: https://www.silvafennica.fi/article/1414/related/1414 (visited on 06/14/2024).

Schmidt, Franziska, Wainwright, Haruko M., Faybishenko, Boris, Denham, Miles, and Eddy-Dilek, Carol (July 3, 2018). "In Situ Monitoring of Groundwater Contamination Using the Kalman Filter". In: *Environmental Science & Technology* 52.13. Publisher: American Chemical Society, pp. 7418–7425. ISSN: 0013-936X. DOI: 10.1021/acs.est.8b00017. URL: https://doi.org/10.1021/acs.est.8b00017 (visited on 10/18/2024).

Schwarz, Gideon (1978). "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2, pp. 461–464. DOI: 10.1214/aos/1176344136. URL: https://doi.org/10.1214/aos/1176344136.

Shih, Tom, Rong, Yue, Harmon, Thomas, and Suffet, Mel (Jan. 1, 2004). "Evaluation of the Impact of Fuel Hydrocarbons and Oxygenates on Groundwater Resources". In: *Environmental Science & Technology* 38.1. Publisher: American Chemical Society, pp. 42–48. ISSN: 0013-936X. DOI: 10.1021/es0304650. URL: https://doi.org/10.1021/es0304650 (visited on 11/05/2024).

Singh, Ravindra and Mangat, Naurang Singh (1996). "Stratified Sampling". In: *Elements of Survey Sampling*. Ed. by Ravindra Singh and Naurang Singh Mangat. Dordrecht: Springer Netherlands, pp. 102–144. ISBN: 978-94-017-1404-4. DOI: 10.1007/978-94-017-1404-4_5. URL: https://doi.org/10.1007/978-94-017-1404-4_5 (visited on 11/05/2024).

Singhal, B. B. S. and Gupta, R. P. (2010). "Groundwater Contamination". In: *Applied Hydrogeology of Fractured Rocks: Second Edition*. Ed. by B.B.S. Singhal and R.P. Gupta. Dordrecht: Springer Netherlands, pp. 221–236. ISBN: 978-90-481-8799-7. DOI: 10.1007/978-90-481-8799-7_12. URL: https://doi.org/10.1007/978-90-481-8799-7_12 (visited on 10/17/2024).

Sorichetta, Alessandro, Ballabio, Cristiano, Masetti, Marco, Robinson Jr., Gilpin R., and Sterlacchini, Simone (2013). "A Comparison of Data-Driven Groundwater Vulnerability Assessment Methods". In: *Groundwater* 51.6, pp. 866–879. DOI: https://doi.org/10.1111/gwat.12012.

Speak, Andrew, Escobedo, Francisco J., Russo, Alessio, and Zerbe, Stefan (2018). "Comparing convenience and probability sampling for urban ecology applications". In: *Journal of Applied Ecology* 55.5, pp. 2332–2342. ISSN: 1365-2664. DOI: 10.1111/1365-2664.13167. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1365-2664.13167 (visited on 11/05/2024).

Squillace, Paul J., Pankow, James F., Korte, Nic E., and Zogorski, John S. (1997). "Review of the environmental behavior and fate of methyl tert-butyl ether". In: *Environmental Toxicology and Chemistry* 16.9, pp. 1836–1844. ISSN: 1552-8618. DOI: 10.1002/etc.5620160911. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/etc.5620160911 (visited on 11/05/2024).

Stehman, Stephen V. and Overton, W. Scott (Jan. 1, 1994). "9 Environmental sampling and monitoring". In: *Handbook of Statistics*. Vol. 12. Environmental Statistics. Elsevier, pp. 263–306. DOI: 10.1016/S0169-7161(05)80011-2. URL: https://www.sciencedirect.com/science/article/pii/S0169716105800112 (visited on 10/22/2024).

Stevens, Don L. and Olsen, Anthony R. (2003). "Variance estimation for spatially balanced samples of environmental resources". In: *Environmetrics* 14.6, pp. 593–610. DOI: https://doi.org/10.1002/env.606. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/env.606. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/env.606.

– (2004). "Spatially Balanced Sampling of Natural Resources". In: *Journal of the American Statistical Association* 99.465, pp. 262–278. DOI: 10.1198/016214504000000250.

Tillé, Yves and Matei, A. (2015). *sampling: survey sampling*. Version R package version 2.7. URL: https://CRAN.%20R-project.org/packageDsampling.

Tutz, Gerhard (Apr. 1, 2023). "Probability and non-probability samples: Improving regression modeling by using data from different sources". In: *Information Sciences* 621, pp. 424–436. ISSN: 0020-0255. DOI: 10.1016/j.ins.2022.11.032. URL: https://www.sciencedirect.com/science/article/pii/S0020025522013214 (visited on 11/05/2024).

U.S. Sustainable Remediation Forum (June 2009). "Sustainable remediation white paper - Integrating sustainable principles, practices, and metrics into remediation projects". In: *Remedi-*

ation Journal 19.3, pp. 5–114. ISSN: 1051-5658, 1520-6831. DOI: 10.1002/rem.20210. URL: https://onlinelibrary.wiley.com/doi/10.1002/rem.20210 (visited on 05/22/2024).

Water Framework Directive - European Commission (Oct. 9, 2024). URL: https://environment.ec.europa.eu/topics/water/water-framework-directive_en (visited on 10/17/2024).

Weisberg, Sanford and Fox, J (2011). *An R Companion to Applied Regression*. English. 2nd ed. Thousand Oaks: Sage.

Wöhling, Thomas, Geiges, Andreas, and Nowak, Wolfgang (2016). "Optimal Design of Multitype Groundwater Monitoring Networks Using Easily Accessible Tools". In: *Groundwater* 54.6, pp. 861–870. DOI: https://doi.org/10.1111/gwat.12430. eprint: https://ngwa.onlinelibrary.wiley.com/doi/pdf/10.1111/gwat.12430. URL: https://ngwa.onlinelibrary.wiley.com/doi/abs/10.1111/gwat.12430.

Wolf, Christof, Joye, Dominique, Smith, Tom W., and Fu, Yang-chih (July 11, 2016). *The SAGE Handbook of Survey Methodology*. Google-Books-ID: g8OMDAAAQBAJ. SAGE. 741 pp. ISBN: 978-1-4739-5905-7.

Wood, S.N. (2004). "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models". In: *Journal of the American Statistical Association* 99.467, pp. 673–686. DOI: https://doi.org/10.1198/016214504000000980.

– (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9781315370279.

– (June 2021). "Package 'mgcv': Mixed GAM Computation Vehicle with Automatic Smoothness Estimation". In: URL: https://CRAN.R-project.org/package=mgcv.

– (Sept. 10, 2024). *On Neighbourhood Cross Validation*. DOI: 10.48550/arXiv.2404.16490. arXiv: 2404.16490. URL: http://arxiv.org/abs/2404.16490 (visited on 10/31/2024).

Wood, S.N., Pya, Natalya, and Säfken, Benjamin (Oct. 1, 2016). "Smoothing Parameter and Model Selection for General Smooth Models". In: *Journal of the American Statistical Association* 111.516, pp. 1548–1563. ISSN: 0162-1459. DOI: 10.1080/01621459.2016.1180986. URL: https://doi.org/10.1080/01621459.2016.1180986 (visited on 11/08/2024).

Wood, Simon N. (Jan. 2003). "Thin Plate Regression Splines". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 65.1, pp. 95–114. ISSN: 1369-7412. DOI: 10.1111/1467-9868.00374. eprint: https://academic.oup.com/jrsssb/article-pdf/65/1/95/49799823/jrsssb_65_1_95.pdf. URL: https://doi.org/10.1111/1467-9868.00374.

Wu, Jianfeng, Zheng, Chunmiao, and Chien, Calvin C. (Mar. 1, 2005). "Cost-effective sampling network design for contaminant plume monitoring under general hydrogeological con-

ditions". In: *Journal of Contaminant Hydrology* 77.1, pp. 41–65. ISSN: 0169-7722. DOI: 10.1016/j.jconhyd.2004.11.006. URL: https://www.sciencedirect.com/science/article/pii/S0169772204001901 (visited on 11/05/2024).

Zamanirad, Mahtab, Sarraf, Amirpouya, Sedghi, Hossein, Saremi, Ali, and Rezaee, Payman (2020). "Modeling the Influence of Groundwater Exploitation on Land Subsidence Susceptibility Using Machine Learning Algorithms". In: *Natural Resources Research* 29, pp. 1127–1141. DOI: https://doi.org/10.1007/s11053-019-09490-9.

Zhang, Chunlong (Feb. 29, 2024). *Fundamentals of Environmental Sampling and Analysis*. Google-Books-ID: luX4EAAAQBAJ. John Wiley & Sons. 581 pp. ISBN: 978-1-119-77859-2.

Zhang, Shu, Mao, Guozhu, Crittenden, John, Liu, Xi, and Du, Huibin (Aug. 1, 2017). "Groundwater remediation from the past to the future: A bibliometric analysis". In: *Water Research* 119, pp. 114–125. ISSN: 0043-1354. DOI: 10.1016/j.watres.2017.01.029. URL: https://www.sciencedirect.com/science/article/pii/S0043135417300350 (visited on 11/05/2024).

Zheng, C. (1990). *MT3D, A modular three-dimensional transport model for simulation of advection, dispersion and chemical reactions of contaminants in groundwater systems*. Rockville, Maryland: S.S. Papadopulos & Associates.