McLean, Marnie Isla (2018) Spatio-temporal models for the analysis and optimisation of groundwater quality monitoring networks. PhD thesis.

https://theses.gla.ac.uk/38975/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses
https://theses.gla.ac.uk/
research-enlighten@glasgow.ac.uk

Spatio-temporal Models for the Analysis and Optimisation of Groundwater Quality Monitoring Networks

by

Marnie Isla McLean

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the
School of Mathematics and Statistics

December 2018

Declaration of Authorship

I, MARNIE ISLA MCLEAN, declare that this thesis titled, 'Spatio-temporal Models for the Analysis and Optimisation of Groundwater Quality Monitoring Networks' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- 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.

Signed: ______________________________

Date: ______________________________

"Life is not easy for any of us. But what of that? We must have perseverance and above all confidence in ourselves. We must believe that we are gifted for something and that this thing must be attained."

Marie Curie

Abstract

Commonly groundwater quality data are modelled using temporally independent spatial models. However, primarily due to cost constraints, data of this type can be sparse resulting in some sampling events only recording a few observations. With data of this nature, spatial models struggle to capture the true underlying state of the groundwater and building models with such small spatial datasets can result in unreliable predictions. This highlights the need for spatio-temporal models which 'borrow strength' from earlier sampling events and which allow interpolations of groundwater concentrations between sampling points.

To compare the relative merits of analysing groundwater quality data using spatial compared to spatio-temporal statistical models, a comparison study is presented using data from a hypothetical contaminant plume along with a real life dataset. In this study, the estimation accuracy of spatial p-spline and Kriging models are compared with spatio-temporal p-spline models. The results show that spatio-temporal methods can increase prediction efficiency markedly so that, in comparison with repeated spatial analysis, spatio-temporal methods can achieve the same level of performance but with smaller sample sizes.

For the comparison study, in the spatio-temporal p-splines model, differing levels of variability over space and time were controlled using different numbers of basis functions rather than separate smoothing parameters due to the computational expense of their optimisation. However, deciding on the number of basis functions for each dimension is subjective due to space and time being measured on different scales, and thus methodology is developed to efficiently tune two smoothing parameters. The proposed methodology exploits lower resolution models to determine starting points for the optimisation procedure allowing for each parameter to be tuned separately.

Working with spatio-temporal models can, however, pose their own problems. Due to the sporadic layout of many monitoring well networks, due to built-up urban areas and transport infrastructure, ballooning can be experienced in the predictions of these models. 'Ballooning' is a term used to describe the event where high or low predictions are made in regions with little data support. To determine when this has occurred a measure is developed to highlight when ballooning may be present in the models predictions. In addition to the measure, to try to eliminate ballooning from happening in the first place, a penalty based on the idea that the total contaminant mass should not change significantly over time is proposed. However, the preliminary results presented here indicate that further work is needed to make this effective.

It is shown that by adopting a spatio-temporal modelling framework a smoother, clearer and more accurate prediction through time can be achieved, compared to spatial modelling of individual time steps, whilst using fewer samples. This was shown using existing sampling schemes where the choice of sampling locations was made by someone with little knowledge or experience in sampling design. Sampling designs on fixed monitoring well networks are then explored and optimised through the minimisation two objective functions; the variance of the predicted plume mass (VM) and the integrated prediction variance (IV). Sampling design optimisations, using spatial and spatio-temporal p-spline models, are carried out, using a variety of numbers of wells and at various future sampling time points. The effects of well-specific sampling frequency are also investigated and it is found that both objective functions tend to propose wells for the next sampling design which have not been sampled recently.

Often, an existing monitoring well network will need to be changed, either by adding new wells or by down-scaling and removing wells. The decision to add wells to the network comes at a financial expense, so it is of paramount importance that wells are added into areas where the gain in knowledge of the region is maximised. The decision to remove a well from the network is equally important and involves a trade-off between costs saved and information lost. The design objective functions suggest a well should be added in an area where the distance to the nearest neighbouring wells is greatest.

Finally, consideration is given to optimal sampling designs when it is assumed the recorded data has multiplicative error - a common assumption in groundwater quality data. When modelling with this type of data, the response is normally log transformed prior to modelling and the predictions are then transformed back onto the original scale for interpretation. Assuming a log transformed response, the objective functions, initially presented, can be used if computation of the objective function is also on the log scale. However, if the desired scale of interpretation of the objective functions is the original scale but modelling was performed on the log scale, the resulting objective function

values cannot simply be exponentiated to give an interpretation on the original scale. Modelling on the log scale while interpreting the objective function on the original scale can be achieved by adopting a lognormal distribution for the predicted response and subsequently numerically integrating its variance to compute the IV objective function. The results indicate that the designs do differ depending on which scale interpretation of the objective function is to be made. When interpreting on the original scale the objective function favours sampling from wells where higher values were previously estimated. Unfortunately, computation of the VM objective function when assuming a lognormal distribution has not been achieved so far.

Acknowledgements

Firstly, I would like to thank my supervisors, Prof. Adrian Bowman and Dr. Ludger Evers for their guidance and help throughout the duration of my Ph.D. I have learned so much from you both and greatly appreciate you sharing your knowledge.

Thank you also to Dr. Wayne Jones and Shell for their help, providing data and for funding my work.

To the staff and students of the Statistics Department, thank you for your friendship and support. In particular, thank you to the lifelong friends I have gained in Craig and Umberto. Thank you for the memorable dining experiences and for teaching me the wonders of Photoshop. There hasn't been a day gone by where we haven't laughed. Thanks also go to Craig, Vinny and Irene for teaching me how to 'properly' go for a drink!

A big thank you to my family and friends back at home. To Catriona, Lauren and Morven, thank you for being more help than you probably realised. To my mum, thank you for putting everything into perspective whenever I felt anxious or overwhelmed and to my dad, thank you for showing me what hard work really is. To you both, thank you for teaching us that nothing is impossible. The love and encouragement we receive from you both is why I am here today.

Finally to Martyn, thank you for quite literally everything. On days when finishing seemed impossible, you were always there with love, positivity and belief. Thank you for solving my problems without saying a word, for always knowing how to make things seem better and most importantly, thank you for always being able to make me smile.

Contents

Declaration of Authorship i

Abstract iii

Acknowledgements vi

List of Figures xi

List of Tables xviii

1 Introduction 1
- 1.1 Groundwater . . . 3
  - 1.1.1 Groundwater Pollution . . . 3
- 1.2 Thesis Overview . . . 5

2 Current Nonparametric Modelling Techniques 7
- 2.1 Kernel Smoothing Methods . . . 8
  - 2.1.1 Kernel Density Estimation . . . 8
  - 2.1.2 Local Linear Regression . . . 9
- 2.2 Regression Splines . . . 10
  - 2.2.1 B-splines . . . 11
  - 2.2.2 Tensor Product Regression Splines for Multi-Dimensional Data . . . 15
  - 2.2.3 Penalised Regression Splines . . . 17
  - 2.2.4 P-splines . . . 19
  - 2.2.5 Tensor Product Penalised Regression Splines . . . 20
  - 2.2.6 Tensor Product P-splines . . . 22
  - 2.2.7 Choosing the Smoothing Parameter . . . 23
- 2.3 Bayesian P-splines . . . 25
  - 2.3.1 Bayesian Regression . . . 25
  - 2.3.2 Derivation of the Posterior Density of the Smoothing Parameter for a P-splines Model . . . 26
- 2.4 Kriging . . . 30
  - 2.4.1 Geostatistical Processes . . . 30
  - 2.4.2 Covariance Functions . . . 31

2.4.3 Prediction with a Gaussian Spatial Process - Kriging 33
2.4.4 Spatio-temporal Geostatistical Processes 34
2.4.5 Prediction with a Gaussian Spatio-temporal Process 35

3 A Comparison of Spatial and Spatio-temporal Modelling Methods for Contaminated Groundwater 37
3.1 Data Simulation 38
3.2 Results 39
3.2.1 Sampling Scenario 1 - A Realistic Design 40
3.2.2 Sampling Scenario 2 - A Full Design 46
3.2.3 Computational Time 51
3.3 Reducing the Number of Samples 51
3.4 Real Application 54
3.4.1 Results 54
3.5 Summary 57

4 Incorporating an Additional Smoothing Parameter for the Temporal Component 58
4.1 Motivation and Model Formulation 58
4.1.1 Formulation 60
4.2 Study of Simulated and Real Datasets 62
4.3 Algorithm for Optimising Two Smoothing Parameters 67
4.3.1 Algorithm 67
4.3.2 Augmented Data Formulation for Optimising λ_{rel} 69
4.4 Improving Computational Speed 70
4.5 Simulation Study 72
4.5.1 Data Simulation 73
4.5.2 Results 74
4.6 Summary 79

5 Ballooning 84
5.1 Basis Functions Simulation Study 85
5.1.1 Simulation Study 1 - A dataset prone to ballooning 86
5.1.2 Simulation Study 2 - A dataset with a more evenly spread design . 87
5.2 A Measure for Detecting Ballooning 89
5.2.1 Ballooning Simulation Study 91
5.3 Summary of Ballooning Properties 93
5.4 Conservation of Plume Mass Penalty 94
5.4.1 Penalty Formulation 94
5.4.2 Two-Dimensional Simulation Study - a toy example 97
5.4.3 Three-Dimensional Simulation Study 101
5.4.4 Summary and Issues 105

6 Optimal Sampling Design of Monitoring Networks with Spatial and Spatio-temporal Models 107
6.1 Review of Current Literature 108
6.1.1 Generic design optimalities extended to geostatistics 114
6.2 Two alternative objective functions 117

6.3 Variance of the Plume Mass (VM) Objective Function 119
6.3.1 VM Objective Function - Spatial Models 119
6.3.2 VM Objective Function - Spatio-temporal Models 123
6.4 Integrated Prediction Variance (IV) Objective Function 126
6.4.1 IV Objective Function - Spatial Models 127
6.4.2 IV Objective Function - Spatio-temporal Models 129
6.5 Simulation Studies 132
6.5.1 Data Simulation 133
6.5.2 Objective Functions in Action 134
6.5.3 The Effect of the Previous Sampling Frequency 140
6.5.4 Increasing the Time to the Next Sampling Event 143
6.6 Using the Objective Functions to Change the Well Network 147
6.6.1 Adding a New Well to the Network 147
6.6.2 Removing a Well from the Network 150
6.7 Summary 153

7 Sampling Design for Data with Multiplicative Errors 155
7.1 The Lognormal Distribution 156
7.2 Predictive Distribution of y 158
7.3 IV Objective Function using the Lognormal Variance 160
7.4 Results 161
7.4.1 Design for a Spatial P-splines Model 163
7.4.2 Design for a Spatio-temporal P-splines Model 166
7.4.2.1 The Effect of Previous Sampling Frequency 169
7.5 Summary 171

8 Discussion & Future Work 173
8.1 Spatio-temporal P-spline Models 173
8.2 Monitoring Network Design 176
8.2.1 Accounting for Multiplicative Errors 178

A Statistical Theorems and Properties 180
A.1 Conditional Distribution Property of a Multivariate Gaussian Distribution 180
A.2 Variance of a linear combination of variables 180

B Matrix Properties 182
B.1 The Spectral decomposition 182
B.2 Kronecker product properties 182

C Predicted Surfaces for the Basis Function Simulation Study in Chapter 5 184

D Analysis of a Second Design using the Objective Functions from Chapter 6 189
D.1 Objective Functions in Action 189
D.1.1 Designs for Spatial Models 189

D.1.2 Designs for a Spatio-temporal Model 193
D.1.3 The Effect of Previous Sampling Frequency 195
D.2 Using the Objective Functions to Change the Well Network 196
D.2.1 Adding a New Well to the Network 196
D.2.2 Removing wells from the network 199

Bibliography **202**

List of Figures

2.1 Some commonly used kernel functions . . . 9

2.2 A B-spline basis function of degree 3 made up of 4 polynomial pieces of degree 3. . . . 12

2.3 Unscaled B-spline basis functions of degree 1, 2 and 3 respectively. . . . 14

2.4 Unscaled tensor product B-spline basis functions of degree 3 . . . 16

2.5 Predicted curves using a splines model without (left) and with (right) a penalty term . . . 20

3.1 True underlying PDE described in Equation 3.1 (PDE1) at times $t \in \{0, 0.25, 0.50, 0.75\}$ . . . 39

3.2 Boxplots of the MSPE for each modelling method at each prediction time under the realistic sampling scenario (200 simulations). The number after each spatio-temporal splines model indicates the number of basis functions for the easting component . . . 42

3.3 Predicted surfaces, of one simulation, for each method at time 100 using only samples prior to this time point under the realistic sampling scenario. The plot in the top left location is the true surface from PDE1. The number after each spatio-temporal splines model indicates the number of basis functions for the easting component. . . . 43

3.4 Predicted surfaces, of one simulation, for each method at time 100 using the entire sampling data under the realistic sampling scenario. The plot in the top left location is the true surface from PDE1. The number after each spatio-temporal splines model indicates the number of basis functions for the easting component. . . . 44

3.5 Predicted surfaces, of one simulation, for each method at time 167 (most recent time point) using the entire sampling data under the realistic sampling scenario. The plot in the top left location is the true surface from PDE1. The number after each spatio-temporal splines model indicates the number of basis functions for the easting component. . . . 45

3.6 Boxplots of the MSPE for each modelling method at each prediction time under the full sampling scenario (200 simulations). The number after each spatio-temporal splines model indicates the number of basis functions for the easting component . . . 47

3.7 Predicted surfaces, of one simulation, for each method at time 100 using only samples prior to this time point under the full sampling scenario. The plot in the top left location is the true surface from PDE1. The number after each spatio-temporal splines model indicates the number of basis functions for the easting component. . . . 48

3.8 Predicted surfaces, of one simulation, for each method at time 100 using the entire sampling data under the realistic sampling scenario. The plot in the top left location is the true surface from the PDE1. The number after each spatio-temporal splines model indicates the number of basis functions for the easting component. 49

3.9 Predicted surfaces, of one simulation, for each method at time 167 (most recent time point) using the entire sampling data under the realistic sampling scenario. The plot in the top left location is the true surface from the PDE1. The number after each spatio-temporal splines model indicates the number of basis functions for the easting component. 50

3.10 Mean total MSPEs over all sampling times for 200 simulations using the 'observation removal' method of data removal. Error bars indicate ±1 standard deviation. 53

3.11 Mean total MSPEs over all sampling times for 200 simulations using the 'well removal' method of data removal. Error bars indicate ±1 standard deviation. 54

3.12 Predicted surfaces at four time points (300, 703, 899 & 1300 days) for each spline model on the real dataset. Filled squares indicate wells that were sampled at the prediction time and their recorded concentration; filled red triangles indicate observations that fall within the 21 day time window to be used in the spatial model and grey circles indicate the last recorded concentration at wells which were not sampled at the prediction time. 56

4.1 Two smoothing parameter grid search optimisation times (in seconds) for spatio-temporal models with increasing numbers of basis functions for a dataset with 4843 observations. For this plot the number of candidate values of λ_{rel} is 30. 61

4.2 Three simulated datasets made up of varying spatial normal densities oscillating at different temporal frequencies. The parameters used to simulate these datasets are shown in Table 4.1 63

4.3 Log posterior surface of the two smoothing parameters for the first simulated dataset using 15 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting arrows show the evolving maximum log posterior value as the number of basis functions in each direction increases from 6 - 9 - 12 - 15. 63

4.4 Log posterior surface of the two smoothing parameters for the second simulated dataset using 15 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting arrows show the evolving maximum log posterior value as the number of basis functions in each direction increases from 6 - 9 - 12 - 15. 64

4.5 Log posterior surface of the two smoothing parameters for the third simulated dataset using 15 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting line show the evolving maximum log posterior value as the number of basis functions in each direction increases from 6 - 9 - 12 - 15. 64

4.6 Log posterior surface of the two smoothing parameters for a dataset simulated from PDE1 with a realistic design, using 24 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting arrows show the evolving maximum log posterior value as the number of basis functions in each direction increases from 9 - 12 - 15 - 18 - 21 - 24. . 65

4.7 Log posterior surface of the two smoothing parameters, for a dataset simulated from PDE1 with a full design, using 24 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting arrows show the evolving maximum log posterior value as the number of basis functions in each direction increases from 9 - 12 - 15 - 18 - 21 - 24. . 65

4.8 Log posterior surface of the two smoothing parameters for the first real dataset using 15 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting arrows show the evolving maximum log posterior value as the number of basis functions in each direction increases from 6 - 9 - 12 - 15. 66

4.9 Log posterior surface of the two smoothing parameters for the second real dataset using 12 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting arrows show the evolving maximum log posterior value as the number of basis functions in each direction increases from 6 - 9 - 12. 66

4.10 True underlying PDE described in Equation 4.15 (PDE2) at times $t \in \{0.25, 0.50, 0.75, 1\}$ 74

4.11 Log posterior surface of the two smoothing parameters for PDE1 with a realistic design over a dense 80 × 30 grid. The point indicates the location of the maximum i.e. the optimal combination, whilst the cross indicates the optimal combination obtained using the algorithm. 76

4.12 Log posterior surface of the two smoothing parameters for PDE1 with a full design over a dense 80 × 30 grid. The point indicates the location of the maximum i.e. the optimal combination, whilst the cross indicates the optimal combination obtained using the algorithm. 76

4.13 Log posterior surface of the two smoothing parameters for PDE2 with a realistic design over a dense 80 × 30 grid. The point indicates the location of the maximum i.e. the optimal combination, whilst the cross indicates the optimal combination obtained using the algorithm. 77

4.14 Log posterior surface of the two smoothing parameters for PDE2 with a full design over a dense 80 × 30 grid. The point indicates the location of the maximum i.e. the optimal combination, whilst the cross indicates the optimal combination obtained using the algorithm. 77

4.15 Comparison of optimisation times (in seconds) for tuning two smoothing parameters using a grid search and the proposed algorithm with increasing numbers of basis functions on a dataset containing ∼ 4800 observations. . 79

4.16 Predicted surfaces for spatio-temporal p-spline models built using the combination of λ and λ_{rel} chosen by a grid search and the proposed algorithm for PDE1 and a realistic design. The plot at the top is the true surface from PDE1. 80

4.17 Predicted surfaces for spatio-temporal p-spline models built using the combination of λ and λ_{rel} chosen by a grid search and the proposed algorithm for PDE1 and a full design. The plot at the top is the true surface from PDE1. Ballooning can be observed in these predictions, this will be investigated in Chapter 5. 81

4.18 Predicted surfaces for spatio-temporal p-spline models built using the combination of λ and λ_{rel} chosen by a grid search and the proposed algorithm for PDE2 and a realistic design. The plot at the top is the true surface from PDE2. 82

4.19 Predicted surfaces for spatio-temporal p-spline models built using the combination of λ and λ_{rel} chosen by a grid search and the proposed algorithm for PDE2 and a full design. The plot at the top is the true surface from PDE2. 83

5.1 Prediction from a one-dimensional simulation where ballooning is evident, along with 95% confidence bands 84

5.2 Predicted surface using a spatio-temporal p-splines model with ballooning evident. Wells are plotted as filled black squares. 90

5.3 Balloons highlighted by the detection mechanism, pixels outlined in pink circles indicate unusually low predictions whilst black circles indicate unusually high predictions. Wells are plotted as filled black squares. 90

5.4 Counts of the number of times ballooning was detected for each combination of basis functions in the spatio-temporal p-splines model, with one smoothing parameter, for each dataset from PDE1 91

5.5 Counts of the number of times ballooning was detected for each combination of basis functions in the spatio-temporal p-splines model, with two smoothing parameters, for each dataset from PDE1 93

5.6 Basis functions of degree 3. Functions plotted in the same colour have the same integral. 96

5.7 True simulated data for the two dimensional study. The data consists of a normal density function over space with increasing standard deviation over time. Points indicate locations where observations were recorded. . . 98

5.8 Predicted surfaces using spline models with the difference penalty (left) and the conservation of mass penalty (middle). Five values of the penalty parameters $\lambda = \{1 \times 10^{-6}, 1 \times 10^{-3}, 1 \times 10^{0}, 1 \times 10^{3}, 1 \times 10^{6}\}$ (rows) are presented along with the true underlying surface (right). 100

5.9 Predicted plume mass at each time for each of the five considered penalised regression spline models with the conservation of mass penalty. . . 101

5.10 True simulated data for the three dimensional study. The data consist of a normal density function over space with a downward spiral motion over time. Points indicate locations where observations were recorded. 102

5.11 Predicted surfaces at each time point for penalised spatio-temporal spline models built with $\lambda_1 \in \left\{0, 1 \times 10^{-6}, 1 \times 10^{-4}, 1 \times 10^{-2}, 10, 1000\right\}$ and $\lambda_2 = 1 \times 10^{-4}$ using the conservation of mass penalty and a spatial difference penalty. Columns represent prediction times and rows correspond to each values of λ_1 for the mass penalty. 104

5.12 Predicted plume mass at each time for each of the five considered spline models with the conservation of mass penalty. 105

6.1 True surface simulated from the PDE at times $t \in \{0.1, 0.4, 0.7, 1\}$. The wells, from Design 1, contained within the white circle are the cluster that are discussed in Section 6.5.3. 133

6.2 Optimal designs for the integrated variance (IV) and variance of the mass (VM) objective functions using a spatial p-splines model when optimising for three proportions of the total number of wells (25%, 50% and 75%) using Design 1. 135

6.3 Optimal designs for the integrated variance (IV) and variance of the mass (VM) objective functions using a spatial Kriging model when optimising for three proportions of the total number of wells (25%, 50% and 75%) using Design 1. 136

6.4 Optimal designs for the integrated variance (IV) and variance of the mass (VM) objective functions using a spatio-temporal p-splines model when optimising for three proportions of the total number of wells (25%, 50% and 75%) using Design 1. 139

6.5 Optimal design of seven wells using the dataset containing seven randomly selected wells in the 'low frequency well' set i.e. first dataset. 141

6.6 Optimal design of seven wells using the dataset containing both of the clustered wells in the 'low frequency well' set i.e. second dataset. 142

6.7 Optimal design of seven wells using the dataset containing one of the clustered wells in the 'low frequency well' set i.e. third dataset. 143

6.8 Temporal B-spline basis functions with two observed time points. Dashed red lines indicate 'active' basis functions (i.e. those positioned over observations). 146

6.9 Change in the VM objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatial p-splines model for Design 1. 148

6.10 Change in the IV objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatial p-splines model for Design 1. 148

6.11 Change in the VM objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatio-temporal p-splines model for Design 1. 149

6.12 Change in the IV objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatio-temporal p-splines model for Design 1. 149

6.13 Difference in predicted variance if each filled monitoring well is removed at the next time point compared with the prediction variance if all wells were sampled for Design 1 and a spatial p-splines model. The facets are ordered in increasing difference in prediction variance. 151

6.14 Difference in predicted variance if each filled monitoring well is removed at the next time point compared with the prediction variance if all wells were sampled for Design 1 and a spatio-temporal p-splines model. The facets are ordered in increasing difference in prediction variance. 152

7.1 Probability density function of the lognormal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$ 157

7.2 Bivariate normal densities overlayed with bivariate student-t densities with increasing degrees of freedom $\nu = 5, 10, 20, 30$. 159

7.3 True surface simulated from the PDE, detailed in Section 4.5.1, at times $t \in \{0.1, 0.4, 0.7, 1\}$ and the well network used for Dataset 2. 162

7.4 Optimal designs for the integrated variance (IV) objective function, computed using the lognormal distribution (ILNV objective function) and the methodology set out in Chapter 6 (IV objective function), when optimising for 25%, 50% and 75% of the wells using a spatial p-splines model on Dataset 1. The surface beneath the wells shows the prediction from the previous sampling time. 163

7.5 Optimal designs for the integrated variance (IV) objective function, computed using the lognormal distribution (ILNV objective function) and the methodology set out in Chapter 6 (IV objective function), when optimising for 25%, 50% and 75% of the wells using a spatial p-splines model on Dataset 2. The surface beneath the wells shows the prediction from the previous sampling time. 166

7.6 Optimal designs for the integrated variance (IV) objective function, computed using the lognormal distribution (ILNV objective function) and the methodology set out in Chapter 6 (IV objective function), when optimising for 25%, 50% and 75% of the wells using a spatio-temporal p-splines model on Dataset 1. The surface beneath the wells shows the prediction from the previous sampling time. 167

7.7 Optimal designs for the integrated variance (IV) objective function, computed using the lognormal distribution (ILNV objective function) and the methodology set out in Chapter 6 (IV objective function), when optimising for 25%, 50% and 75% of the wells using a spatio-temporal p-splines model on Dataset 2. The surface beneath the wells shows the prediction from the previous sampling time. 170

7.8 Optimal design for the integrated variance (IV) objective function computed using the log-normal distribution and a spatial p-spline model when optimising for 7 wells using the dataset with a ‘low frequency well’ set. . . 171

8.1 Rectangle domain currently used for design optimisation and the proposed convex hull domain 177

8.2 Flow chart for deciding which objective function to use based on the assumed error type. 179

C.1 Predicted surfaces for spatio-temporal spline models with one smoothing parameter and an increasing number of basis functions for the 10th time point in PDE2 185

C.2 Predicted surfaces for spatio-temporal spline models with one smoothing parameter and an increasing number of basis functions for the 32nd time point in PDE2 186

C.3 Predicted surfaces for spatio-temporal spline models with two smoothing parameters and an increasing number of basis functions for the 10th time point in PDE2 187

C.4 Predicted surfaces for spatio-temporal spline models with two smoothing parameters and an increasing number of basis functions for the 32nd time point in PDE2 188

D.1 Optimal designs for the integrated variance (IV) and variance of the mass (VM) objective functions using a spatial p-spline model when optimising for three proportions of the total number of wells (25%, 50% and 75%) using Design 2. 190

D.2 Optimal designs for the integrated variance (IV) and variance of the mass (VM) objective functions using a spatial Kriging model when optimising for three proportions of the total number of wells (25%, 50% and 75%) using Design 2. 191

D.3 Optimal designs for the integrated variance (IV) and variance of the mass (VM) objective functions using a spatio-temporal p-spline model when optimising for three proportions of the total number of wells (25%, 50% and 75%) using Design 2. 194

D.4 Optimal design of seven wells using the low frequency dataset simulated on Design 2. 195

D.5 Change in the VM objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatial p-splines model for Design 2. 196

D.6 Change in the IV objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatial p-splines model for Design 2. 197

D.7 Change in the VM objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatio-temporal p-splines model for Design 2. 198

D.8 Change in the IV objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatio-temporal p-splines model for Design 2. 198

D.9 Difference in predicted variance if each filled monitoring well is removed at the next time point compared with the prediction variance if all wells were sampled for Design 2 and a spatial p-splines model. The facets are ordered in increasing difference in prediction variance. 200

D.10 Difference in predicted variance if each filled monitoring well is removed at the next time point compared with the prediction variance if all wells were sampled for Design 2 and a spatio-temporal p-splines model. The facets are ordered in increasing difference in prediction variance. 201

List of Tables

3.1 Mean Square Prediction Errors (MSPEs) and standard errors (SEs) at the 100th (using prior samples only), 100th (using all samples) and 167th time points under the realistic sampling scenario (200 simulations). Note: in the spline models the values in brackets indicate the basis functions for each direction and for Kriging in brackets is the covariance function used. 41

3.2 Mean Square Prediction Errors (MSPEs) and standard errors (SEs) at the 100th (using prior samples only), 100th (using all samples) and 167th time points under the full sampling scenario (200 simulations) Note: in the spline models the values in brackets indicate the basis functions for each direction and for Kriging in brackets is the covariance function used. 46

3.3 Mean computational time in seconds for each spatio-temporal model used in Section 3.2. The timings are separated by sampling design, where the first design contained 1400 samples and the second 4843 samples. The number of basis functions given are for each dimension i.e. (easting, northing, time). 51

3.4 Well-based cross validation scores for each prediction time for each spline model on the real dataset. 55

4.1 Combinations of standard deviations of the four normal densities plotted in space and the corresponding frequencies of their oscillations over time for the simulated data. 62

4.2 Optimal values of λ and λ_{rel} chosen by a classical grid search and the proposed algorithm along with MSPEs for predictions made with the resulting models for the PDE introduced in Chapter 3, described in Equation 3.1, (PDE1) and the PDE introduced in this chapter, described in Equation 4.15 (PDE2). 75

4.3 Values of the maximum log posterior values for the optimal combinations of λ and λ_{rel} for each PDE, design and optimisation method. 78

5.1 Mean Square Prediction Errors (MSPEs) and standard errors (SEs) at the 100th and 167th (last) time points of PDE1, Equation 3.1, under realistic and full sampling scenarios (200 simulations) for one smoothing parameter spatio-temporal p-spline models with differing numbers of basis functions (northing (N), easting (E), time (T)). 87

5.2 Mean Square Prediction Errors (MSPEs) and standard errors (SEs) at the 100th and 167th (last) time points of PDE1 under realistic and full sampling scenarios (200 simulations) for two smoothing parameter spatio-temporal p-spline models with differing numbers of basis functions (northing (N), easting (E), time (T)). 87

5.3 Mean Square Prediction Error (MSPE) and standard errors (SEs) at the 10th and 32nd (last) time points of PDE2 under realistic and full sampling scenarios (200 simulations) for the one smoothing parameter spatio-temporal p-spline models with differing numbers of basis functions (northing (N), easting (E), time (T)). 88

5.4 Mean Square Prediction Error (MSPE) and standard errors (SEs) at the 10th and 32nd (last) time points of PDE2 under realistic and full sampling scenarios (200 simulations) for the two smoothing parameter spatio-temporal p-spline models with differing numbers of basis functions (northing (N), easting (E), time (T)). 88

5.5 Mean proportion of ballooned pixels for those simulations which detected ballooning at the 100^{th} and 167^{th} time points of PDE1 under realistic and full sampling scenarios (200 simulations) for one smoothing parameter spatio-temporal p-spline models with differing numbers of basis functions (northing (N), easting (E), time (T)). 92

5.6 Mean proportion of ballooned pixels for those simulations which detected ballooning at the 100^{th} and 167^{th} time points of PDE1 under realistic and full sampling scenarios (200 simulations) for two smoothing parameter spatio-temporal p-spline models with differing numbers of basis functions (northing (N), easting (E), time (T)). 93

5.7 Mean Square Prediction Errors (MSPEs) for increasing values of $\lambda \in \{1 \times 10^{-6}, 1 \times 10^{-3}, 1 \times 10^{0}, 1 \times 10^{3}, 1 \times 10^{6}\}$ using the mass penalty and a difference penalty. 99

6.1 Cross-tabulation of the objective function values for each objective functions optimal design when optimising for 25%, 50% and 75% of the total number of wells using a spatial p-splines model using Design 1. Bold entries indicate the design with the minimum of each objective function. . 137

6.2 Cross-tabulation of the value of the objective functions for each objective functions optimal design when optimising for 25%, 50% and 75% of the total number of wells using a spatial Kriging model using Design 1. Bold entries indicate the design with the minimum of each objective function. . 137

6.3 Cross-tabulation of the value of the objective functions for each objective functions optimal design when optimising for 25%, 50% and 75% of the wells using a spatio-temporal p-splines model for Design 1. Bold entries indicate the design with the minimum of each objective function. 140

7.1 Cross-tabulation of the Integrated lognormal Variance (ILNV) and Integrated Variance (IV) objective functions and their corresponding optimal designs for Dataset 1. Optimisation was carried out for 25%, 50% and 75% of the total number of wells using a spatial p-splines model, with modelling being performed on the log scale. Bold entries indicate the design with the minimum of each objective function. 164

7.2 Cross-tabulation of the Integrated lognormal Variance (ILNV) and Integrated Variance (IV) objective functions and their corresponding optimal designs for Dataset 2. Optimisation is carried out for 25%, 50% and 75% of the total number of wells using a spatial p-splines model, with modelling being performed on the log scale. Bold entries indicate the design with the minimum of each objective function. 165

7.3 Cross-tabulation of the Integrated lognormal Variance (ILNV) and Integrated Variance (IV) objective functions and their corresponding optimal designs for Dataset 1. Optimisation is carried out for 25%, 50% and 75% of the total number of wells using a spatio-temporal p-splines model, with modelling being performed on the log scale. Bold entries indicate the design with the minimum of each objective function. 168

7.4 Cross-tabulation of the Integrated lognormal Variance (ILNV) and Integrated Variance (IV) objective functions and their corresponding optimal designs for Dataset 2. Optimisation is carried out for 25%, 50% and 75% of the total number of wells using a spatio-temporal p-splines model, with modelling being performed on the log scale. Bold entries indicate the design with the minimum of each objective function. 169

D.1 Cross-tabulation of the objective function values for each objective functions optimal design when optimising for 25%, 50% and 75% of the total number of wells using a spatial p-splines model using Design 2. Bold entries indicate the design with the minimum of each objective function. . 192

D.2 Cross-tabulation of the value of the objective functions for each objective functions optimal design when optimising for 25%, 50% and 75% of the total number of wells using a spatial Kriging model using Design 2. Bold entries indicate the design with the minimum of each objective function. . 192

D.3 Cross-tabulation of the value of the objective functions for each objective functions optimal design when optimising for 25%, 50% and 75% of the wells using a spatio-temporal splines model for Design 2. Bold entries indicate the design with the minimum of each objective function. 195

Chapter 1

Introduction

Spatio-temporal models have become a prevalent theme across many fields with their application being seen in economics (Holly et al. [2010]), epidemiology (Waller et al. [1997]), ecology (Malchow et al. [2007]) and environmental studies (Miller et al. [2014]). They are used to model data collected in space at several time points and are designed to describe the spatial and temporal correlations which are often present.

Spatio-temporal data arise most commonly in the environmental setting; for example, hourly measurements taken at meteorological stations (Environmental Protection Agency [2018]); or daily measurements of particle matter in the air (Air Quality in Scotland [2018]). The observed data can take various forms. Areal data are particularly common, where a study region is split into non-overlapping areal units and observations are made at each unit; for example, health boards in Scotland (National Health Service [2018]). This type of data is most frequently used in epidemiological modelling across a geographical region; see Shaddick and Zidek [2015]. In addition to areal data, geostatistical data are also very widely used. This is data which can be recorded at infinitely many locations across the study region. In reality these measurements are taken at predefined locations, for example groundwater monitoring wells around a refinery or air quality monitoring stations. Alternatively both types of data can be used collectively, see Lee et al. [2017].

Several goals can be achieved by making use of a spatio-temporal model. Typically they are used for prediction; however, they can also be used to track temporal changes and for design optimisation. Observed spatio-temporal trends can be explained by a multitude of both Bayesian and frequentist models; two of the most common methods are Kriging (Gaussian Processes), Krige [1951], and smoothing methods. See Wood [2006], Cressie and Wikle [2011] and Fahrmeir et al. [2013] for further details. These models are able to capture the non-linear trends often exhibited in both the spatial and temporal components of the data and allow for a more flexible representation compared with standard regression techniques. This thesis will focus primarily on spline-based models. Application of some of these smoothing methods can be seen in, Lee and Durban [2011], who use p-spline models in a mixed model framework to investigate ozone over Europe and similarly, Bowman et al. [2009], who model air pollution across Europe by utilising an extended additive splines model. O'Donnell et al. [2013] use spatio-temporal p-splines to model nitrate pollution in the river Tweed, while Ventrucci et al. [2014] use a spatial parametric function combined with a smooth temporal function for modelling neuronal activity in the brain.

Modelling spatio-temporal data can be computationally demanding. Very often the data are made up of observations at many locations across a large study region and at several time points. With the ability now to store large volumes of data, and to collect data automatically, the dense nature of the data can pose computational problems when estimating model parameters, such as the smoothing parameter/parameters in a penalised regression splines model. Almost all parameter estimation methods require the inversion of an appropriate design matrix and as more basis functions are incorporated to allow for more modelling flexibility with increased amounts of data, the dimensions of this matrix increase exponentially, resulting in computational times which also increase exponentially. Several authors have tackled the issue of poor computational speed, including Wood [2011], who proposed an efficient restricted maximum likelihood (REML) method for determining the smoothing parameters in an additive smoothing splines model. Molinari [2014] made use of eigenvalue decompositions and linear algebra formulations to tune the smoothing parameter efficiently in a p-splines model. Further details of this method will be given in Section 2.3.2.

In other cases, data are collected by hand, resulting in sparse datasets, usually with several missing values. This can pose problems of a different nature, with some sampling times only containing one or two observations.

Time and space should be treated and modelled differently as they are measured on different scales that have no association. However, deciding how to treat these two components is very often a subjective matter. Differing covariance structures can be adopted for each component if a Kriging-based model is used; see Cressie and Wikle [2011] for more details. In a spline paradigm, separate smoothing parameters can be used to penalise the smoothness across each dimension separately. Alternatively, Evers et al. [2015] suggest scaling the number of basis function for each dimension to reflect the level of variation believed to be present e.g. if it is assumed that there is less variation over time compared with space, then the temporal dimension is assigned a smaller number of basis functions compared with the spatial dimensions. Approaches for dealing with space and time data will be discussed further in Chapter 4.

1.1 Groundwater

Groundwater is a term used to describe all water stored beneath the surface of the earth in geological formations known as aquifers, located in the saturated zone below the water table (Bear [1979]). It is a vital and essential resource which is widely used to supply water for drinking, industry and agriculture and it makes up around 97% of all the available freshwater on earth, excluding glaciers and ice caps (Hornberger et al. [2014]).

In many countries groundwater is the main source of drinking water. With rapid population growth, the demand for this is continually increasing (Arnell [1999]), with currently around 75% of EU inhabitants depending on groundwater for their drinking water (European Commission [2018]). In addition to its use by the population, it also maintains wetlands and river flow during periods of drought and is a vital factor in the sustainability of their biodiversity and ecology (Scottish Environment Protection Agency [2018]).

1.1.1 Groundwater Pollution

Groundwater quality can be affected by pollutants from many sources. Therefore, several countries have legislation in place and protect groundwater from contaminants; for example, the European Union Water Framework Directive (European Union [2016]).

Groundwater moves slowly through the subsurface, resulting in pollutants remaining for long periods of time, sometimes even for decades.

Contamination of groundwater is difficult to avoid and thus, with such a high demand for the resource, it is of paramount importance that pollution within groundwater is monitored and controlled. Monitoring groundwater is not an easy operation given its 'hidden' nature. Locating contamination can be difficult and, once located, it is challenging to access and assess its implications. This highlights the need for a regular monitoring schedule to allow for early detection of a pollution incident. A more in-depth discussion of the different categories of groundwater monitoring along with a discussion of current methods, is presented at the beginning of Chapter 6.

Some of the more common sources of groundwater pollution are outlined below:

- **Natural Sources**

  Contamination can arise from substances found naturally in soil or rocks such as iron, manganese and arsenic. Excessive consumption of these chemicals can cause health problems; see Ng et al. [2003] and Appelo and Postma [2004]. Particles of decaying organic matter can also pollute the groundwater.

- **Septic Systems and Landfill sites**

  Effluent from septic tanks and sewage works are one of the main causes of groundwater contamination. Incorrectly installed waste water disposal systems in homes and businesses can leak bacteria and household chemicals into the groundwater (Groundwater Foundation [2018]).

  Chemicals can leach from landfill sites into the ground and subsequently the groundwater. New landfill sites are required to be lined with a synthetic or clay material to prevent these hazardous chemicals filtering into the groundwater; for the EU regulations, see European Union [1999]. However, if this layer is damaged the chemicals are able to enter the groundwater. Similarly, old landfill sites may still emit chemicals.

- **Agriculture**

  Groundwater can become contaminated as a result of many agricultural practices. Often pesticides or fertilizers sprayed on the crops can seep into the soil and eventually reach the groundwater. The most common pollutant is nitrate, a by product from nitrogen-rich fertilizers and animal waste (Singh and Sekhon [1979]). If concentrations of nitrate rise too high this can pose a risk to human health.

- **Releases and spills from petroleum products**

  Underground and above-ground storage tanks are often used for petroleum products; for example, oil or gas tanks for central heating or fuel tanks at petrol stations. As tanks age they can be subject to corrosion which in turn can result in leaks into the groundwater. Most of these petroleum based chemicals do not dissolve and disperse into the water but instead travel in a cloud formation. Benzene, toluene, ethylbenzene, and total xylenes (BTEX), which come from gasoline refining, and methyltert-butyl-ether (MTBE), which is a fuel additive, are common contaminants in urban areas. These contaminants are collectively called NAPLs (Non-aqueous phase liquids) (Mackay and Cherry [1989]).

Contamination by petroleum based chemicals will be the main focus of this thesis.

1.2 Thesis Overview

The primary aim of this thesis is to develop methodology for determining optimal sampling designs for groundwater quality monitoring. These optimal designs will look to reduce sample sizes and subsequently costs by utilising spatio-temporal models in preference to the more commonly used spatial models.

In preparation for this aim, Chapter 3 will assess the benefits of using a spatio-temporal model over a spatial model for prediction, through a comparative study of current modelling methods described in Chapter 2.

Chapter 4 will then adapt the spatio-temporal p-splines model by Molinari [2014] to allow for more flexibility in choosing the degree of smoothness in the spatial and temporal components. This will be done by incorporating an additional smoothing parameter for the temporal dimension. Adding this extra flexibility comes at a computational expense, thus an algorithm for determining the optimal combination of smoothing parameters is also presented.

Due to the nature of the monitoring well networks and erratic sampling frequencies, a phenomenon known as 'ballooning' can be seen in the predictions of the one and two smoothing parameter spatio-temporal p-spline models. A simulation study is conducted in Chapter 5 with the aim of determining which components of the model specification account for this. A measure for detecting when ballooning might be present is also introduced along with a 'Conservation of Mass' penalty which aims to penalise sudden changes in the contaminant plume mass over time.

Chapter 6 uses spatial Kriging and spatial and spatio-temporal p-spline models to optimise sampling designs based on minimising two objective functions, namely, the Variance of the Plume Mass (VM) and the Integrated Prediction Variance (IV). Properties of designs resulting from these objective functions are investigated.

Finally, Chapter 7 adapts the objective functions for data with multiplicative error since this is what is commonly appropriate for groundwater quality data. The resulting designs are compared to the designs from Chapter 6 to determine key differences.

Chapter 2

Current Nonparametric Modelling Techniques

Environmental data, such as measurements of contamination in groundwater, exhibit many complex features which make classical parametric methods such as linear regression inappropriate. Thus, environmental data are often analysed through nonparametric modelling methods. Utilising these approaches allows the assumption of linearity to be relaxed and more flexible smooth functions to be fitted instead.

Given a set of observed data containing response y_i and a single covariate x_i; $i \in \{1, \cdots, n\}$, a nonparametric model can be denoted as

$$y_i = m(x_i) + \epsilon_i, \tag{2.1}$$

where $\epsilon_i \sim N(0, \sigma^2)$ are the independent error terms and $m(x_i)$ is a nonparametric regression function of the covariate which can be estimated by some smooth function, $\hat{m}(x_i)$. In the following sections, a brief overview will be given of some of the smoothing techniques used to estimate $\hat{m}(x_i)$. A more in-depth description of splines is also included, as this is the primary method used throughout this thesis.

2.1 Kernel Smoothing Methods

2.1.1 Kernel Density Estimation

Kernel density estimation allows the detailed shape of the underlying density function of a set of data to be estimated. Given observed data, $\mathbf{x} = (x_1, x_2, \cdots, x_n)$, the density function can be estimated by:

$$\begin{aligned}\hat{m}(x) =& \frac{1}{n}\sum_{i=1}^{n} w(x - x_i; \lambda),\\ =& \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\lambda} K\left(\frac{x - x_i}{\lambda}\right),\end{aligned} \tag{2.2}$$

where $K()$ is known as a kernel function, with each observation having a kernel function centred around it. Kernels are symmetric, non-negative functions which assign weights to the neighbouring observations. As the distance to the neighbouring observation increases, the weight assigned decreases monotonically. The bandwidth or the smoothing parameter, λ, defines the size of the 'window' surrounding each observation. As λ increases the density becomes more smooth. Some common kernel functions are detailed below:

- Uniform/Rectangular
$$K(u) = \frac{1}{2}\mathbb{I}\{|u| \leq 1\}$$
- Triangular
$$K(u) = (1 - |u|)\mathbb{I}\{|u| \leq 1\}$$
- Epanechnikov
$$K(u) = \frac{3}{4}(1 - u^2)\mathbb{I}\{|u| \leq 1\}$$
- Tricube
$$K(u) = (1 - |u|^3)^3\mathbb{I}\{|u| \leq 1\}$$
- Gaussian
$$K(u) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}u^2\right)$$

where $\mathbb{I}$ is an indicator function, taking the value 1 when x_i falls within the window, of width 2λ, around the observation, and 0 otherwise. Figure 2.1 illustrates the shape of each of these kernels.

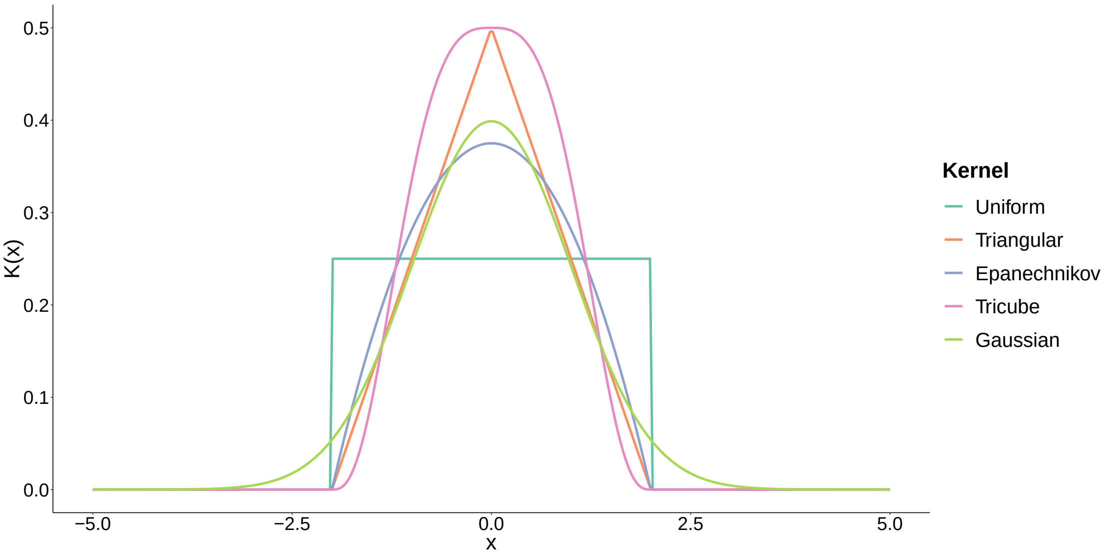

FIGURE 2.1: Some commonly used kernel functions

2.1.2 Local Linear Regression

Local linear regression utilises the weights produced by a kernel function to fit regression models locally to a set of data. Given a set of observed data containing responses y_i and a single covariate x_i; $i \in \{1, \cdots, n\}$, local linear regression solves a weighted least squares problem at each x_i:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_i - x)\}^2 w(x - x_i; \lambda). \tag{2.3}$$

The estimated value at x, $\hat{m}(x)$, is then taken as $\hat{\alpha}$. As in the case of kernel density estimation, λ acts as a smoothing parameter and controls the smoothness of the function. As λ increases, the width of the kernel increases and thus the smoother the estimated function becomes. The kernel also ensures that observations close to x_i are given the most weight in determining the estimate. For a more detailed discussion see Bowman and Azzalini [1997].

2.2 Regression Splines

Regression splines are another common nonparametric regression approach which allow the relationships between a response and covariates to be described in a flexible manner. Regression splines are constructed by joining a set of known functions at points called *knots*, often these functions are referred to as *basis functions*.

The estimate of the nonparametric function, $\hat{m}(x_i)$, is a function of the form $\sum_{j=1}^{m} \alpha_j B_j(x_i)$, B_j; $j \in \{1, \cdots, m\}$ are the basis functions constructed over knots, $\boldsymbol{\kappa} = (\kappa_1, \cdots, \kappa_q)$ and α_j are the corresponding basis coefficients. For a polynomial spline of degree p, the number of basis functions is $m = (p + q - 1)$. The model can be expressed in vector matrix form:

$$\mathbf{y} = \mathbf{B}(\mathbf{x})\boldsymbol{\alpha} + \mathbf{e}, \tag{2.4}$$

where $\mathbf{B} = \mathbf{B}(\mathbf{x})$ is an $(n \times m)$ basis matrix with each column corresponding to a basis function, $\boldsymbol{\alpha}$ is an $(m \times 1)$ vector of basis coefficients and $\mathbf{x}$ is the covariate vector. This matrix and vector are treated in the same way as the design matrix and vector of coefficients from a linear model, respectively. To obtain the estimates for the basis coefficients and hence the estimated fitted values, the least squares function (LS) shown in Equation 2.5 is minimised,

$$\mathrm{LS}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \left(y_i - \underbrace{\sum_{j=1}^{m} \alpha_j B_j(x_i)}_{\hat{m}(x_i)} \right)^2 . \tag{2.5}$$

Alternatively this can be written in vector-matrix notation as:

$$\mathrm{LS}(\boldsymbol{\alpha}) = (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha})^{\top} (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) . \tag{2.6}$$

Consequently, by minimising the expression in Equation 2.6 with respect to $\boldsymbol{\alpha}$, the vector of basis coefficient estimates, $\hat{\boldsymbol{\alpha}}$, is computed as:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{B}^\top \mathbf{B})^{-1}\mathbf{B}^\top \mathbf{y} \tag{2.7}$$

Several decisions have to be made when modelling with regression splines, one of which is choosing an appropriate number and location of the knots. The choice is difficult, but crucial, as the knots can dramatically change the shape of the function. The decision involves a bias-variance trade off. A large number of knots will produce a rougher model which tracks the data closely. This model will have a low bias, but the variance of this model is likely to be high. On the other hand, a model with a small number of knots will have a low variance but is likely to have a high bias. This model will be smoother than the model with a large number of knots.

There are also many options for the locations at which the knots are placed. The simplest and most routinely used method is to have equally spaced knots, but an alternative is to position the knots according to quantiles of the covariates or, 'by eye' i.e. subjectively.

To resolve these issues, a model selection criterion can be adopted and several models can be compared. A further approach which is also commonly used is to introduce a penalty term on the basis coefficients to control the smoothness of the function and prevent over-fitting. This penalty approach is known as *penalised regression splines* and is discussed in more detail in Section 2.2.3.

There are several different types of basis functions that can be used, including; polynomial splines, natural cubic splines, truncated power basis, thin-plate splines and B-splines. The later are described in Section 2.2.1. For further information on the other types of spline bases see Fahrmeir et al. [2013] and Wood [2006].

2.2.1 B-splines

B-splines are a set of basis functions which are commonly chosen as an alternative to the truncated power basis. They are preferred as they are more stable numerically; see Fahrmeir et al. [2013]. The main advantage of B-splines is that they are a local basis i.e. they are only non-zero over a small range of the data, making them computationally efficient. For a given B-spline basis function of degree p and set of knots $\boldsymbol{\kappa} = (\kappa_1, \cdots, \kappa_q)$, the following properties hold:

- A B-spline basis function of degree p is made up of p + *1* polynomial pieces of degree p. This is depicted in Figure 2.2, where a B-spline of degree 3 is seen to be made up of 4 polynomial pieces of degree 3.
- Each basis function is non-zero over a range of p + *2* adjacent knots.
- At any point within the range of the data, p + *1* basis functions are non-zero.
- Every basis function overlaps with *2p* other bases.
- A further *2p* knots are required outside of the domain $[a, b]$.
- For every point $x \in [a, b]$:

$$\sum_{j=1}^{m} B_j(x) = 1$$

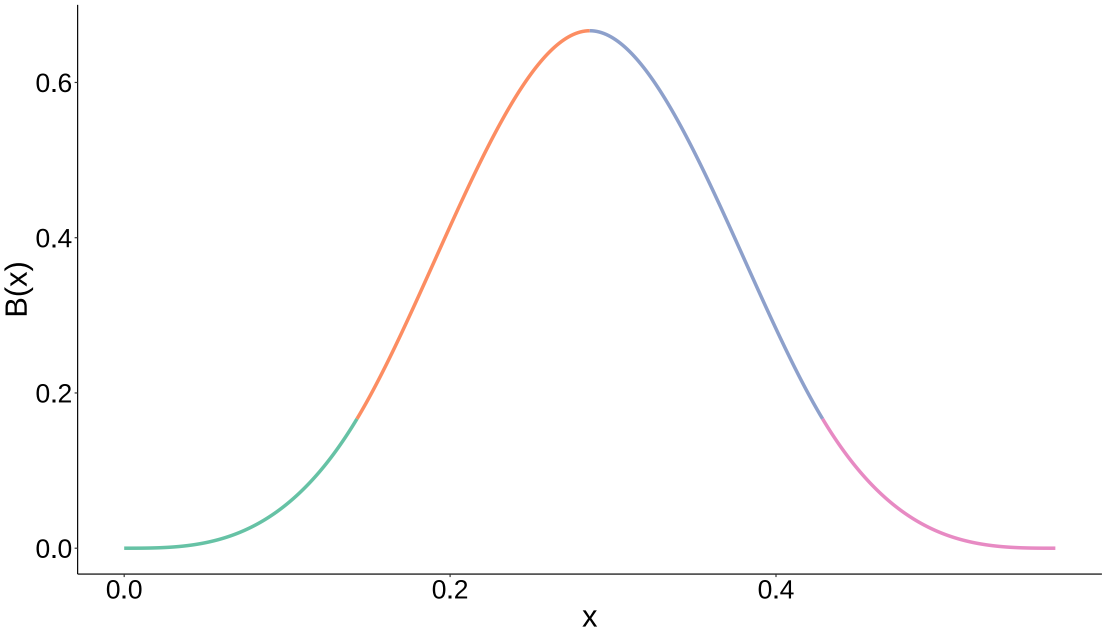

FIGURE 2.2: A B-spline basis function of degree 3 made up of 4 polynomial pieces of degree 3.

B-splines can be defined recursively as:

- Given a set of q knots the B-spline basis of degree 0 is given by the functions $\big(B_1^0(x), \cdots, B_{q-1}^0(x)\big)$ with

$$B_j^0(x) = \begin{cases} 1 & \kappa_j \leq x \leq \kappa_{j+1} \\ 0 & otherwise. \end{cases} \tag{2.8}$$

- Given a set of q knots the B-spline basis of degree $p > 0$ is given by the functions $\Big(B_1^p(x), \cdots, B_{p+q-1}^p(x)\Big)$ with

$$B_j^p(x) = \frac{x - \kappa_{j-p}}{\kappa_j - \kappa_{j-p}} B_{j-1}^{p-1}(x) + \frac{\kappa_{j+1} - x}{\kappa_{j+1} - \kappa_{j+1-p}} B_j^{p-1}(x). \tag{2.9}$$

Thus the resulting matrix $\mathbf{B} = \mathbf{B}(\mathbf{x})$ is:

$$\mathbf{B} = \begin{bmatrix} B_1^p(x_1) & \cdots & \cdots & B_m^p(x_1) \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ B_1^p(x_n) & \cdots & \cdots & B_m^p(x_n) \end{bmatrix}$$

From this matrix it is apparent that there is no intercept term for a B-splines model. However, the intercept is implicitly contained in the span of the basis; see Fahrmeir et al. [2013]. As B-splines are a local basis, the matrix $\mathbf{B}$ is made up of mainly 0's.

Figure 2.3 illustrates B-spline basis functions in one dimension, before they are scaled by the basis coefficients. As shown in this plot, the basis functions are all the same shape, they are simply shifted along the x-axis. The distance between each function depends on the distance at which the knots are placed. In Figure 2.3 equally spaced knots are used, and it is clear from the plot that as the number of knots increases the basis functions would become closer together.

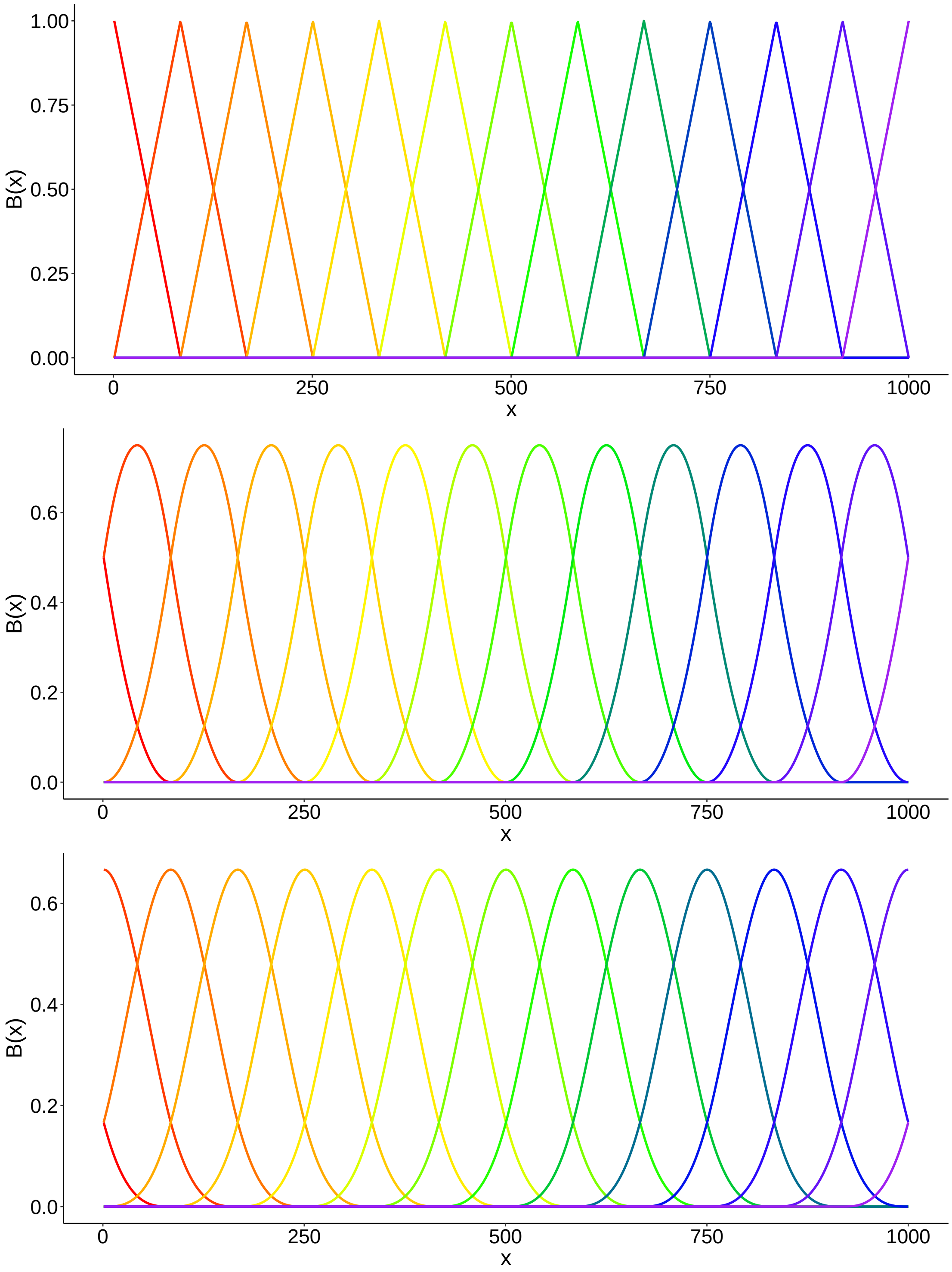

Figure 2.3: Unscaled B-spline basis functions of degree 1, 2 and 3 respectively.

Derivative of a B-spline

The construction of B-splines from polynomial pieces makes their derivatives simple to compute. It can be shown that the first derivative of a B-spline of degree p is:

$$\frac{\partial}{\partial x} B_j^p(x) = \frac{p}{\kappa_j - \kappa_{j-p}} B_{j-1}^{p-1}(x) + \frac{p}{\kappa_{j+1} - \kappa_{j+1-p}} B_j^{p-1}(x). \tag{2.10}$$

When the case of equally-spaced knots is considered Equation 2.10 can be simplified to:

$$\frac{\partial}{\partial x} B_j^p(x) = \frac{1}{\delta} B_{j-1}^{p-1}(x) + \frac{1}{\delta} B_j^{p-1}(x). \tag{2.11}$$

Where $\delta = \kappa_j - \kappa_{j-1}$. The derivative of the non-parametric regression function, $f(x)$, is then given as:

$$\frac{\partial}{\partial x} m(x) = \sum_{j=1}^{m-1} B_j^{p-1}(x) \frac{\alpha_{j+1} - \alpha_j}{\delta}. \tag{2.12}$$

In vector-matrix form Equation 2.12 becomes:

$$\frac{\partial}{\partial x} m(x) = \frac{1}{\delta} \mathbf{B}^{p-1} \mathbf{D}_1 \boldsymbol{\alpha}, \tag{2.13}$$

where $\mathbf{B}^{p-1}$ is a matrix of basis functions of degree $p-1$ and $\mathbf{D}_1$ is a 1^{st} order difference matrix. Similarly, the r^{th} derivative can be defined as:

$$\frac{\partial^r}{\partial x^r} m(x) = \frac{1}{\delta^r} \mathbf{B}^{p-r} \mathbf{D}_r \boldsymbol{\alpha}. \tag{2.14}$$

2.2.2 Tensor Product Regression Splines for Multi-Dimensional Data

Generalising regression splines to data of a higher dimension, i.e. models with more than one covariate, is relatively straightforward. To extend the approach to 2-dimensions, with data indexed over spatial coordinates (x_{1i}, x_{2i}), $m(x_{1i}, x_{2i})$ can then be estimated as:

$$\hat{m}(x_{1i}, x_{2i}) = \sum_{j}\sum_{k} \alpha_{jk} B_{jk}(x_{1i}, x_{2i}) = \sum_{j}\sum_{k} \alpha_{jk} B_j(x_{1i}) B_k(x_{2i}) \tag{2.15}$$

where $m(x_{1i}, x_{2i})$ is a non-parametric regression function of the spatial coordinates x_{1i} and x_{2i}, α_{jk} are the basis coefficients, and the basis functions, B_j and B_k, are for the easting and northing components respectively. The basis functions can be constructed efficiently using row-wise Kronecker products of the marginal B-spline bases; Lee and Durban [2011]. Figure 2.4 illustrates the construction of the unscaled tensor product B-splines.

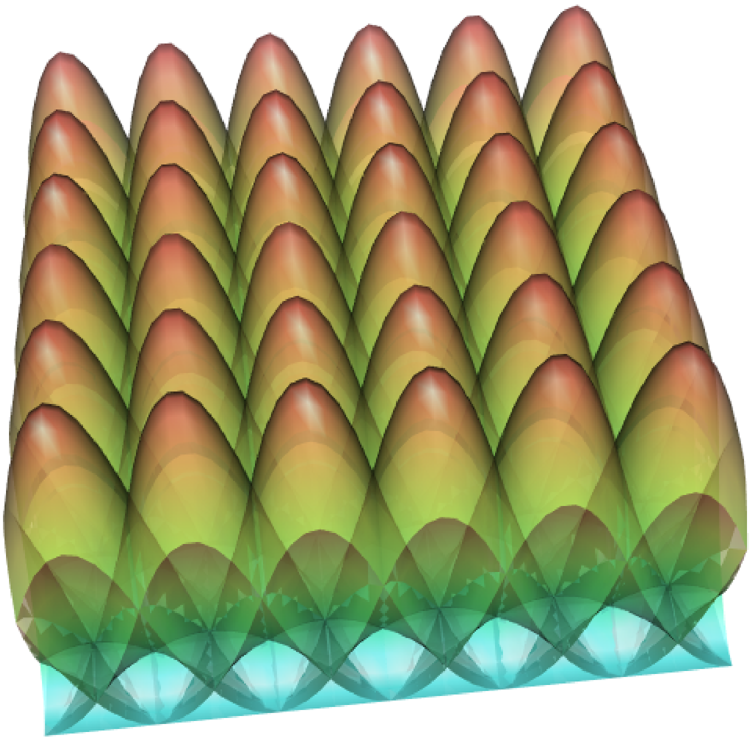

FIGURE 2.4: Unscaled tensor product B-spline basis functions of degree 3

The matrix of basis functions **B**, is of dimension $n \times (m_1, m_2)$, where m_1 and m_2 are the numbers of basis functions for each component.

To extend this approach to a 3^{rd} dimension an additional basis function, B_l, is added to Equation 2.15 for the temporal component, as detailed in Equation 2.16. Again, the basis functions are constructed using row-wise Kronecker products of the marginal B-spline bases:

$$\begin{aligned}\hat{m}(x_{1i}, x_{2i}, t_i) &= \sum_j \sum_k \sum_l \alpha_{jkl} B_{jkl}(x_{1i}, x_{2i}, t_i),\\ &= \sum_j \sum_k \sum_l \alpha_{jkl} B_j(x_{1i}) B_k(x_{2i}) B_l(t_i).\end{aligned} \tag{2.16}$$

Similarly, the matrix of basis functions, $\mathbf{B}$, is of dimension $n \times (m_1, m_2, m_3)$, where again m_1, m_2 and m_3 are the numbers of basis functions for each component. For simplicity, each dimension can be given an equal number of basis functions, but, we will see later that it can be useful for each dimension to have a different number of basis functions.

2.2.3 Penalised Regression Splines

To overcome the issue of choosing an appropriate number of knots a penalty term which prevents over-fitting can be introduced. The main advantage of this penalised approach is that the smoothness of the function is no longer dependent on the number of knots, but rather on a single smoothing parameter. The method involves minimising the penalised least squares criterion (PLS), shown in Equation 2.17 in the same way as the least square criterion is minimised in Equation 2.5. To fit a spline model with a penalty term a reasonably large number of equidistant knots is commonly chosen (∼20 - 40) and Equation 2.17 is minimised with respect to $\boldsymbol{\alpha}$, the basis coefficients.

$$\text{PLS}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \alpha_j B_j(x_i) \right)^2 + \lambda \mathbf{PEN} \tag{2.17}$$

where λ, the *smoothing parameter*, is a non-negative value which penalises the overall smoothness of the function. When $\lambda = 0$ there is no penalty attached and the function has the ability to interpolate the data. As $\lambda \to \infty$ the function becomes increasingly smooth.

Determining the most appropriate smoothing parameter is imperative. Section 2.2.7 outlines several criteria that can be used to determine the optimal smoothing parameter.

A common choice of penalty, shown in Equation 2.18, is the integral of the squared second derivative of the non-parametric function. The second derivative is deemed a suitable choice of penalty as it is a measure of curvature.

The penalty term can be denoted as

$$\begin{aligned}\lambda \int_x m''(x)^2 dx &= \lambda \int_x \left(\sum_{j=1}^{m} \alpha_j B_j''(x) \right)^2 dx \\ &= \lambda \int_x \left(\sum_{j=1}^{m} \sum_{k=1}^{m} \alpha_j \alpha_k B_j''(x) B_k''(x) \right) dx \\ &= \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha},\end{aligned} \tag{2.18}$$

where $\mathbf{K}[j,k] = \int_x B_j''(x) B_k''(x) dx$. The quadratic form nature of the penalty allows for easier computation. The PLS can then be written in vector-matrix notation as shown in Equation 2.19.

$$\text{PLS}(\boldsymbol{\alpha}) = ||\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}||^2 + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}. \tag{2.19}$$

The expression for $\boldsymbol{\alpha}$ which minimises the PLS criterion is:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{K})^{-1} \mathbf{B}^\top \mathbf{y}. \tag{2.20}$$

The covariance matrix of the PLS estimates, conditional on the observed data, is formulated as:

$$\mathbf{C}_{\hat{\alpha}|y} = \text{cov}(\hat{\boldsymbol{\alpha}}) = \sigma^2 (\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{K})^{-1} \mathbf{B}^\top \mathbf{B} (\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{K})^{-1}. \tag{2.21}$$

From Equation 2.21 it is apparent that the PLS estimates are not unbiased. However, in comparison with the un-penalised LS estimates, the PLS estimates have lower variance, Fahrmeir et al. [2013].

Finally, the fitted values can be obtained:

$$\hat{\mathbf{y}} = \mathbf{B}\hat{\boldsymbol{\alpha}} = \mathbf{B}(\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{K})^{-1}\mathbf{B}^\top\mathbf{y} = \mathbf{S}\mathbf{y}. \tag{2.22}$$

Here $\mathbf{S} = \mathbf{B}(\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{K})^{-1}\mathbf{B}^\top$ is known as the *smoothing matrix* (see Fahrmeir et al. [2013]) whose trace, $\text{tr}(\mathbf{S})$, is defined as the *effective degrees of freedom (e.d.f)* and can be interpreted as the equivalent number of parameters in the model, giving an estimation of the model's complexity. An estimate of the variance of the fitted values is then obtained through Equation 2.23. This formulation exploits the fact that $\mathbf{S}$ is a symmetric matrix.

$$\text{var}(\hat{\mathbf{y}}) = \text{var}(\mathbf{S}\mathbf{y}) = \mathbf{S}\sigma^2\mathbf{I}_n\mathbf{S}^\top = \sigma^2\mathbf{S}\mathbf{S}^\top. \tag{2.23}$$

2.2.4 P-splines

An alternative approach is to use p-splines, proposed by Eilers and Marx [1996], which adds a penalisation term based on order differences between adjacent coefficients in the bases of the B-splines. For p-splines with order one differences, the penalty term is

$$\lambda||\mathbf{D}_1\boldsymbol{\alpha}||^2 = \lambda\sum_{j=1}^{q+p-2}(\alpha_{j+1} - \alpha_j)^2. \tag{2.24}$$

Thus, the penalty in vector-matrix form can be denoted as:

$$\lambda\boldsymbol{\alpha}^\top\underbrace{\mathbf{D}_d^\top\mathbf{D}_d}_{\mathbf{K}}\boldsymbol{\alpha}, \tag{2.25}$$

where $\mathbf{D}_d$ is a d^{th} order difference matrix. Note, by denoting $\mathbf{K} = \mathbf{D}_d^\top\mathbf{D}_d$, the penalty term is of the same form as that given in Equation 2.19 and hence the same equations can be used to obtain the basis coefficient estimates and the fitted values.

Commonly 1^{st} or 2^{nd} order differences are used; for example a second order penalty is computed as:

$$\mathbf{D}_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & \vdots \\ 0 & 1 & -2 & 1 & 0 & \ddots & \vdots \\ 0 & 0 & 1 & -2 & 1 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

The difference penalty is used as it is a good discrete approximation to the integrated square of the d^{th} derivative, see Eilers and Marx [1996].

Figure 2.5 demonstrates the difference penalty in action. The panel on the left shows the result of fitting a regression spline with a B-spline basis. Here the curve is clearly over-fitted. In contrast, the panel on the right shows the same basis but for a penalised regression spline fit with a first order difference penalty. Adding the penalty with a suitable penalisation parameter forces the curve to be smooth.

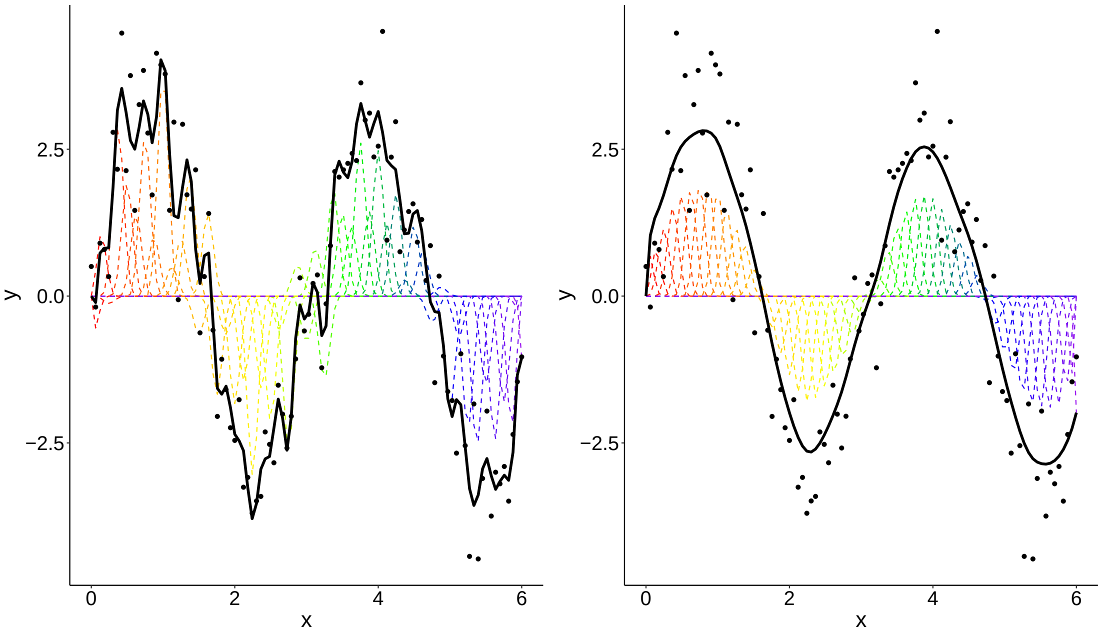

FIGURE 2.5: Predicted curves using a splines model without (left) and with (right) a penalty term

2.2.5 Tensor Product Penalised Regression Splines

The penalties described above can be easily adapted for multidimensional data, building on their spline bases detailed in Section 2.2.2. Separate smoothing parameters can be used for each dimension; alternatively one global smoothing parameter can also be used.

Given data which are indexed over two spatial dimensions, i.e. (x_{1i}, x_{2i}), from Equation 2.15 the non-parametric function $m(x_{1i}, x_{2i})$ can be estimated as:

$$\hat{m}(x_{1i}, x_{2i}) = \sum_j \sum_k \alpha_{jk} B_j(x_{1i}) B_k(x_{2i}).$$

The penalised least squares criterion, defined in Equation 2.17 for one dimension, can be written for two dimensions as:

$$\mathrm{PLS}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \left(y_i - \sum_j \sum_k \alpha_{jk} B_j(x_{1i}) B_k(x_{2i}) \right)^2 + \lambda \mathbf{PEN}. \tag{2.26}$$

As shown by Wood [2006], the corresponding integrated squared second derivative penalty can be computed as:

$$\begin{aligned}
\lambda \mathbf{PEN} &= \lambda \int_{x_1} \int_{x_2} \left[\left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 \right] dx_1 dx_2 \\
&= \lambda \int_{x_1} \int_{x_2} \left(\sum_j \sum_k \alpha_{jk} B_j''(x_1) B_k(x_2) \right)^2 \\
&\qquad + \left(\sum_j \sum_k \alpha_{jk} B_j(x_1) B_k''(x_2) \right)^2 dx_1 dx_2 \\
&= \lambda \sum_{jk} \sum_{lm} \alpha_{jk} \alpha_{lm} \left(\int_{x_1} B_j''(x_1) B_l''(x_1) dx_1 \int_{x_2} B_k(x_2) B_m(x_2) dx_2 \right. \\
&\qquad \left. + \int_{x_1} B_j(x_1) B_l(x_1) dx_1 \int_{x_2} B_k''(x_2) B_m''(x_2) dx_2 \right) \\
&= \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}
\end{aligned}$$

where

$$\mathbf{K} = (\widetilde{\mathbf{A}}_1 \otimes \mathbf{A}_2) + (\mathbf{A}_1 \otimes \widetilde{\mathbf{A}}_2)$$

and

$$\mathbf{A}_1[j,l] = \int_{x_1} B_j(x_1)B_l(x_1)dx_1, \qquad \mathbf{A}_2[k,m] = \int_{x_2} B_k(x_2)B_m(x_2)dx_2, \tag{2.27}$$

$$\widetilde{\mathbf{A}}_1[j,l] = \int_{x_1} B_j''(x_1)B_l''(x_1)dx_1, \qquad \widetilde{\mathbf{A}}_2[k,m] = \int_{x_2} B_k''(x_2)B_m''(x_2)dx_2. \tag{2.28}$$

The penalty term for data indexed over three variables, i.e. space and time, can be denoted similarly, with matrix $\mathbf{K}$ taking the form

$$\mathbf{K} = (\widetilde{\mathbf{A}}_1 \otimes \mathbf{A}_2 \otimes \mathbf{A}_3) + (\mathbf{A}_1 \otimes \widetilde{\mathbf{A}}_2 \otimes \mathbf{A}_3) + (\mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \widetilde{\mathbf{A}}_3),$$

where the matrices $\mathbf{A}_1$, $\widetilde{\mathbf{A}}_1$, $\mathbf{A}_2$ and $\widetilde{\mathbf{A}}_2$ have the same entries as in Equations 2.27 and 2.28, and the matrices $\mathbf{A}_3$ and $\widetilde{\mathbf{A}}_3$, i.e. those for the temporal dimension, take the form

$$\mathbf{A}_3[l,o] = \int_t B_l(t)B_o(t)dt, \qquad \widetilde{\mathbf{A}}_3[l,o] = \int_t B_l''(t)B_o''(t)dt. \tag{2.29}$$

2.2.6 Tensor Product P-splines

The first order difference penalty for data indexed over two variables is computed by summing over all of the squared row-wise and column-wise differences i.e.

$$\sum_j \sum_k (\alpha_{(j+1)k} - \alpha_{jk})^2 + \sum_j \sum_k (\alpha_{j(k+1)} - \alpha_{jk})^2. \tag{2.30}$$

In terms of matrices, this can be written as:

$$\boldsymbol{\alpha}^\top[(\mathbf{D}_{(1)}^\top\mathbf{D}_{(1)} \otimes \mathbf{I}_2) + (\mathbf{I}_1 \otimes \mathbf{D}_{(2)}^\top\mathbf{D}_{(2)})]\boldsymbol{\alpha} \tag{2.31}$$

where $\mathbf{D}_{(i)}^\top\mathbf{D}_{(i)}$ is the cross product of the matrix of differences for each of the indexing variables and $\mathbf{I}_i$ are identity matrices of dimension equal to the number of basis functions for each subscripted variables.

To obtain the difference penalty for data indexed over 3 variables i.e. spatio-temporal data, the penalty is similar with the addition of another matrix of differences for the third dimension i.e.

$$\boldsymbol{\alpha}^\top[(\mathbf{D}_{(1)}^\top\mathbf{D}_{(1)}\otimes\mathbf{I}_2\otimes\mathbf{I}_3)+(\mathbf{I}_1\otimes\mathbf{D}_{(2)}^\top\mathbf{D}_{(2)}\otimes\mathbf{I}_3)+(\mathbf{I}_1\otimes\mathbf{I}_2\otimes\mathbf{D}_{(3)}^\top\mathbf{D}_{(3)})]\boldsymbol{\alpha} \tag{2.32}$$

The matrices 2.31 and 2.32 are then multiplied by the smoothing parameter, λ, to give the model penalty term. Alternatively, separate smoothing parameters can be used to apply different levels of smoothness to each dimension, we will see later that this can be useful.

2.2.7 Choosing the Smoothing Parameter

Choosing the optimal smoothing parameter, λ, can be tackled through an optimality criterion. Several commonly used criterion for determining the optimal value of λ are outlined below. Here n is the number of observations, p is the number of parameters, RSS $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares and L is the value of the likelihood function for the fitted model.

- **Akaike's Information Criterion (AIC)** (see Akaike [1973])

  $$\text{AIC} = 2n - 2\log(L) \tag{2.33}$$

  A 'corrected' version of this criterion (AICc) was later proposed by Sugiura [1978]:

  $$\text{AICc} = \text{AIC} + \frac{2n(p+1)}{n-p-1}. \tag{2.34}$$

- **Bayesian Information Criterion (BIC)**(see Schwarz [1978])

  $$\text{BIC} = -2\log(L) + p\log(n). \tag{2.35}$$

  BIC generally imposes a stronger penalty on the number of parameters than AIC resulting in BIC preferring less complex models compared with AIC.

- **Cross-Validation (CV)** involves leaving out each data point in turn, building a model with the remaining data then using this model to predict the value for the omitted observation. The CV score is computed as:

$$\text{CV} = \frac{1}{n} \sum_{i} \left(y_i - \hat{y}_i^- \right)^2 , \tag{2.36}$$

where $\hat{y}_i^-$ is the predicted value for the i^{th} observation using a model that was built without the i^{th} observation; y_i is the observed value at this location and n is the number of observations. Computing the CV score can be time-consuming when large datasets are involved as it requires n models to be built for each value of of the smoothing parameter being assessed. To improve computational time and effort, the observations can be divided into k groups (folds) and each group of observations is left out in turn, the CV score is then computed in the same way but is averaged over k rather than n.

- **Generalised Cross Validation (GCV)**

$$\text{GCV} = \frac{n\text{RSS}}{\left(n - tr(\mathbf{S})\right)^2} \tag{2.37}$$

GCV provides a more computationally efficient version of CV. Rather than n computations of the criterion for each value of λ as required by CV, GCV only requires one computation of the criterion for each value of λ.

See Wood [2006] for further details.

Alternatively a Bayesian approach can be adopted as described by Evers et al. [2015], where λ is chosen as the value that maximises the posterior density. This is known as the MAP (maximum a posteriori) estimate. This approach to choosing λ is described further in Section 2.3.2.

2.3 Bayesian P-splines

2.3.1 Bayesian Regression

In the classical regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.38}$$

where $\mathbf{X}$ is an $n \times m$ matrix of explanatory variables with corresponding unknown model parameters $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)$; $\mathbf{y}$ is an $n \times 1$ vector of the response variable and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ is an $n \times 1$ vector of observation errors which are assumed to be independent with constant variance, σ^2. Thus:

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n). \tag{2.39}$$

Bayesian regression treats the unknown model parameters $(\boldsymbol{\beta}, \sigma^2)$ as random variables allowing for them to be described by a probability distribution, $f_{\boldsymbol{\beta},\sigma^2}$. A common choice of prior distribution for these parameter variables is the Normal-Inverse-Gamma distribution

$$f_{\boldsymbol{\beta},\sigma^2} \sim \mathcal{NIG}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}}, a, b), \tag{2.40}$$

which is equivalent to the product between a normal prior on $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \sigma^2\mathbf{V}_{\boldsymbol{\beta}})$ and an inverse gamma prior on $\sigma^2 \sim \mathcal{IG}(a, b)$ where $a, b > 0$. This is a common choice because it is conjugate.

Bayes' theorem then enables the joint posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$, conditional on the observed explanatory variables $\mathbf{X}$, and response variable $\mathbf{Y} = \mathbf{y}$, to be expressed as:

$$f_{\boldsymbol{\beta},\sigma^2|\mathbf{X},\mathbf{Y}} = \frac{f_{\mathbf{Y}|\mathbf{X},\boldsymbol{\beta},\sigma^2} f_{\boldsymbol{\beta},\sigma^2}}{f_{\mathbf{Y}}}. \tag{2.41}$$

This formulation allows $(\boldsymbol{\beta}, \sigma^2)$ to be updated by the observed data through the prior distribution of the model parameters, $f_{\boldsymbol{\beta},\sigma^2}$, which reflects initial beliefs about the model parameters, the likelihood function of the data, $f_{\mathbf{Y}|\boldsymbol{\beta},\sigma^2}$, and the marginal distribution of the data, $f_{\mathbf{Y}} = \int f_{\boldsymbol{\beta},\sigma^2} f_{\mathbf{Y}|\boldsymbol{\beta},\sigma^2} \; d\boldsymbol{\beta} d\sigma^2$.

2.3.2 Derivation of the Posterior Density of the Smoothing Parameter for a P-splines Model

Outlined in this section is the derivation of the posterior density of the smoothing parameter, λ proposed by Molinari [2014] for a p-splines model. The *maximum a posteriori* value of the derived distribution of λ is subsequently used to define the optimal smoothing parameter. This model is also used as the basis for the material in Chapter 5 where a second smoothing parameter is considered.

Model Summary

Building on the model specification detailed in Equation 2.39, the observation model is assumed as $\mathbf{Y}|\boldsymbol{\alpha}, \sigma^2, M_\lambda \sim \mathcal{N}_n(\mathbf{B}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_n)$ where M_λ is the model for a particular penalisation parameter term λ; $\mathbf{B} \in \mathbb{R}^{n\times m}$ is a matrix of B-spline basis functions and $\boldsymbol{\alpha} \in \mathbb{R}^m$ are the corresponding basis coefficients i.e. for $\mathbf{Y} = \mathbf{y}$,

$$f_{\mathbf{Y}|\boldsymbol{\alpha},\sigma^2,M_\lambda} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{B}\boldsymbol{\alpha})^\top(\mathbf{y}-\mathbf{B}\boldsymbol{\alpha})\right\}, \tag{2.42}$$

with $\mathbf{y} \in \mathbb{R}^n$.

A normal inverse gamma prior is placed on the parameters $\boldsymbol{\alpha}, \sigma^2 \sim \mathcal{NIG}(\boldsymbol{\mu}, \mathbf{V}(\lambda), a, b)$ i.e.

$$f_{\boldsymbol{\alpha},\sigma^2} = \frac{b^a}{(2\pi)^{m/2}\Gamma(a)|\mathbf{V}(\lambda)|^{1/2}}[\sigma^2]^{-(a+m/2+1)} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\alpha}-\boldsymbol{\mu})^\top\mathbf{V}(\lambda)^{-1}(\boldsymbol{\alpha}-\boldsymbol{\mu})+2b\right\}, \tag{2.43}$$

with $\boldsymbol{\mu} \in \mathbb{R}^m$, and a and b both in $\mathbb{R}^+$. In this model, $\boldsymbol{\mu}$ is set to 0 and the inverse of the hyperparameter $\mathbf{V}(\lambda)$ is set to $\mathbf{V}(\lambda)^{-1} = \lambda\mathbf{D}^\top\mathbf{D}$, where $\mathbf{D}$ is a d^{th} order difference matrix. An improper uniform prior is put on the penalty parameter λ, f_{M_λ}.

However using this formulation of $\mathbf{V}(\lambda)^{-1}$ results in singularity issues as $\mathbf{D}^\top\mathbf{D}$ is not of full rank and hence is not invertible. To overcome this issue, $\mathbf{V}(\lambda)^{-1}$ is reformulated as:

$$\mathbf{V}(\lambda)^{-1} = \lambda\mathbf{D}^\top\mathbf{D} + \tau\mathbf{I}_m. \tag{2.44}$$

The posterior distribution of the model M_λ, $f_{M_\lambda|\mathbf{y}}$ is then defined as the expression when $\tau \to 0$. Through derivation, the model parameters joint posterior distribution can be shown to be:

$$f_{\boldsymbol{\alpha},\sigma^2|\mathbf{Y},M_\lambda} \sim \mathcal{NIG}(\boldsymbol{\mu}^*, \mathbf{V}^*(\lambda), a^*, b^*) \tag{2.45}$$

Where, with $\boldsymbol{\mu} = \mathbf{0}$;

$$\begin{aligned}
\mathbf{V}^*(\lambda) &= (\mathbf{B}^\top\mathbf{B} + \mathbf{V}^{-1})^{-1} = (\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{D}^\top\mathbf{D})^{-1} \\
\boldsymbol{\mu}^* &= \mathbf{V}^*(\mathbf{B}^\top\mathbf{y} + \mathbf{V}^{-1}\boldsymbol{\mu}) = \mathbf{V}^*(\mathbf{B}^\top\mathbf{y}) \\
a^* &= a + \frac{n}{2} \\
b^* &= b + \frac{1}{2}\left[\mathbf{y}^\top\mathbf{y} + \boldsymbol{\mu}^\top\mathbf{V}^{-1}\boldsymbol{\mu} - (\boldsymbol{\mu}^*)^\top(\mathbf{V}^*)^{-1}\boldsymbol{\mu}^*\right] \\
&= b + \frac{1}{2}\left[\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{B}\mathbf{V}^*\mathbf{B}^\top\mathbf{y}\right] \\
&= b + \frac{1}{2}\mathbf{y}^\top\left[\mathbf{I}_n - \mathbf{B}(\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{B}^\top\right]\mathbf{y}
\end{aligned} \tag{2.46}$$

The likelihood, $f_{\mathbf{Y}|M_\lambda}$, is obtain as a by-product in this derivation. By Bayes' theorem

$$f_{M_\lambda|\mathbf{Y}} \propto f_{\mathbf{Y}|M_\lambda} f_{M_\lambda} \tag{2.47}$$

and so it can be shown that, in general,

$$f_{M_\lambda|\mathbf{Y}} \propto \frac{\Gamma(a^*)|\mathbf{V}^*(\lambda)|^{1/2}}{[b^*]^{a^*}|\mathbf{V}(\lambda)|^{1/2}} f_{M_\lambda}. \tag{2.48}$$

Retaining only the terms that depend on λ and substituting in the expressions for $\mathbf{V}^*(\lambda)$, $\boldsymbol{\mu}^*$, a^* and b^*, the posterior distribution for model M_λ is proportional to:

$$f_{M_\lambda|\mathbf{Y}} \propto \lambda^{\frac{rank(\mathbf{D}^\top\mathbf{D})}{2}} \times \frac{\Gamma(a^*)|\mathbf{B}^\top\mathbf{B}+\lambda\mathbf{D}^\top\mathbf{D}|^{-1/2}}{\left\{b+\frac{1}{2}\mathbf{y}^\top\left[\mathbf{I}_n-\mathbf{B}(\mathbf{B}^\top\mathbf{B}+\lambda\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{B}^\top\right]\mathbf{y}\right\}^{a^*}} f_{M_\lambda} \tag{2.49}$$

The optimal λ is taken as the MAP estimate of the log posterior distribution.

Improving Computational Efficiency

To evaluate $f_{M_\lambda|\mathbf{Y}}$ for each value of λ, the inverse and determinant of the $f \times f$ matrix $\mathbf{B}^\top\mathbf{B}+\lambda\mathbf{D}^\top\mathbf{D}$ must be computed, with $f = m^3$, where m is the number of one dimensional basis functions being used for each dimension. For l candidate values of λ the naive approach is of complexity $\mathrm{O}(l \times f^3)$. By utilising linear algebra methodology, the computational effort can be reduced to a single $\mathrm{O}(f^3)$ calculation followed by, for each candidate λ, a $\mathrm{O}(f)$ calculation, i.e. $\mathrm{O}(f^3 + f \times l)$, for l candidate values of λ.

To obtain this reduction in computational effort, $\mathbf{B}^\top\mathbf{B}+\lambda\mathbf{D}^\top\mathbf{D}$, can be jointly diagonalised in a similar way to Eldén [1977] and Wood [2000].

Since $\omega_0 = \mathbf{B}^\top\mathbf{B}+\mathbf{D}^\top\mathbf{D}$ is strictly positive definite, the Theorem of Spectral Decomposition (see Appendix B.1) can be applied to give:

$$\omega_0 = \mathbf{B}^\top\mathbf{B}+\mathbf{D}^\top\mathbf{D} = \mathbf{\Gamma}_0\mathbf{\Delta}_0\mathbf{\Gamma}_0{}^\top, \tag{2.50}$$

where the orthogonal matrix $\mathbf{\Gamma}_0$ contains the eigenvectors of $\mathbf{B}^\top\mathbf{B}+\mathbf{D}^\top\mathbf{D}$, and $\mathbf{\Delta}_0$ is a diagonal matrix containing the corresponding eigenvalues. Thus we can define $\omega_B = \left(\mathbf{B}\mathbf{\Gamma}_0\mathbf{\Delta}_0^{-1/2}\right)^\top\left(\mathbf{B}\mathbf{\Gamma}_0\mathbf{\Delta}_0^{-1/2}\right)$ and similarly apply the aforementioned theorem, to obtain:

$$\begin{aligned}\omega_B &= (\mathbf{B}\mathbf{\Gamma}_0\mathbf{\Delta}_0{}^{-1/2})^\top(\mathbf{B}\mathbf{\Gamma}_0\mathbf{\Delta}_0{}^{-1/2})\\ &= \mathbf{\Delta}_0{}^{-1/2}\mathbf{\Gamma}_0{}^\top\mathbf{B}^\top\mathbf{B}\mathbf{\Gamma}_0\mathbf{\Delta}_0{}^{-1/2}\\ &= \mathbf{\Gamma}_B\mathbf{\Delta}_B\mathbf{\Gamma}_B{}^\top,\end{aligned} \tag{2.51}$$

where $\mathbf{\Gamma}_B$ is orthogonal and $\mathbf{\Delta}_B$ is diagonal. Solving for $\mathbf{B}^\top\mathbf{B}$:

$$\mathbf{B}^\top\mathbf{B} = \underbrace{\boldsymbol{\Gamma}_0{\boldsymbol{\Delta}_0}^{1/2}\boldsymbol{\Gamma}_B}_{\mathbf{U}}\boldsymbol{\Delta}_B\underbrace{{\boldsymbol{\Gamma}_B}^\top{\boldsymbol{\Delta}_0}^{1/2}{\boldsymbol{\Gamma}_0}^\top}_{\mathbf{U}^\top} = \mathbf{U}\boldsymbol{\Delta}_B\mathbf{U}^\top \tag{2.52}$$

Thus, from Equation 2.50 it can be shown that:

$$\mathbf{B}^\top\mathbf{B} + \mathbf{D}^\top\mathbf{D} = \mathbf{U}\mathbf{U}^\top \tag{2.53}$$

Consequently

$$\mathbf{D}^\top\mathbf{D} = \mathbf{U}\mathbf{U}^\top - \mathbf{U}\boldsymbol{\Delta}_B\mathbf{U}^\top \tag{2.54}$$

and thus, combining Equations 2.51 and 2.54,

$$\begin{aligned}\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{D}^\top\mathbf{D} &= \mathbf{U}\boldsymbol{\Delta}_B\mathbf{U}^\top + \lambda(\mathbf{U}\mathbf{U}^\top - \mathbf{U}\boldsymbol{\Delta}_B\mathbf{U}^\top)\\ &= \mathbf{U}\underbrace{\left[\boldsymbol{\Delta}_B + \lambda(\mathbf{I} - \boldsymbol{\Delta}_B)\right]}_{\boldsymbol{\Delta}_\lambda}\mathbf{U}^\top\\ &= \mathbf{U}\boldsymbol{\Delta}_\lambda\mathbf{U}^\top.\end{aligned} \tag{2.55}$$

Thus,

$$\begin{aligned}(\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{D}^\top\mathbf{D})^{-1} &= (\mathbf{U}^\top)^{-1}\boldsymbol{\Delta}_\lambda^{-1}\mathbf{U}^{-1}\\ &= (\mathbf{U}^{-1})^\top\boldsymbol{\Delta}_\lambda^{-1}\mathbf{U}^{-1}.\end{aligned} \tag{2.56}$$

It can be shown that $|\mathbf{U}|^2 = |\boldsymbol{\Delta}_0|$; hence

$$|\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{D}^\top\mathbf{D}| = |\mathbf{U}|^2|\boldsymbol{\Delta}_\lambda| = |\boldsymbol{\Delta}_0||\boldsymbol{\Delta}_\lambda|. \tag{2.57}$$

Letting $\mathbf{w} = {\boldsymbol{\Gamma}_B}^\top{\boldsymbol{\Delta}_0}^{-1/2}{\boldsymbol{\Gamma}_0}^\top\mathbf{B}^\top\mathbf{y}$, from Equation 2.46, allows $f_{M_\lambda|\mathbf{y}}$ to be calculated for each value of λ without needing to compute $\hat{\boldsymbol{\alpha}}$ each time.

$$\begin{aligned} b^*(\lambda) &= b + \frac{1}{2}\mathbf{y}^\top \left[\mathbf{I}_n - \mathbf{B}(\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{B}^\top\right]\mathbf{y} \\ &= b + \frac{1}{2}\mathbf{y}^\top\mathbf{y} - \frac{1}{2}\mathbf{y}^\top\mathbf{B}(\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{B}^\top\mathbf{y} \\ &= b + \frac{1}{2}||\mathbf{y}||^2 - \frac{1}{2}\mathbf{w}^\top\boldsymbol{\Delta}_\lambda^{-1}\mathbf{w} \end{aligned} \tag{2.58}$$

Bringing together Equations 2.57 and 2.58 with Equation 2.49, the posterior distribution, $f_{M_\lambda|\mathbf{Y}}$ is obtained

$$f_{M_\lambda|\mathbf{Y}} \propto \lambda^{\frac{rank(\mathbf{D}^\top\mathbf{D})}{2}} \times \frac{[\Gamma(a^*)|\boldsymbol{\Delta}_0|^{-1/2}]|\boldsymbol{\Delta}_\lambda|^{-1/2}}{\left\{\left[b + \frac{1}{2}||\mathbf{y}||^2\right] - \frac{1}{2}\mathbf{w}^\top\boldsymbol{\Delta}_\lambda^{-1}\mathbf{w}\right\}^{a^*}} \; f_{M_\lambda}. \tag{2.59}$$

The posterior distribution of λ in the form shown in Equation 2.59 now only depends on λ through the inverse and determinant of $\boldsymbol{\Delta}_\lambda$, which are efficient to compute since $\boldsymbol{\Delta}_\lambda$ is diagonal. The expressions in square brackets, along with $\mathbf{w}$, only need to be computed once since they do not depend on λ.

2.4 Kriging

2.4.1 Geostatistical Processes

The geostatistical approach adopts the idea that the spatially distributed variable of interest, $\{Y(\mathbf{s}); \mathbf{s} \in D\}$, is a realisation from a spatial stochastic process (random field) indexed over spatial locations $\mathbf{s}$ within a fixed continuous study region $D \subset \mathbb{R}^2$. In reality, data are observed at a finite subset of locations n and are denoted as $\mathbf{y} = (y(\mathbf{s}_1), \cdots y(\mathbf{s}_n))$.

A common model assumes that the joint distribution of these observations is multivariate Gaussian. The process is then a Gaussian process which is completely defined by its mean function or first moment, $\mu(\mathbf{s}) = \mathbb{E}(y(\mathbf{s}))$, and its covariance or second moment, $C(\mathbf{s}, \mathbf{s}') = \text{cov}(y(\mathbf{s}), y(\mathbf{s}'))$.

A Geostatistical process can be described as *stationary* if the following assumptions are satisfied; Diggle and Ribeiro [2007]:

1. $\mu_y(\mathbf{s}) = \mu$ i.e. the mean of the process is constant for all spatial locations $\mathbf{s}$

2. The covariance function $\text{cov}(y(\mathbf{s}), y(\mathbf{s}')) = C_Y(h)$, where $h = \mathbf{s} - \mathbf{s}'$ i.e. the covariance only depends on the distance between the observation locations.

Moreover, this process can be described as *isotropic* if $C_y(h) = C_y(||h||)$ where $||.||$ denotes the Euclidean distance, i.e. the covariance does not depend on the direction between two spatial locations.

2.4.2 Covariance Functions

Covariance functions can be used to model the correlation between observations. There are parametric families of functions used to define an appropriate class of covariance functions; Diggle and Ribeiro [2007]. Described below are some of the more commonly used covariance functions which are known to be positive definite, a necessary condition of the covariance function.

- Exponential
$$C_y(h) = \begin{cases} \sigma^2 + \tau^2 & h = 0 \\ \sigma^2 \text{exp}(-h/\phi) & h > 0 \end{cases}$$

- Gaussian
$$C_y(h) = \begin{cases} \sigma^2 + \tau^2 & h = 0 \\ \sigma^2 \text{exp}(-(h/\phi)^2) & h > 0 \end{cases}$$

- Power Exponential
$$C_y(h) = \begin{cases} \sigma^2 + \tau^2 & h = 0 \\ \sigma^2 \text{exp}(-|h/\phi|^r) & h > 0 \end{cases}$$
  where $0 < r \leq 2$ i.e. non integers.

- Spherical
$$C_y(h) = \begin{cases} \sigma^2 + \tau^2 & h = 0 \\ \sigma^2 [1 - \frac{3}{2}(h/\phi) + \frac{1}{2}(h/\phi)^3] & 0 < h \leq \phi \\ 0 & h > \phi \end{cases}$$

- Matérn
$$C_y(h) = \begin{cases} \sigma^2 + \tau^2 & h = 0 \\ \sigma^2 \frac{2^{1-\kappa}}{\Gamma(\kappa)} (h/\phi)^\kappa K_\kappa(h/\phi) & h > 0 \end{cases}$$

where, for all functions, $h = ||\mathbf{s}_i - \mathbf{s}_j||$.

In each of these functions, σ^2 denotes the 'partial sill' parameter which is the limit of the covariance as the distance tends towards 0; ϕ denotes the range parameter, which is the distance between observations at which the covariance is close to 0 and τ^2 is the nugget parameter which quantifies the measurement error or non-spatial variation.

In the Matérn covariance, $K_\kappa()$ denotes a modified Bessel function of order κ where $\kappa > 0$ is a shape parameter. When $\kappa = 0.5$, the Matérn function reduces to the exponential function, while as $\kappa \rightarrow \infty$ the Matérn function approaches the Gaussian covariance function.

Parameter Estimation using Maximum Likelihood

A geostatistical process can be modelled as:

$$Y(\mathbf{s}) = \boldsymbol{\mu}_y(\mathbf{s}) + \boldsymbol{\epsilon}_y(\mathbf{s}) \tag{2.60}$$

where:

$$\hat{\boldsymbol{\mu}}_y = (\hat{\mu}_y(\mathbf{s}_1), \cdots, \hat{\mu}_y(\mathbf{s}_n)) = \mathbf{X}\boldsymbol{\beta} \tag{2.61}$$

and

$$\boldsymbol{\epsilon}_y = (\epsilon_y(\mathbf{s}_1), \cdots, \epsilon_y(\mathbf{s}_n)) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \tag{2.62}$$

where $\mathbf{X}$ is a design matrix of the covariates with corresponding coefficients $\boldsymbol{\beta}$, $\boldsymbol{\theta} = (\sigma^2, \tau^2, \phi)$ are the covariance parameters and $\boldsymbol{\Sigma}(\boldsymbol{\theta})_{ij} = C_y(||\mathbf{s}_i - \mathbf{s}_j||, \boldsymbol{\theta})$, where $C_y()$ is a chosen covariance function. Consequently, given geostatistical data $\mathbf{y} = (y(\mathbf{s}_1), \cdots y(\mathbf{s}_n))$, the geostatistical model considered is:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta})). \tag{2.63}$$

Classically, the parameters $(\boldsymbol{\beta}, \sigma^2, \tau^2, \phi)$ in this model are estimated using maximum likelihood which involves choosing the parameter values which maximise the log likelihood function of $\mathbf{y}$ based on the multivariate Gaussian assumption i.e.

$$\ln(f(\mathbf{y})) \propto -\frac{1}{2}\ln(|\boldsymbol{\Sigma}(\boldsymbol{\theta})|) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.64}$$

Maximisation can be performed by computer optimisation.

2.4.3 Prediction with a Gaussian Spatial Process - Kriging

Kriging was first proposed by Krige [1951] as a method for prediction from a spatial Gaussian process. The Kriging predictor is based on deriving the Best Linear Unbiased Predictor (BLUP) for a new spatial location, $\mathbf{s}_0$, given current observations $\mathbf{y}$ and is obtained by minimising the mean square prediction error (MSPE):

$$\text{MSPE} = \mathbb{E}\left[(y(\mathbf{s}_0) - \hat{y}(\mathbf{s}_0))^2\right] \tag{2.65}$$

It can be shown that $\hat{y}(\mathbf{s}_0) = \mathbb{E}\left(y(\mathbf{s}_0)|\mathbf{y}(\mathbf{s})\right)$ and hence the Conditional Distribution Property of a Multivariate Gaussian Distribution can be used, see A.1.

Application of the aforementioned property allows the optimal predictor to be derived for $y(\mathbf{s}_0)$ given $\mathbf{y} = (y(\mathbf{s}_1), ..., y(\mathbf{s}_n))^\top$. The joint geostatistical process of a new observation at location $\mathbf{s}_0$ and the current observations can be defined as,

$$\begin{pmatrix} y(\mathbf{s}_0) \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_y \\ \boldsymbol{\mu}_y \end{pmatrix}, \begin{pmatrix} k & \mathbf{c}_0^\top \\ \mathbf{c}_0 & \mathbf{K} \end{pmatrix}\right). \tag{2.66}$$

Here

$$k = C_y(||\mathbf{s}_0 - \mathbf{s}_0||; \boldsymbol{\theta})$$

$$\mathbf{c}_0 = (C_y(||\mathbf{s}_0 - \mathbf{s}_1||; \boldsymbol{\theta}), \ldots, C_y(||\mathbf{s}_0 - \mathbf{s}_n||; \boldsymbol{\theta}))$$

$$\mathbf{K}_{ij} = \boldsymbol{\Sigma}(\boldsymbol{\theta})_{ij} = C_y(||\mathbf{s}_i - \mathbf{s}_j||; \boldsymbol{\theta}) \qquad i, j \in 1, \cdots, n$$

where $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2)$ are the covariance model parameters and $C_y()$ is a chosen covariance function (see Section 2.4.2).

It follows that

$$\mathbb{E}[y(\mathbf{s}_0)|\mathbf{y}] = \hat{\mu}_y + \mathbf{c}_0^\top \mathbf{K}^{-1}(\mathbf{y} - \hat{\mu}_y \mathbf{1}) \tag{2.67}$$

and,

$$\mathbf{C}_{\mathbf{s}_0|\mathbf{y}} = \text{var}(y(\mathbf{s}_0)|\mathbf{y}) = k - \mathbf{c}_0^\top \mathbf{K}^{-1}\mathbf{c}_0. \tag{2.68}$$

Equation 2.67 is the **Ordinary Kriging predictor**, which assumes μ_y is constant but unknown. When μ_y is non-constant, i.e. includes location specific covariate values, and is unknown we have the **Universal Kriging predictor**.

2.4.4 Spatio-temporal Geostatistical Processes

Suppose the data are now collected over space and time i.e. indexed as $(\mathbf{s}_i, t_i)$. In this instance, the geostatistical process for spatio-temporal data, $\mathbf{y} = (y(\mathbf{s}_1, t_1), ..., y(\mathbf{s}_n, t_m))^\top$ is

$$Y(\mathbf{s}) = \boldsymbol{\mu}_y(\mathbf{s}, \mathbf{t}) + \boldsymbol{\epsilon}_y(\mathbf{s}, \mathbf{t}) \tag{2.69}$$

where

$$\hat{\boldsymbol{\mu}}_y = (\hat{\mu}_y(\mathbf{s}_1, t_1), \cdots, \hat{\mu}_y(\mathbf{s}_n, t_m)) = \mathbf{X}\boldsymbol{\beta} \tag{2.70}$$

and

$$\boldsymbol{\epsilon}_y = (\epsilon_y(\mathbf{s}_1, t_1), \cdots, \epsilon_y(\mathbf{s}_n, t_m)) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})). \tag{2.71}$$

The geostatistical model is then of the same form as that for spatial data denoted in Equation 2.63 and the model parameters can be estimated by maximum likelihood.

The covariance structures for space and time should be treated differently; Cressie and Wikle [2011] describe several of these structures in detail. One of the most common and simple covariance structures assumes that the spatial and temporal covariances are separable, i.e. they act independently, so that

$$C_y((\mathbf{s},t),(\mathbf{s}',t')) = C_y(t,t')C_y(\mathbf{s},\mathbf{s}'). \tag{2.72}$$

The spatio-temporal process is said to be *stationary* if both the spatial and temporal covariances satisfy the stationarity assumptions detailed in the spatial Kriging section (Section 2.4.1). Under this assumption of stationarity and isotropy, a separable spatio-temporal covariance function can be denoted as:

$$C_y(t,t') \otimes C_y(\mathbf{s},\mathbf{s}') = C_y(u) \otimes C_y(h) \tag{2.73}$$

where $h = ||\mathbf{s}-\mathbf{s}'||$ and $u = ||t-t'||$. To compute the separable covariance matrix, the Kronecker product, $\otimes$, is used to obtain the covariances between all possible space and time combinations.

2.4.5 Prediction with a Gaussian Spatio-temporal Process

Assuming a separable covariance structure for the current observations, the joint geostatistical process of a new observation at location $\mathbf{s}_0$ and time t_0 can be defined similarly to spatial Kriging in Section 2.4.3, as:

$$\begin{pmatrix} y(\mathbf{s}_0,t_0) \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_y \\ \boldsymbol{\mu}_y \end{pmatrix}, \begin{pmatrix} k & \mathbf{c}_0^\top \\ \mathbf{c}_0 & \mathbf{K} \end{pmatrix}\right). \tag{2.74}$$

Here

$$k = C_y(t_0 - t_0; \boldsymbol{\theta}_t) \otimes C_y(\mathbf{s}_0 - \mathbf{s}_0; \boldsymbol{\theta}_s),$$

$$\mathbf{c}_0 = (C_y(t_0 - t_1; \boldsymbol{\theta}_t) \otimes C_y(\mathbf{s}_0 - \mathbf{s}_1; \boldsymbol{\theta}_s), \ldots, C_y(t_0 - t_m; \boldsymbol{\theta}_t) \otimes C_y(\mathbf{s}_0 - \mathbf{s}_n; \boldsymbol{\theta}_s)),$$

$$\mathbf{K} = \boldsymbol{\Sigma}^{(t)} \otimes \boldsymbol{\Sigma}^{(s)},$$

where $\boldsymbol{\Sigma}^{(s)}_{ij} = C_y(||\mathbf{s}_i - \mathbf{s}_j||; \boldsymbol{\theta}_s)$ is the spatial covariance matrix, $\boldsymbol{\theta}_s = (\sigma^2_s, \tau^2_s, \phi_s)$ are the spatial covariance parameters, $\boldsymbol{\Sigma}^{(t)}_{ij} = C_y(||t_i - t_j||; \boldsymbol{\theta}_t)$ is the temporal covariance matrix, $\boldsymbol{\theta}_t = (\sigma^2_t, \tau^2_t, \phi_t)$ are the temporal covariance parameters and $C_y()$ is a covariance function from Section 2.4.2.

It follows, from the Conditional Distribution Property of a Multivariate Gaussian distribution, that the **Ordinary Kriging Predictor** for spatio-temporal data is

$$\mathbb{E}[y(\mathbf{s}_0, t_0)|\mathbf{y}] = \hat{\mu}_y + \mathbf{c}_0^\top \mathbf{K}^{-1}(\mathbf{y} - \hat{\mu}_y \mathbf{1}) \tag{2.75}$$

and

$$\mathbf{C}_{\mathbf{s}_0, t_0|\mathbf{y}} = \text{var}(y(\mathbf{s}_0, t_0)|\mathbf{y}) = k - \mathbf{c}_0^\top \mathbf{K}^{-1}\mathbf{c}_0. \tag{2.76}$$

Chapter 3

A Comparison of Spatial and Spatio-temporal Modelling Methods for Contaminated Groundwater

Modelling groundwater contamination can be difficult due to the impracticalities and cost of obtaining samples from every monitoring well at every sampling period. In some cases only a very small proportion of the wells may be sampled at one time or the samples can be sporadic with some wells remaining unsampled for long periods of time due to their proximity to other wells.

The main objective of this study was to compare the predictive performance of spatial and spatio-temporal modelling techniques, to determine whether the added computational complexity of constructing spatio-temporal models has any increased benefits on the resulting predictions compared with spatial models. Spatial models for interpolating contaminant plumes in groundwater are already widely used e.g. Elumalai et al. [2017], Reed et al. [2004], *Surfer® 16 from Golden Software, LLC* [2018], whereas methods which also take into account temporal information i.e. spatio-temporal models, are a lot less common.

3.1 Data Simulation

For this study, groundwater data were simulated from the partial differential equation (PDE) detailed in Equation 3.1. A slightly modified version of this data was used by Evers et al. [2015] for a comparison of methods for selecting a smoothing parameter for a spatio-temporal p-splines model.

$$\frac{\partial y}{\partial t} = D \cdot \left(\frac{\partial^2 y}{\partial x_1^2} + \frac{\partial^2 y}{\partial x_2^2} \right) + \omega_1(x_1, x_2)\frac{\partial y}{\partial x_1} + \omega_2(x_1, x_2)\frac{\partial y}{\partial x_2} \tag{3.1}$$

where y are the contaminant concentrations, x_1 and x_2 are the spatial coordinates and $t \in [0, 1]$ denotes time. In the first term, D is a constant controlling how quickly the solute spreads. This is combined with the sum of the 2^{nd} partial derivatives to give a term which describes the spread by diffusion of the contaminant in the groundwater. The remaining two advection terms describe how the contaminant is affected by groundwater flow, where ω_1 and ω_2 describe its direction and velocity in each direction respectively. These functions were chosen based on observed groundwater levels at a current site. An additional spatial Matérn effect was added to the simulated data.

Observed measurements were generated by interpolating the PDE at sampling locations obtained from a set of real site locations. The true concentrations (i.e. test data) were obtained by interpolating the numerical solution to the PDE, computed over a $100 \times 100 \times 100$ grid. Once these measurements were generated, well-specific and measurement noise were added. The initial contaminant plume is shown in the top left panel of Figure 3.1. The remaining three plots show the spread of the plume at subsequent times, $t \in \{0.25, 0.50, 0.75\}$.

Two sampling scenarios were created from a network of 29 wells (displayed as points in Figure 3.1). The first (scenario 1) used the exact design obtained from a current site. It is often very impractical and not always viable to obtain samples from every well in the network during each sampling period, therefore the simulated data have several incomplete sampling periods. For this scenario there were a total of 1400 observations spread over 167 sampling periods. The second scenario (scenario 2) consisted of every well being sampled at all 167 sampling periods i.e. 4843 observations. In terms of

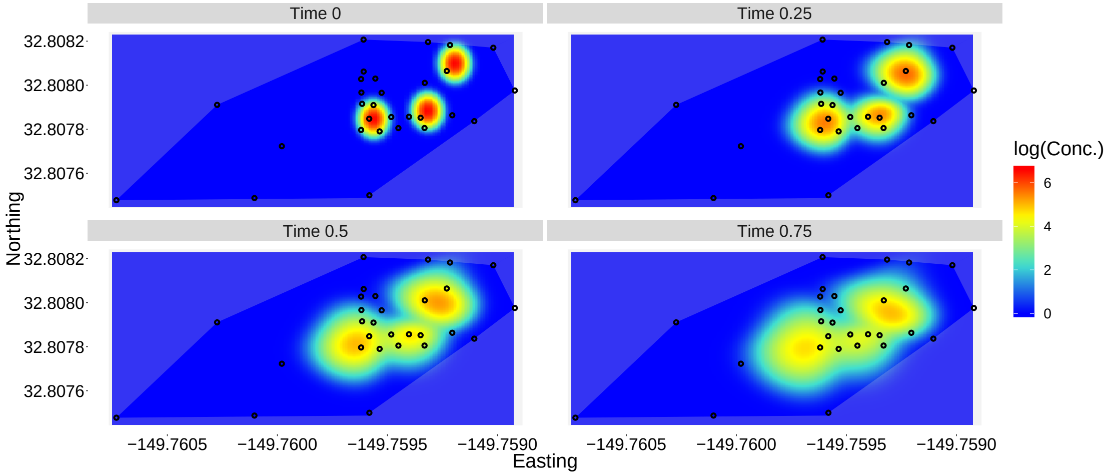

FIGURE 3.1: True underlying PDE described in Equation 3.1 (PDE1) at times $t \in \{0, 0.25, 0.50, 0.75\}$

gathering as much information as possible, this scenario is more appealing; however, it is not practical or cost-effective.

Before the data were used for analysis a $\log(y + 1)$ transform was applied to the concentration values. The $+1$ was included to account for the simulation occasionally producing concentration values at, or very close to, 0.

3.2 Results

Spatial and spatio-temporal p-spline models, along with spatial Kriging, were applied to the data simulated in Section 3.1. When implementing spatial Kriging, exponential and Matérn (with κ fixed to 2) covariance functions were considered. Kriging was included in the study as it is one of the most common spatial interpolation methods Li and Heap [2014]. For spatial and spatio-temporal p-splines a first order difference penalty was used along with B-spline basis functions of degree three. Two combinations of basis functions were used for the spatio-temporal model, the first had (14, 8, 3) basis functions corresponding to the easting, nothing and time components respectively and the second had (25, 15, 3) basis functions. The temporal component was given a lower number of basis functions to reflect the fact that in this dataset the contaminant concentrations vary more over space than they do over time. Computation of the spatio-temporal model is very time consuming and dependent on the number of basis functions, thus a model with fewer basis functions was used to assess the accuracy of the results with a model that takes less time to compute. As mentioned in Section 2.2.4 there are several ways to determine the most appropriate smoothing parameter for penalised spline-based models. In this study the Bayesian MAP estimate was used as described in Section 2.3.2 and by Evers et al. [2015] for both the spatial and spatio-temporal p-spline models.

To determine the predictive performance of each of the methods, mean square prediction errors (MSPE) were computed and compared. The MSPE at time t is defined as:

$$\text{MSPE}_t = \frac{1}{n} \sum_j \left(y_{\mathbf{s}_j t} - \hat{y}_{\mathbf{s}_j t} \right)^2 , \tag{3.2}$$

where $y_{\mathbf{s}_j t}$ is the true value from the test data at spatial prediction location $\mathbf{s}_j = (x_{1j}, x_{2j})$ and prediction time t, $\hat{y}_{\mathbf{s}_j t}$ is the fitted value from the model at spatial location $\mathbf{s}_j$ and prediction time t, and n is the total number of prediction locations at time t. A low value of MSPE indicates that the model has predicted well.

For each sampling scenario, predictions were obtained for three time points; time 100 which is located approximately half way through the data, and has samples from 11 wells under scenario 1; time 167 which is the final time for which data are available i.e. the most recent time point - under scenario 1 this time point had observations from 14 wells; and finally time 100, but only using the data up until this time point - this time point was included as only 16 of a possible 29 wells had samples taken by this point.

3.2.1 Sampling Scenario 1 - A Realistic Design

Table 3.1 shows the MSPE for predictions at each time point. At all three prediction times, the spatio-temporal p-spline models perform best, with the model containing (25, 15, 3) basis functions performing best at time 100 and the model with (14, 8, 3) performing best for time 100 when only prior samples are used and also time 167. The spatial methods perform less well. At time 100, the two Kriging models perform similarly with a MSPE of ~ 1.05 whilst the spatial p-splines performed slightly better with a MSPE of 0.93. At time 167 the three spatial methods have very similar MSPEs of ~ 1.1, which is significantly higher than the spatio-temporal methods with a MSPE of ~ 0.4.

The difference in MSPE for the spatio-temporal and spatial methods is probably caused by the sampling scheme missing areas of high contaminant concentration due to wells not being sampled in certain locations. The spatio-temporal methods capture these areas by exploiting temporal smoothness in the models, so that information from earlier sampling events aids prediction.

When predicting for time 100 using prior samples only, the spatio-temporal methods do not have the significant improvement over the spatial methods as was seen for the other two prediction times. This can be explained by looking at the sampling scheme. By time 100, only 16 of the 29 available wells had samples recorded and, of these 16 wells, 11 were sampled at time 100. Thus, at this time point, the spatio-temporal models only have extra information from 5 wells. The location of these additional five wells is around the perimeter of the study region where there is little to no contamination present and thus they provide little additional information about where the contaminant cloud is located.

Table 3.1: Mean Square Prediction Errors (MSPEs) and standard errors (SEs) at the 100th (using prior samples only), 100th (using all samples) and 167th time points under the realistic sampling scenario (200 simulations). Note: in the spline models the values in brackets indicate the basis functions for each direction and for Kriging in brackets is the covariance function used.

Modelling Method	Time 100 (Prior)	Time 100	Time 167
Spatio-temporal P-splines (25, 15, 3)	0.8281 (0.0095)	0.3526 (0.0089)	0.4878 (0.0075)
Spatio-temporal P-splines (14, 8, 3)	0.7329 (0.0125)	0.4864 (0.0146)	0.3736 (0.0106)
Spatial P-splines	0.9372 (0.0158)	0.9372 (0.0158)	1.0907 (0.0224)
Spatial Kriging (Matern)	1.0471 (0.0257)	1.0471 (0.0257)	1.0594 (0.0231)
Spatial Kriging (Exponential)	1.0798 (0.0216)	1.0798 (0.0216)	1.1294 (0.0365)

From the prediction surfaces for one simulation displayed in Figures 3.3, 3.4 and 3.5 at all prediction times the models are able to identify the main mass of contamination in the centre of the study region where most wells are located.

The differences in the prediction surfaces lie around the boundary where at certain times no wells are sampled. For time 100 (Figure 3.4) the spatial methods are unable to predict the diffusion of the contaminant cloud into the north region of the plot due to no wells being sampled in this area at this time point. This is also the case for the

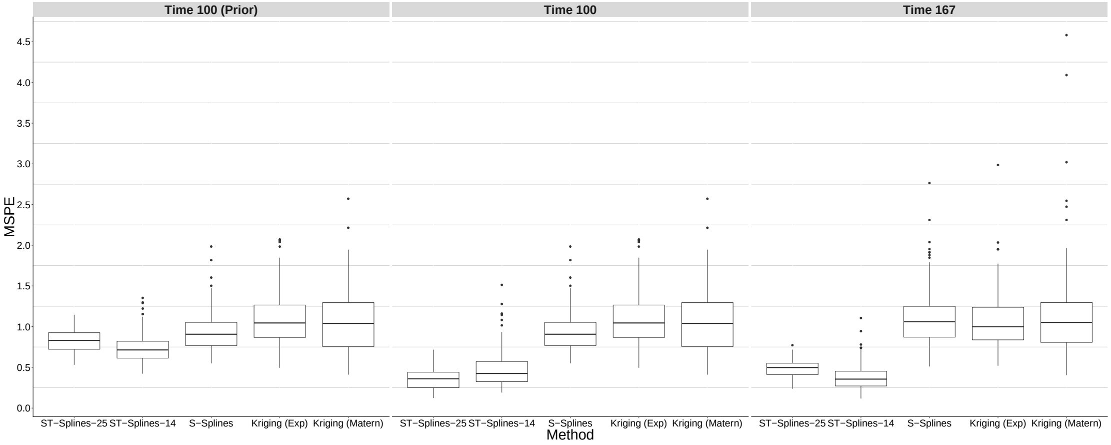

FIGURE 3.2: Boxplots of the MSPE for each modelling method at each prediction time under the realistic sampling scenario (200 simulations). The number after each spatio-temporal splines model indicates the number of basis functions for the easting component

spatio-temporal p-spline models when predicting at time 100 using prior samples only (Figure 3.3) since by this time only 16 of the 29 wells have had samples taken and none are located in this area. In comparison, when predicting at time 100 using all of the data, the spatio-temporal p-spline models are able to identify the restricted spread in the northerly direction by using information from later sampling times. A similar case is present when predicting at time 167 (Figure 3.5) where again the spatial methods are unable to predict the diffusion of the contaminant; however, this time it is in a southerly direction. The spatio-temporal methods have the advantage that earlier observations at these boundary wells inform the model more accurately of where the contaminant plume lies. Comparing the predicted surfaces from the two spatio-temporal models, the improved MSPE for the lower resolution model over the higher resolution model appears to stem from the higher resolution model tracking the data too closely.

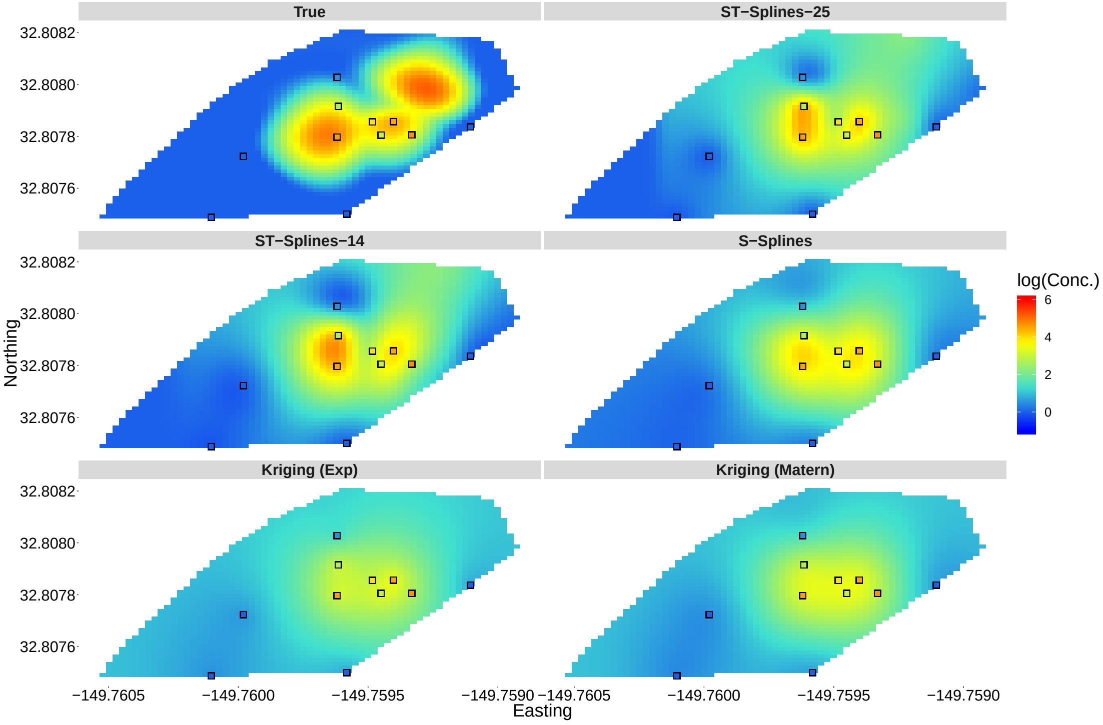

FIGURE 3.3: Predicted surfaces, of one simulation, for each method at time 100 using only samples prior to this time point under the realistic sampling scenario. The plot in the top left location is the true surface from PDE1. The number after each spatio-temporal splines model indicates the number of basis functions for the easting component.

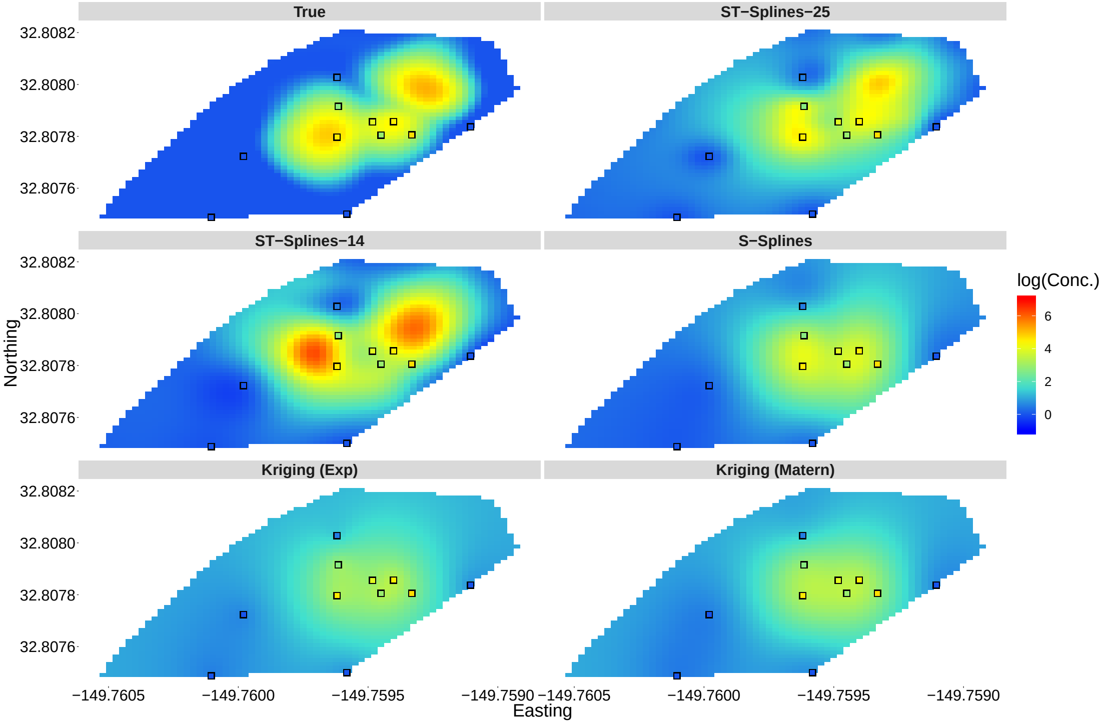

FIGURE 3.4: Predicted surfaces, of one simulation, for each method at time 100 using the entire sampling data under the realistic sampling scenario. The plot in the top left location is the true surface from PDE1. The number after each spatio-temporal splines model indicates the number of basis functions for the easting component.

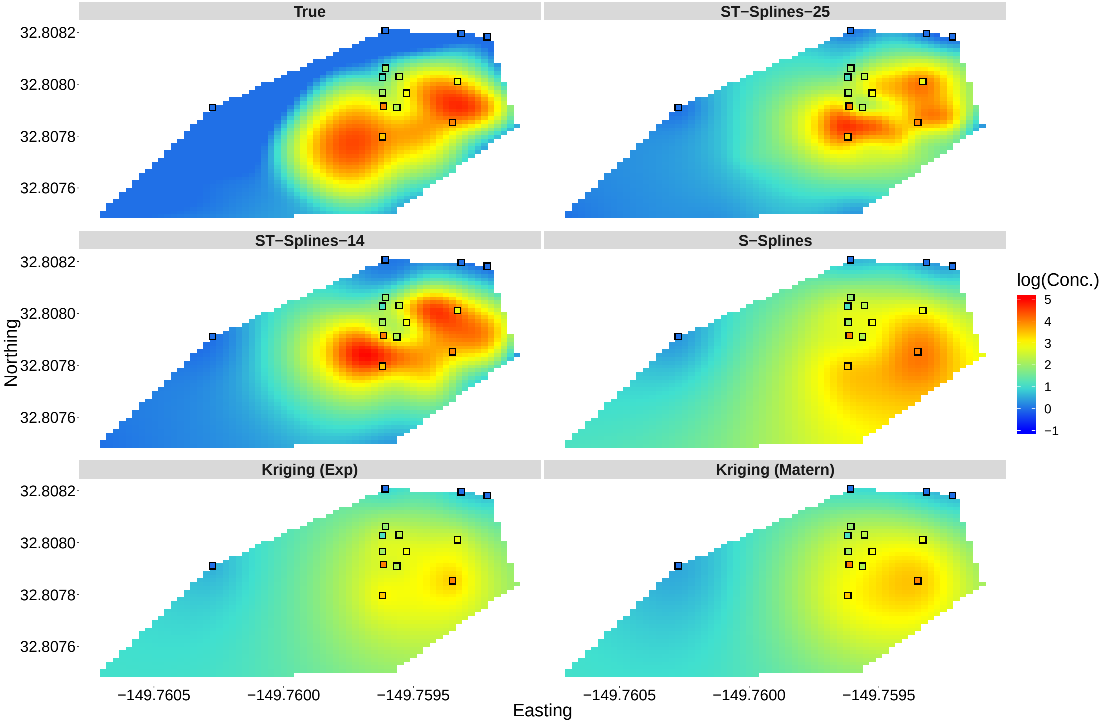

FIGURE 3.5: Predicted surfaces, of one simulation, for each method at time 167 (most recent time point) using the entire sampling data under the realistic sampling scenario. The plot in the top left location is the true surface from PDE1. The number after each spatio-temporal splines model indicates the number of basis functions for the easting component.

3.2.2 Sampling Scenario 2 - A Full Design

The same modelling methods were then applied to data obtained from a full sampling design. For this set of data the best performing method for all time points was spatio-temporal p-splines with (25, 15, 3) basis functions. However, as expected, the improvement in predictive performance of the spatio-temporal methods over the spatial methods is less significant compared with sampling scenario 1. As in the first scenario, the three spatial methods had similar results with a MSPE of ~ 0.50 at time 100 and ~ 0.6 at time 167. The spatio-temporal p-splines model with (14, 8, 3) basis functions had a higher MSPE of 1.20 at time 100 and 0.71 at time 167. The boxplots of the simulation results shown in Figure 3.6 highlight several outliers for the lower resolution spatio-temporal p-splines model. It is likely that the poor performance of this model is due to an effect known as 'ballooning', where unusually high or low predictions are made in areas with little data support.

Table 3.2: Mean Square Prediction Errors (MSPEs) and standard errors (SEs) at the 100th (using prior samples only), 100th (using all samples) and 167th time points under the full sampling scenario (200 simulations) Note: in the spline models the values in brackets indicate the basis functions for each direction and for Kriging in brackets is the covariance function used.

Modelling Method	Time 100 (Prior)	Time 100	Time 167
Spatio-temporal P-splines (25, 15, 3)	0.3001 (0.0067)	0.3071 (0.0065)	0.4827 (0.0070)
Spatio-temporal P-splines (14, 8, 3)	0.6418 (0.0391)	1.2012 (0.0945)	0.7090 (0.0422)
Spatial P-splines	0.5013 (0.0097)	0.5013 (0.0097)	0.6677 (0.0094)
Spatial Kriging (Matern)	0.4610 (0.0137)	0.4610 (0.0137)	0.5931 (0.0153)
Spatial Kriging (Exponential)	0.4881 (0.0120)	0.4881 (0.0120)	0.6358 (0.0139)

The prediction surfaces for a single simulation are displayed in Figures 3.7, 3.8 and 3.9. In comparison with the realistic sampling scenario, under the full design the spatial methods predict surfaces which are much more similar to those of the spatio-temporal methods. This is not surprising given every well is sampled at every time point and so the amount of additional information brought by a spatio-temporal model is limited. Looking at the prediction surfaces for the spatio-temporal p-splines model with a lower number of basis functions (top right panel of each surface plot), ballooning is evident in

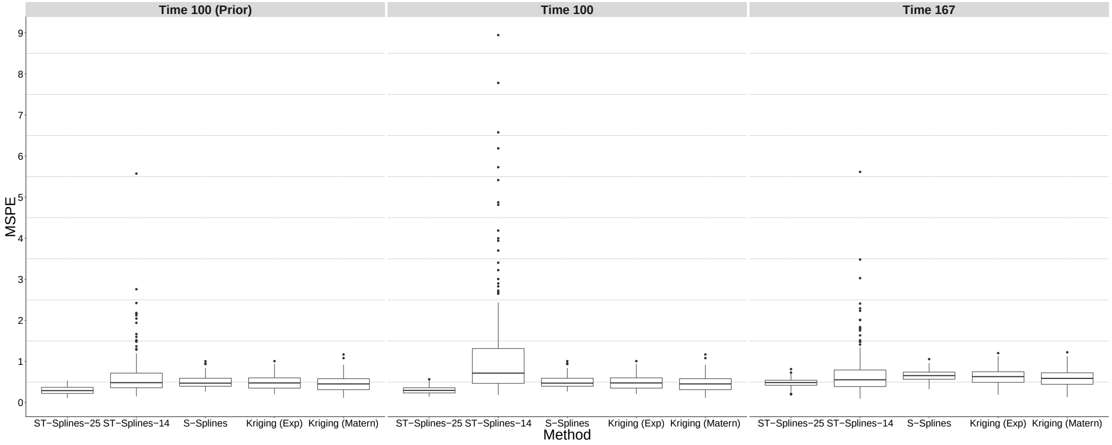

FIGURE 3.6: Boxplots of the MSPE for each modelling method at each prediction time under the full sampling scenario (200 simulations). The number after each spatio-temporal splines model indicates the number of basis functions for the easting component

the central region of the plot. Chapter 5 investigates ballooning further and a measure for its detection is presented.

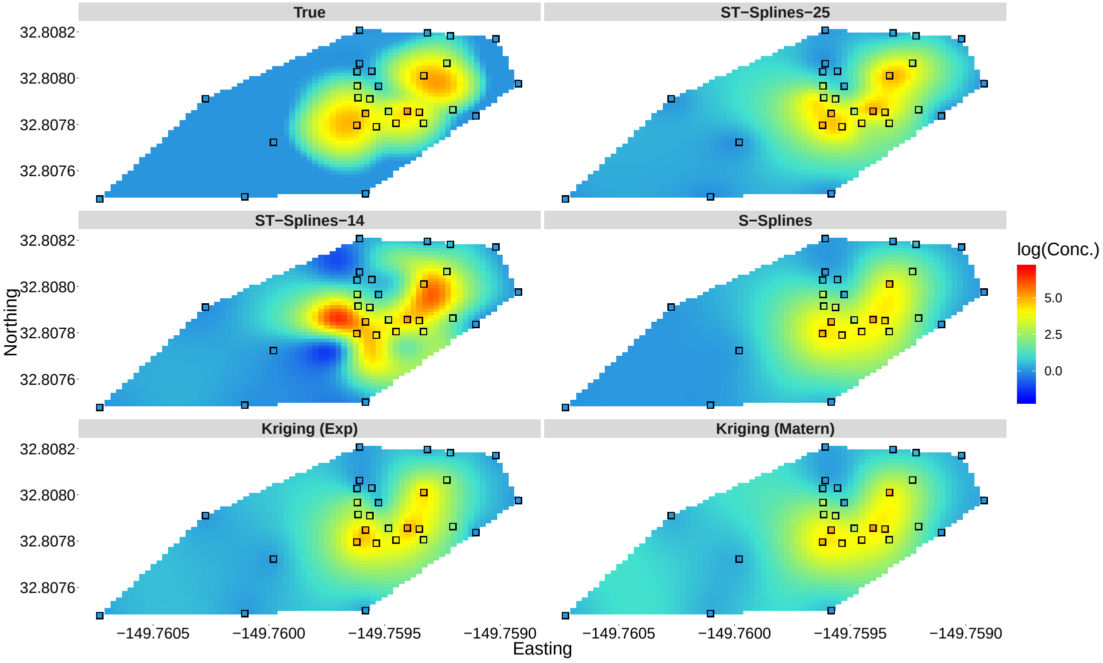

FIGURE 3.7: Predicted surfaces, of one simulation, for each method at time 100 using only samples prior to this time point under the full sampling scenario. The plot in the top left location is the true surface from PDE1. The number after each spatio-temporal splines model indicates the number of basis functions for the easting component.

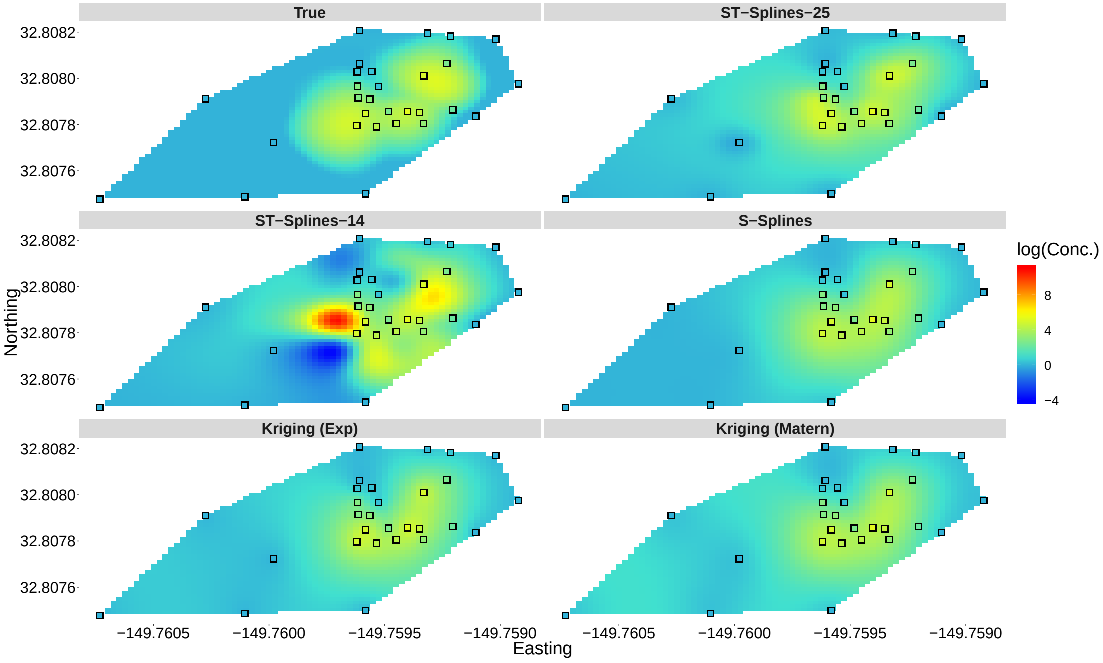

FIGURE 3.8: Predicted surfaces, of one simulation, for each method at time 100 using the entire sampling data under the realistic sampling scenario. The plot in the top left location is the true surface from the PDE1. The number after each spatio-temporal splines model indicates the number of basis functions for the easting component.

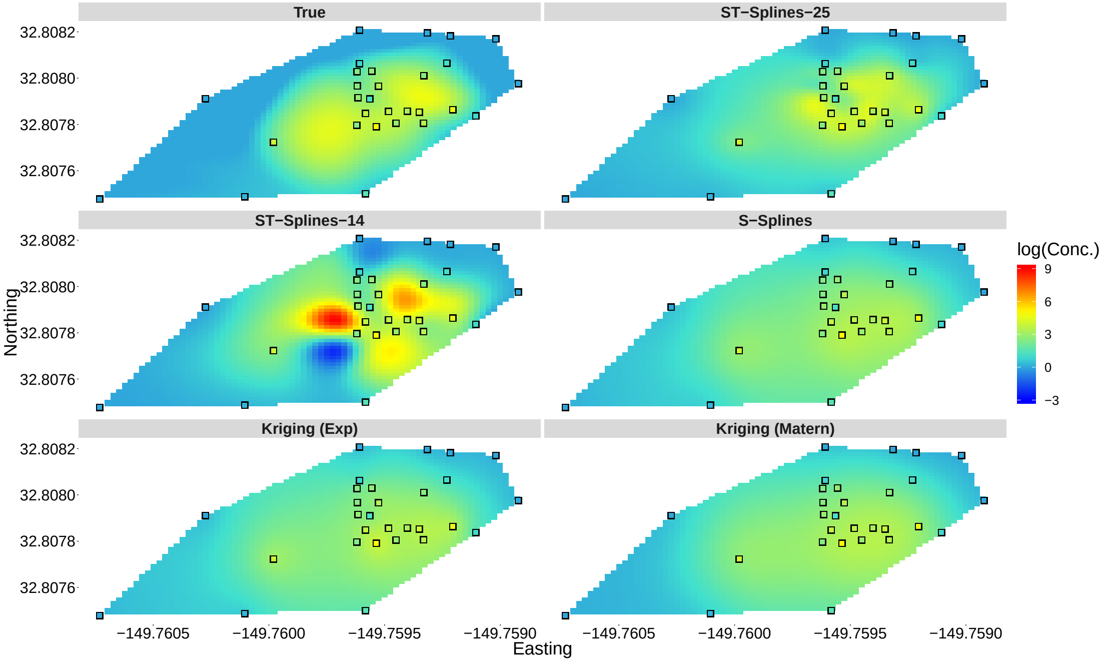

Figure 3.9: Predicted surfaces, of one simulation, for each method at time 167 (most recent time point) using the entire sampling data under the realistic sampling scenario. The plot in the top left location is the true surface from the PDE1. The number after each spatio-temporal splines model indicates the number of basis functions for the easting component.

3.2.3 Computational Time

From the results in Section 3.2 it is apparent that using a spatio-temporal model with a significant number of basis functions improves predictions over a spatial model. However, as would be expected, the computation times for determining parameters in the spatio-temporal models are considerably greater than those of the spatial models. Evers et al. [2015] used several linear algebra methods to significantly improve this computation time; see Section 2.3.2 for details. However, the time taken for fitting a spatial model will never be matched. Table 3.3 contains the mean computation times for each of the spatio-temporal models considered in Section 3.2, separated by each sampling scenario. For comparison the computation time of the spatial methods is almost instantaneous (less than 1 second).

Table 3.3: Mean computational time in seconds for each spatio-temporal model used in Section 3.2. The timings are separated by sampling design, where the first design contained 1400 samples and the second 4843 samples. The number of basis functions given are for each dimension i.e. (easting, northing, time).

No. of Basis functions	Realistic Design (1400 samples)	Full Design (4843 samples)
25, 15, 3	25.81	44.41
14, 8, 3	3.68	4.32

3.3 Reducing the Number of Samples

The primary aim of this study, from an industry perspective, was to determine whether the quantity of data collected can be reduced, whilst retaining the prediction accuracy, by predicting the state of the study region using a spatio-temporal model rather than a spatial model. By reducing the number of samples, the costs associated with data collection will also reduce. To assess whether this could be achieved, two data removal techniques were considered and simulations were performed. The first, referred to as 'observation removal', involved starting with a full design i.e. samples were obtained from every well at every sampling time, then iteratively data from two randomly selected wells were removed from each sampling event ($\sim 5\%$ of data) in a stratified approach. This corresponds to an engineer collecting two fewer samples at each sampling event. The second, referred to as 'well removal', again started with a full design and involved iteratively removing all data associated with two wells. At each round of data removals models (spatial splines and spatio-temporal p-splines) were built and subsequently predictions were made at each sampling time. Given that the interest here is in determining by how much the quantity of data can be reduced through predicting with a spatio-temporal model rather than a spatial model, only one of the spatial models (p-splines) was used for comparison. The mean square prediction error at time t was then calculated for each model using the formula detailed in Equation 3.2 and these MSPEs were then summed over all time points and averaged over the simulations to give mean total MSPEs for each stage of data removal, shown in Equation 3.3:

$$\text{MSPE} = \frac{1}{M}\sum_{m=1}^{M}\sum_{t=1}^{T}\text{MSPE}_{tm} \tag{3.3}$$

where M is the number of simulations and T is the number of time points for which predictions were made.

Figure 3.10 shows that as data are removed by the observation removal method, the performances of the spatial model deteriorates markedly in contrast with a much gentler rate of decline for the spatio-temporal model. The almost constant total MSPE over all stages of observation removals for the spatio-temporal model demonstrates the benefit of the spatio-temporal models ability to ‘borrow strength’ over time.

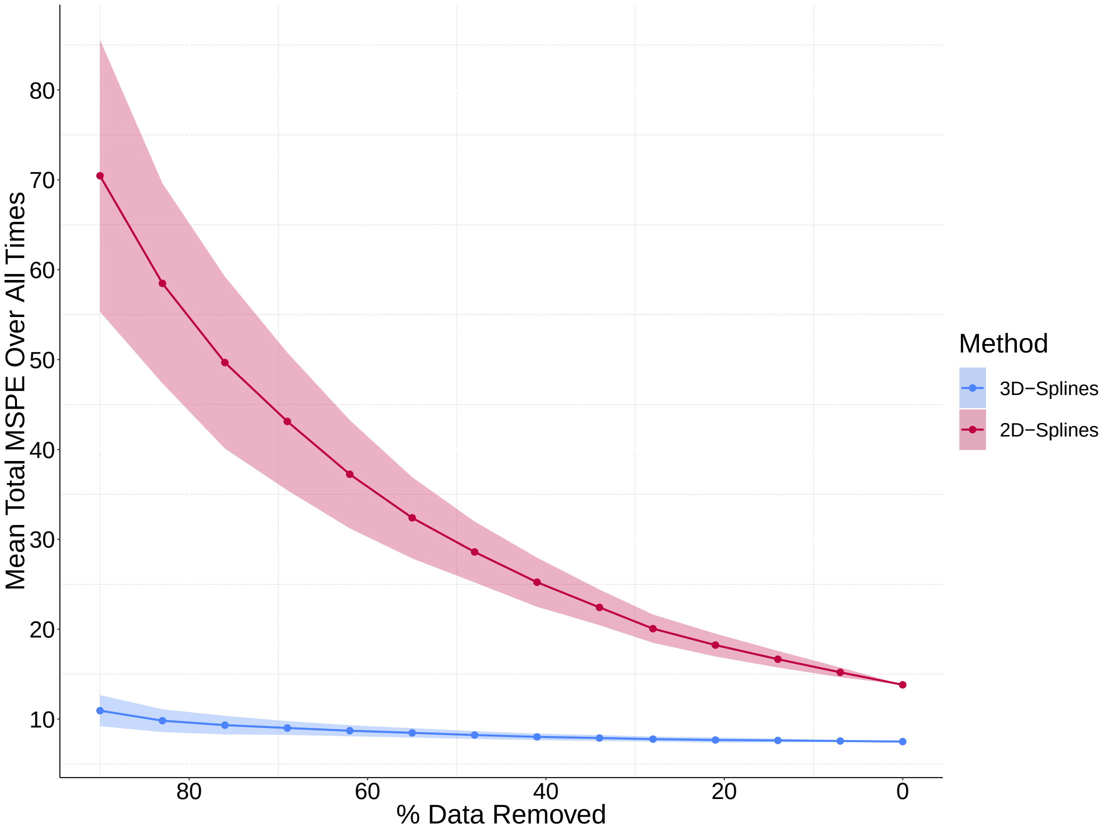

Figure 3.10: Mean total MSPEs over all sampling times for 200 simulations using the 'observation removal' method of data removal. Error bars indicate ±1 standard deviation.

Figure 3.11 shows results from the scenario where wells are removed rather than observations. This is a more challenging situation for the spatio-temporal model as removing the data from entire wells removes all knowledge of what is happening in these areas, and so the spatio-temporal model has no observations from which to 'borrow strength' over time. This is reflected in the MSPE results where the spatio-temporal model still delivers the best performance but the rate of deterioration with well removal, while slower, is more similar to those of the spatial model.

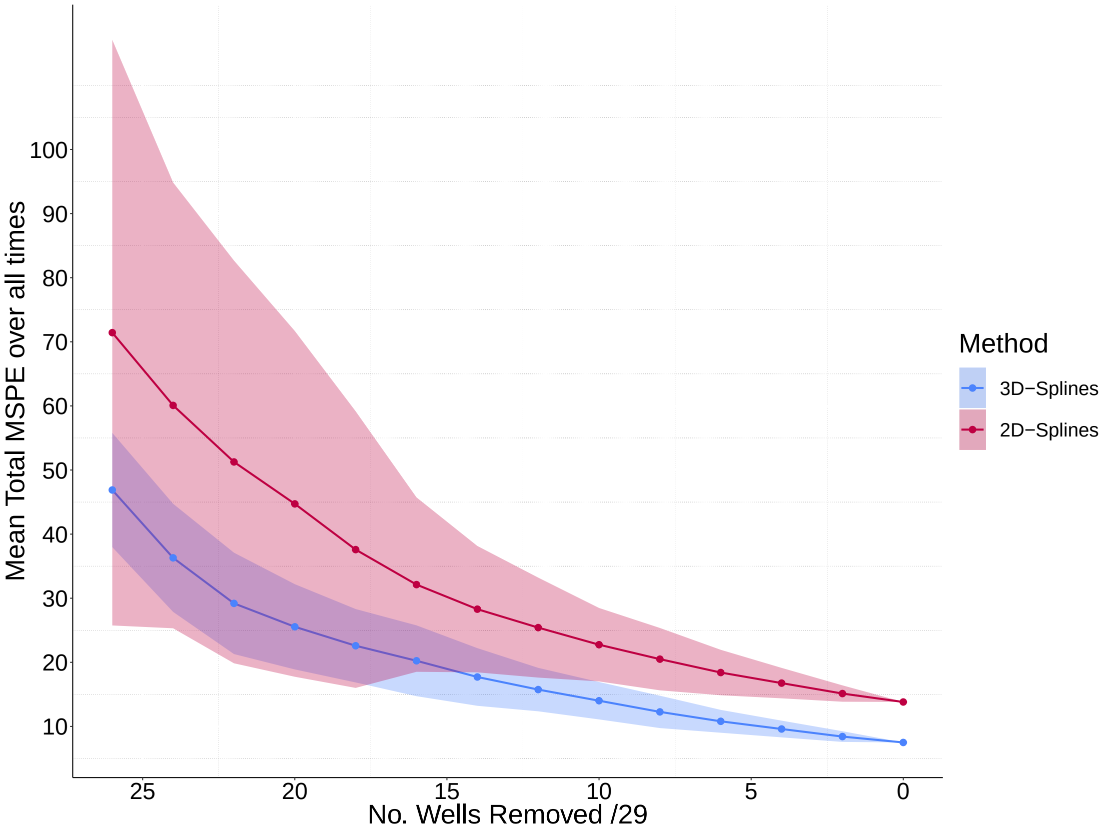

FIGURE 3.11: Mean total MSPEs over all sampling times for 200 simulations using the 'well removal' method of data removal. Error bars indicate ±1 standard deviation.

3.4 Real Application

The spatial and spatio-temporal p-spline models were used to analyse a dataset on a pollution incident at a refinery. MTBE (methyl tertiary butyl ether) is added to petrol to reduce noxious emissions as well as engine knocking. MTBE is no longer used at this site but it was in use at the time the data were collected. Due to MTBE having a high aqueous solubility and low retardation potential, its transit through groundwater is steady and degrades only under anaerobic conditions.

3.4.1 Results

For each of the p-spline models the degree of basis functions and order of penalty were kept the same as those in used in the simulation study. However, this time 18 basis

Table 3.4: Well-based cross validation scores for each prediction time for each spline model on the real dataset.

Modelling Method	Time			
	300	700	900	1300
Spatial P-splines	0.4142	2.4630	1.5438	4.1243
Spatio-temporal P-splines	0.8072	1.6380	1.8401	2.0396

functions were used for the easting component, 22 for northing and 14 for time. These numbers of basis functions were chosen to reflect the spatial aspect ratio of the study region and they took into account the fact that measurements were taken over a significantly longer period of time compared with the simulation study. As before, the MAP estimate of the smoothing parameter was used for both the spatial and spatio-temporal p-spline models.

Due to the sparsity of the data, each of the chosen times where predictions were made had only $\sim$ 10 or fewer observations, many of which were in very close vicinity. This is a very small number of observations from which to make reliable predictions and so, for the spatial p-splines model, observations within a 3-week time window were also included to improve stability.

Figure 3.12 shows the predicted surfaces from spatial and spatio-temporal p-spline models at 4 time points throughout the data. Table 3.4 shows 10-fold well-based cross-validation (CV) scores for each model at each prediction time shown in Figure 3.12. The shape and direction of the predicted contaminant plume is consistent with the south-east/north-west gradient of the groundwater flow for both p-spline models. Looking at the first time point, the spatio-temporal p-splines model is able to detect the release of the contaminant in the south east corner, while this is not the case for the spatial model due to a lack of samples in this region. Following the location of the leak being identified, both models are able to track the depletion of the plume. The spatio-temporal model provides a more definitive plume shape in comparison with the spatial model. The spatial model performs particularly poorly at time 1300 days. Interpretation of the plot suggests this is likely to be due to the model over-smoothing.

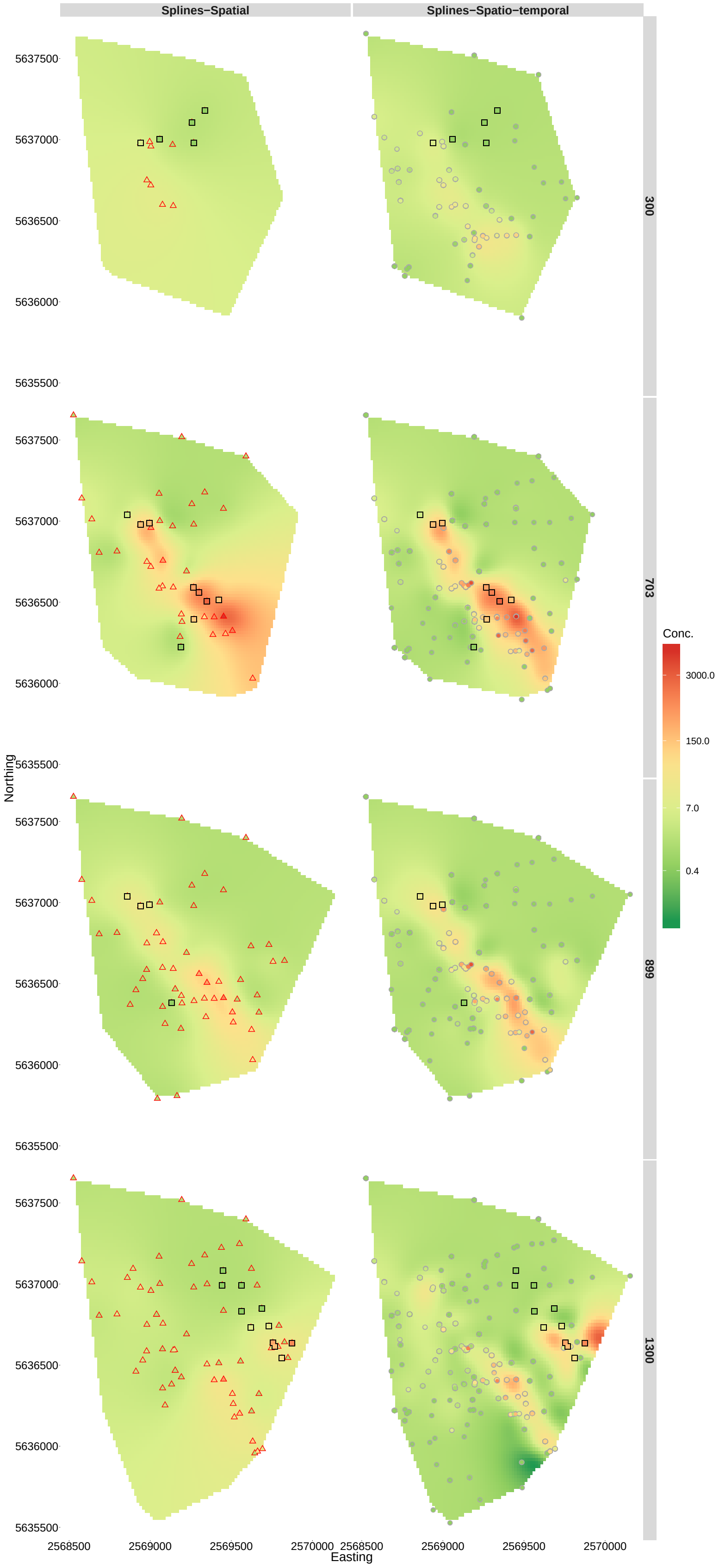

FIGURE 3.12: Predicted surfaces at four time points (300, 703, 899 & 1300 days) for each spline model on the real dataset. Filled squares indicate wells that were sampled at the prediction time and their recorded concentration; filled red triangles indicate observations that fall within the 21 day time window to be used in the spatial model and grey circles indicate the last recorded concentration at wells which were not sampled at the prediction time.

3.5 Summary

In conclusion, the results show that in general there is an added benefit in using a spatio-temporal model which borrows strength over time as opposed to the currently more frequently used spatial methods which treat time independently. The prediction surfaces for varying time points when using an incomplete sampling design (Figures 3.3, 3.4, 3.5) highlighted the ability of a spatio-temporal model to use earlier sampling information to improve its predictions. However, Figure 3.3 also illustrated that even the spatio-temporal model cannot predict what is going on in a region where there are no data available. The spatio-temporal splines model using (14, 8, 3) basis functions highlighted an effect known as ballooning in some of its predictions. Studies in Chapter 5 suggest that increasing the number of basis functions helps to prevent this from happening and this was also apparent when comparing the MSPEs for the two spatio-temporal models under sampling scenario 2.

The studies looking into data removal, presented in Section 3.3, clearly demonstrate the potential cost savings that can be made by reducing the number of samples and adopting a spatio-temporal model. This study also highlighted that the reduction in data very much depends on how data are removed. A spatio-temporal model is more advantageous if observations spread out across the study region are removed rather than entire sets of observations from specific wells. To achieve the equivalent accuracy with a spatial model, the network needs to be sampled much more extensively. It is worth noting however, that the computational effort required in a spatio-temporal model is significantly greater than that of a spatial model. To avoid ballooning being triggered in the spatio-temporal model, a large number of basis functions needs to be used and as the number of basis functions for each component increases, the computational time also increases exponentially.

There are several other methods that can be used for modelling spatio-temporal data such as Kriging. Evidence of ballooning has also been witnessed in spatio-temporal Kriging models when a Matérn covariance structure is used. One benefit to modelling with a spline-based model over Kriging is that there is no assumption of stationarity and isotropy.

Chapter 4

Incorporating an Additional Smoothing Parameter for the Temporal Component

4.1 Motivation and Model Formulation

Data obtained from several groundwater sites indicated that contaminant concentrations varied more across space than they did over time, highlighting the need for a model which controls the smoothness over space and time separately. In the case of a p-splines model smoothness across space and time could be controlled independently with separate smoothing parameters for each component. The single smoothing parameter spatio-temporal p-splines model used in Chapter 3, which was developed by Evers et al. [2015], utilises efficient linear algebra to obtain the optimal smoothing parameter. Unfortunately however, their method can only be used to efficiently optimise one smoothing parameter and not the desired two parameters being suggested. To compensate for this restriction, they suggest scaling the number of basis functions in each component to emulate the 'wigglyness' of the function i.e. the time component is given a smaller number of basis functions to reflect the fact that the contaminant concentrations vary more over space than they do over time. However there is no detailed criteria on how the numbers of basis functions should be assigned. Given that space and time are measured on different scales deciding on a suitable set of rules for the number of basis functions in each dimension is subjective and complex and thus, a model with two smoothing parameters seems more appropriate.

Tuning a single smoothing parameter is relatively fast when working with one dimensional data, and many steps have been taken to improve the computational speed further when working with higher dimensioned data. However, as the number of smoothing parameters being used increases, obtaining the optimal combination requires a computational efforts which very quickly become unmanageable. The optimisation of multiple smoothing parameters has been discussed by many authors. To improve computational speed, Wood [2000] propose a general multiple smoothing parameter selection method based on minimising the Generalised Cross Validation (GCV) score for Generalised Additive Models (GAMs). Computing the GCV score many times is very time consuming due to the trace of the smoothing matrix being required in the denominator, see Equation 2.37. To improve the computational speed, matrix decompositions and transformations are used. This methodology is built on by Wood [2004] who deal with the numerical instabilities caused by rank deficiency in the original methodology. This new method provides numerical robustness by allowing for a fixed penalty term i.e. a ridge penalty. They use pivoted QR or singular value decompositions (SVD) of the smoothing matrix, to improve the computational time of the GCV score. They then went on to publish a methodology aimed at semiparametric Generalised Linear Models (GLMs) (Wood [2011]). A restricted maximum likelihood (REML) or maximum likelihood (ML) method is adopted, with optimisation through a Newton method. This new method does not suffer from the occasional under-smoothing experienced by GCV and AIC.

Here a method is proposed for tuning two smoothing parameters in a tensor product p-splines model. This builds on the methodology set out by Evers et al. [2015], detailed in Section 2.3.2. There is potential to consider a third smoothing parameter to allow for differing levels of smoothness in the northing and easting components. However, since each spatial component is measured on the same scale, this can more easily be accounted for by scaling the numbers of basis functions by the length of the study region in each direction.

4.1.1 Formulation

The penalised least squares expression to be minimised for a spatio-temporal p-splines model with two smoothing parameters can be formulated as:

$$\text{PLS}(\boldsymbol{\alpha}) = (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha})^\top (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) + \lambda(\boldsymbol{\alpha}^\top \mathbf{P}_1^{\text{F}} \boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbf{P}_2^{\text{F}} \boldsymbol{\alpha} + \lambda_{rel} \boldsymbol{\alpha}^\top \mathbf{P}_3^{\text{F}} \boldsymbol{\alpha}) \tag{4.1}$$

with

$$\begin{aligned}\mathbf{P}_1^{\text{F}} &= \mathbf{P}_1 \otimes \mathbf{I}_2 \otimes \mathbf{I}_3\\ \mathbf{P}_2^{\text{F}} &= \mathbf{I}_1 \otimes \mathbf{P}_2 \otimes \mathbf{I}_3\\ \mathbf{P}_3^{\text{F}} &= \mathbf{I}_1 \otimes \mathbf{I}_2 \otimes \mathbf{P}_3.\end{aligned}$$

Kronecker products of the difference penalty matrices, $\mathbf{P}_1 = \mathbf{D}_{s_1}^\top \mathbf{D}_{s_1}$, $\mathbf{P}_2 = \mathbf{D}_{s_2}^\top \mathbf{D}_{s_2}$ and $\mathbf{P}_3 = \mathbf{D}_t^\top \mathbf{D}_t$, and identity matrices, of dimension equivalent to the number of basis functions in each dimension, create the required penalty structure for the spatio-temporal data. Here λ is the overall smoothing parameter and λ_{rel} is the scaling factor of λ for the temporal component.

The criterion, adopted by Evers et al. [2015], for determining the optimal single smoothing parameter, λ, can be utilised in the two smoothing parameter model. The posterior distribution for λ, detailed in Equation 2.49, is adapted for λ and λ_{rel} to become

$$f_{M_\lambda|\mathbf{y}} \propto \lambda^{\frac{rank(\mathbf{D}^\top\mathbf{D})}{2}} \times \frac{\Gamma(a^*)|\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{P}|^{-1/2}|\mathbf{P}|^{1/2}}{b + \frac{1}{2}\mathbf{y}^\top \left[\mathbf{I}_n - \mathbf{B}(\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{P})^{-1}\mathbf{B}^\top\right]\mathbf{y}} f_{M_\lambda} \tag{4.2}$$

where

$$\mathbf{P} = \mathbf{P}_1^{\text{F}} + \mathbf{P}_2^{\text{F}} + \lambda_{rel}\mathbf{P}_3^{\text{F}}. \tag{4.3}$$

Optimising λ_{rel} is computationally expensive since it is embedded within the penalty matrix $\mathbf{P}$. Conversely, the linear algebra methods used in the model proposed by Evers et al. [2015] allow λ to be optimised efficiently through expressing Equation 4.2 in a manner that depends only on λ, through the inverse and the determinant of an inverse of a diagonal matrix. This formulation is given in detail in Equation 2.59.

This method, aided by a grid search, can be used to obtain the combination of λ and λ_{rel} that maximises the log posterior distribution given in Equation 4.2. Optimisation this way, for k candidate values of λ_{rel} and l candidate values of λ, is of complexity $\mathcal{O}(r \times (f^3 + l \times f))$. Here $f = \prod_{k=1}^{3} m_k$ where m_k is the number of 1-dimensional basis functions for component k. This is already a reduction on the naive computational complexity of $\mathcal{O}(r \times l \times f^3)$. However, as the number of basis functions in each direction and the number of candidate values of λ_{rel} increases, the increase in computational expense of optimising both smoothing parameters is cubic. This is highlighted in Figure 4.1, which shows the total optimisation time via a grid search for increasing numbers of basis functions and 30 candidate values of λ_{rel}.

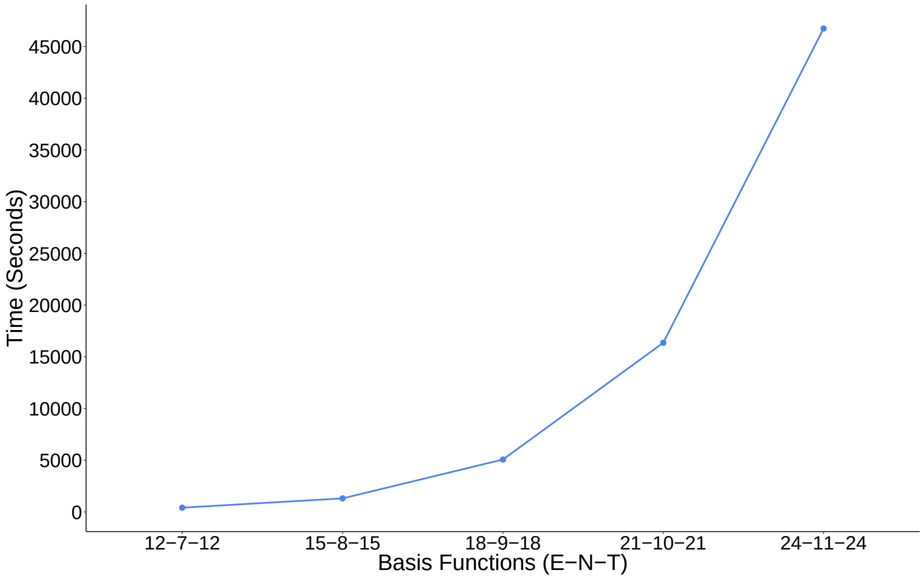

Figure 4.1: Two smoothing parameter grid search optimisation times (in seconds) for spatio-temporal models with increasing numbers of basis functions for a dataset with 4843 observations. For this plot the number of candidate values of λ_{rel} is 30.

To avoid the need for an expensive grid search, where both parameters are tuned together, a new formulation of the PLS expression (Equation 4.1) can be used to allow each parameter to be tuned separately using the methodology set out by Evers et al. [2015], this is detailed later in Section 4.3.2. However, for this to work effectively the surface being optimised over needs to have contours that are parallel to one of the axes. This allows the identification of starting points for each parameter that would not result in the optimisation getting ‘stuck’ in a contour. The following section seeks to determine whether the log posterior surface of the two smoothing parameters exhibits the necessary parallel contours to allow the parameters to be tuned separately. This is done by performing grid searches on several simulated and real datasets to find the shape of the log posterior surface. The results of this simulation study are then used to develop a more efficient optimisation procedure for the two smoothing parameters.

4.2 Study of Simulated and Real Datasets

Grid searches over combinations of λ and λ_{rel} were performed on simulated and real life datasets to determine the shape of the log posterior surfaces and also to try to detect any trends in the optimal combination of λ and λ_{rel} for varying numbers of basis functions.

Three simulated datasets were created based on four normal densities plotted at different spatial locations with temporal sine curves oscillating each of these densities separately. Table 4.1 contains the different combinations of standard deviation and frequency that were used for the densities and sine waves respectively. Figure 4.2 shows the simulated datasets at one time point.

Table 4.1: Combinations of standard deviations of the four normal densities plotted in space and the corresponding frequencies of their oscillations over time for the simulated data.

Simulation	Standard Deviation	Frequency
1	4, 2, 0.5, 4	0.4, 0.4, 0.4, 0.4
2	4, 4, 4, 4	0.2, 0.4, 0.2, 0.4
3	4, 2, 0.5, 4	0.2, 0.4, 0.2, 0.4

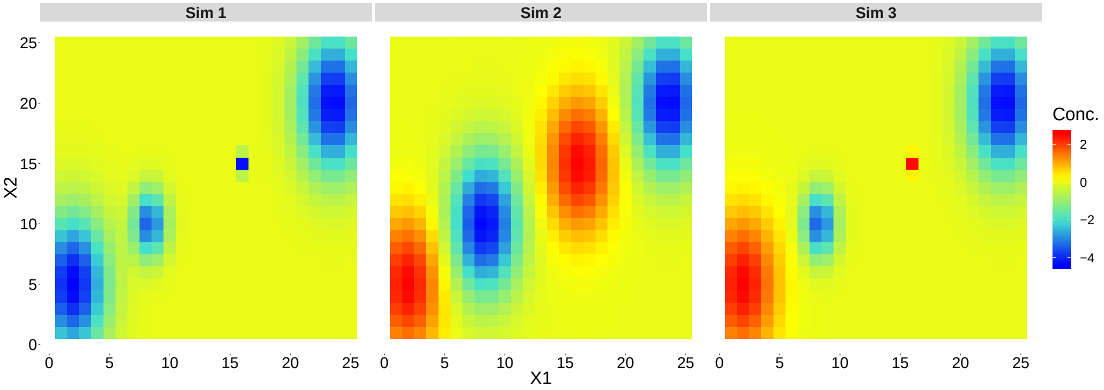

FIGURE 4.2: Three simulated datasets made up of varying spatial normal densities oscillating at different temporal frequencies. The parameters used to simulate these datasets are shown in Table 4.1

Figures 4.3, 4.4 and 4.5 show the log posterior surfaces for each simulated dataset over a 80 × 30 grid of values of λ and λ_{rel}. In Figures 4.6 and 4.7 the log posterior surfaces for data simulated from the PDE used in Chapter 3 are shown, whilst Figures 4.8 and 4.9 are the log posterior surfaces for two real datasets obtained from two different sites. The black arrows track the optimal combination of parameters as the number of basis functions is increased.

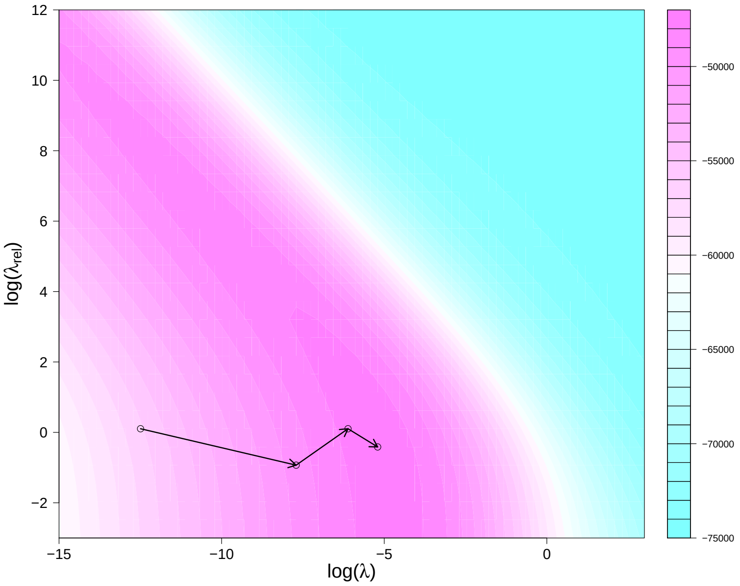

FIGURE 4.3: Log posterior surface of the two smoothing parameters for the first simulated dataset using 15 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting arrows show the evolving maximum log posterior value as the number of basis functions in each direction increases from 6 - 9 - 12 - 15.

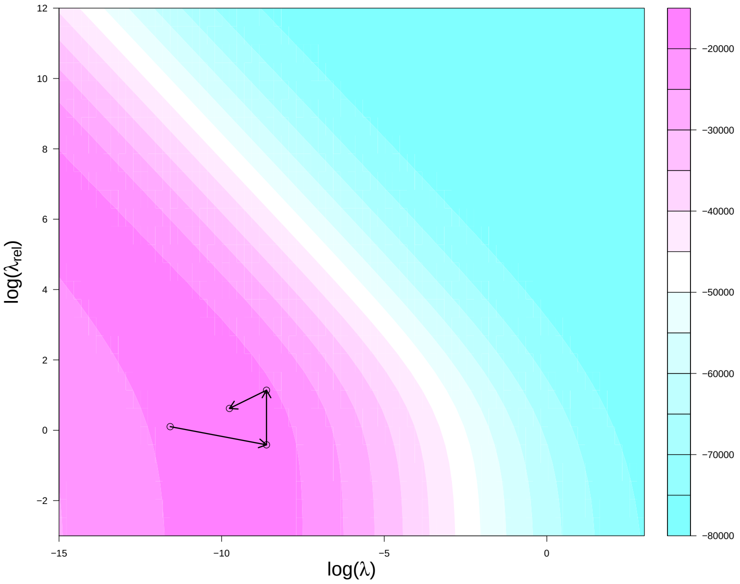

FIGURE 4.4: Log posterior surface of the two smoothing parameters for the second simulated dataset using 15 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting arrows show the evolving maximum log posterior value as the number of basis functions in each direction increases from 6 - 9 - 12 - 15.

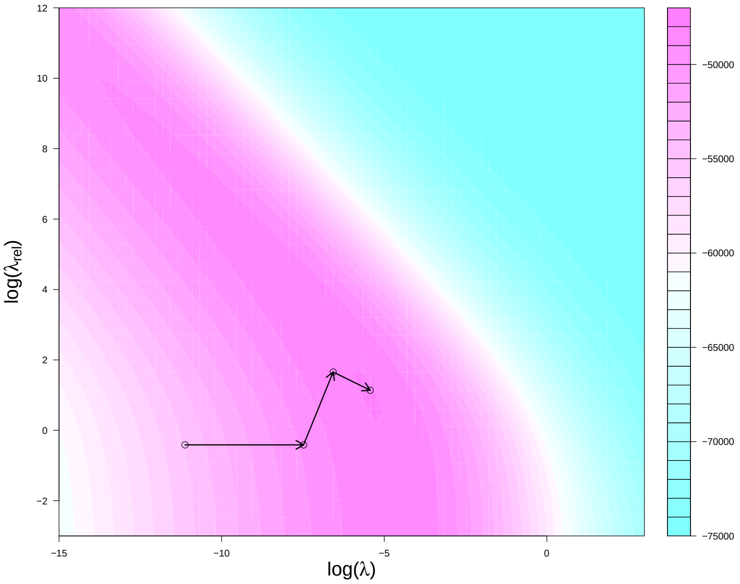

FIGURE 4.5: Log posterior surface of the two smoothing parameters for the third simulated dataset using 15 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting line show the evolving maximum log posterior value as the number of basis functions in each direction increases from 6 - 9 - 12 - 15.

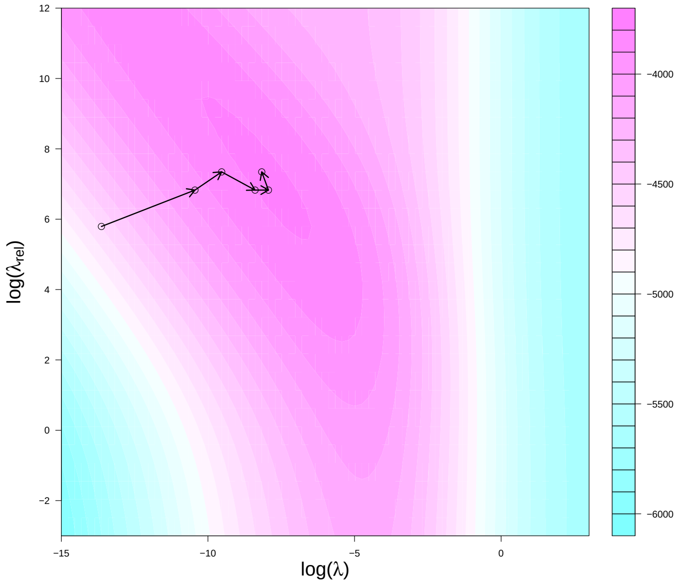

Figure 4.6: Log posterior surface of the two smoothing parameters for a dataset simulated from PDE1 with a realistic design, using 24 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting arrows show the evolving maximum log posterior value as the number of basis functions in each direction increases from 9 - 12 - 15 - 18 - 21 - 24.

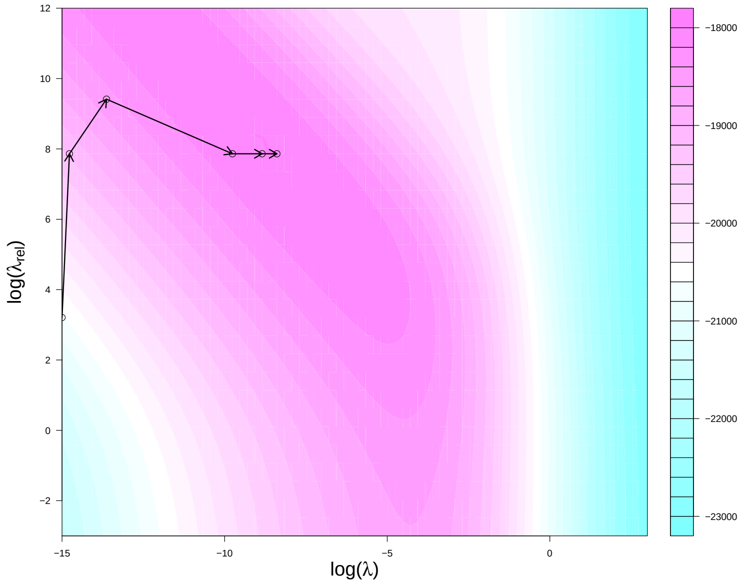

Figure 4.7: Log posterior surface of the two smoothing parameters, for a dataset simulated from PDE1 with a full design, using 24 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting arrows show the evolving maximum log posterior value as the number of basis functions in each direction increases from 9 - 12 - 15 - 18 - 21 - 24.

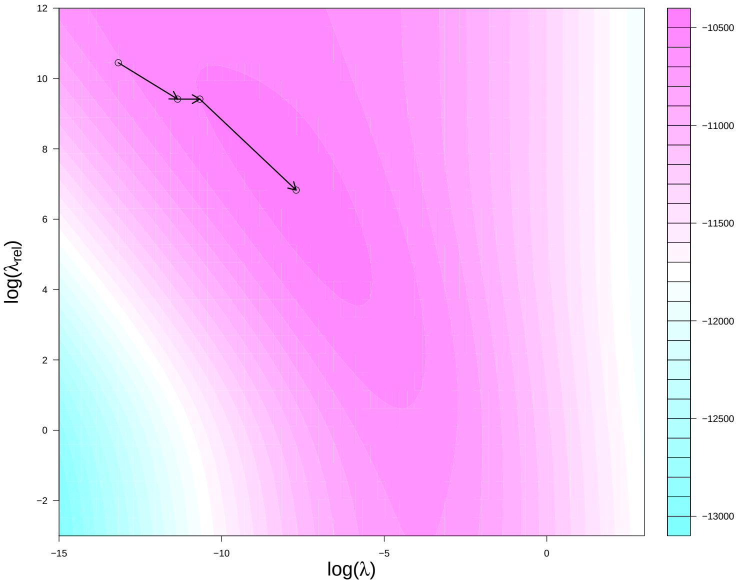

FIGURE 4.8: Log posterior surface of the two smoothing parameters for the first real dataset using 15 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting arrows show the evolving maximum log posterior value as the number of basis functions in each direction increases from 6 - 9 - 12 - 15.

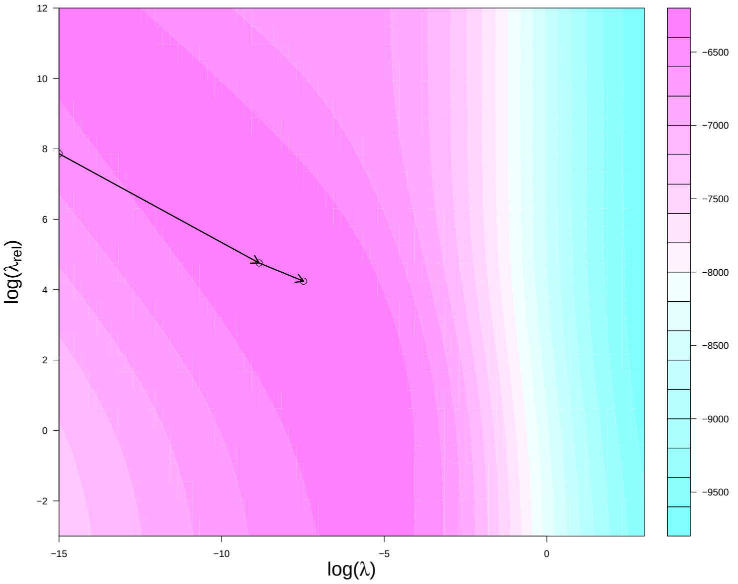

FIGURE 4.9: Log posterior surface of the two smoothing parameters for the second real dataset using 12 basis functions in each direction over a dense 80 × 30 grid. The black points and connecting arrows show the evolving maximum log posterior value as the number of basis functions in each direction increases from 6 - 9 - 12.

In order to employ the methodology previously described, using the method of Evers et al. [2015], each parameter needs to be tuned separately whilst fixing the other. To achieve this, the contours of the log posterior surface need to lie parallel to one of the axes. This makes the optimisation less dependent on the starting point and also helps to prevent the optimisation becoming 'stuck' in a contour. From the figures depicted above it is apparent that the contours generally do not follow the desired shape and are instead curved along the main diagonal, making it difficult to tune each parameter individually.

However, there does appear to be a trend present when looking at the optimal locations as the number of basis functions increases. With the exception of the second simulated dataset, as the number of basis functions in each dimension increases the optimal value of λ changes substantially. On the contrary the optimal value of λ_{rel} does not change as substantially, particularly if the optimum for the model with the lowest number of basis functions is ignored. Therefore, it is proposed that this trend is exploited to give starting points for tuning the two smoothing parameters.

4.3 Algorithm for Optimising Two Smoothing Parameters

The study conducted in Section 4.2 highlighted that although λ was different for each combination of basis functions, λ_{rel} did not appear to change as substantially compared with λ as the number of basis functions in the model increased.

It was therefore proposed that initially a coarse grid search would be performed using a model with a lower number of basis functions to obtain approximate values of λ_{rel} and λ. The approximate value of λ_{rel} would then be used to tune λ for a model with the desired number of basis functions and similarly, the newly obtained optimal value of λ would be used to tune λ_{rel} for the desired model. This allowed λ to be tuned efficiently as before. However, λ_{rel} can now also be tuned in the same way as λ using an augmented data notation. The adopted algorithm is detailed in the following section.

4.3.1 Algorithm

1. Determine each dimensions reduced number of basis functions, $\widetilde{m}_k$, for the grid search, by scaling the desired number of basis functions by γ

$$\widetilde{m}_k = \gamma \cdot m_k \qquad k = 1, 2, 3.$$

2. Use this lower number of basis functions to approximately optimise λ and λ_{rel} on a $q \times q$ grid, where $q \sim 10$. (Evaluation at different values of λ are not computationally costly so the number of these can be increased if required.)

3. Using the optimised values of λ and λ_{rel} and the original number of basis functions (m_1, m_2, m_3):

   3.1 Tune λ whilst fixing λ_{rel} to the value from step 2

   3.2 Tune λ_{rel} using a new formulation of the PLS expression (Equation 4.10), detailed in Section 4.3.2, and fixing λ to the value from step 3.1

Choosing the Reduced Number of Basis Functions for Step 1

To decide on an appropriate reduced number of basis functions for the initial tuning step, an effort reduction κ, for the expensive inversion of the matrix $(\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{P})$ needs to be chosen, e.g. $\kappa = 1/10$.

To begin with, the estimated computational effort for a single inversion of the matrix $(\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{P})$, given there are m_k basis functions for each of the k components, is of order

$$\left(\prod_{k=1}^{3} m_k\right)^3. \tag{4.4}$$

This formula stems from the fact that matrix inversions are of complexity $\mathcal{O}(f^3)$, where f is the dimension of the matrix. Here $f = \prod_{k=1}^{3} m_k$ since three dimensional tensor product basis functions are being used.

The scaling factor γ, which controls the reduction in the number of basis functions, depends on the desired effort reduction κ.

To achieve the chosen effort reduction κ, the scaling factor γ needs to be chosen such that:

$$\left(\prod_{k=1}^{3} \gamma \cdot m_k\right)^3 = \left(\prod_{k=1}^{3} m_k\right)^3 \cdot \kappa. \tag{4.5}$$

Thus the scaling factor for the number of basis functions is

$$\gamma = \sqrt[9]{\kappa}. \tag{4.6}$$

Example

To reduce the computation time on a model with $m_k = 18$ to a $1/10^{th}$ of the effort, the number of basis functions would need to be scaled by:

$$\gamma = \sqrt[9]{\kappa} = \sqrt[9]{1/10} = 0.77$$

for the initial tuning step i.e. $\widetilde{m}_k = 18 \times 0.77 \simeq 14$.

4.3.2 Augmented Data Formulation for Optimising λ_{rel}

An augmented data notation can be used to allow the penalised least squares expression for two smoothing parameters (Equation 4.1) to be re-expressed with λ_{rel} in the position of λ. To obtain the augmented data, Equation 4.1 can be alternatively denoted as:

$$(\mathbf{y} - \mathbf{B}\boldsymbol{\alpha})^\top (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) + (\mathbf{0} - \mathbf{D}_s\boldsymbol{\alpha}\sqrt{\lambda})^\top(\mathbf{0} - \mathbf{D}_s\boldsymbol{\alpha}\sqrt{\lambda}) + \lambda\lambda_{rel}\boldsymbol{\alpha}^\top\mathbf{D}_t^\top\mathbf{D}_t\boldsymbol{\alpha}. \tag{4.7}$$

which is equivalent to

$$\left(\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{B}\\ \sqrt{\lambda}\mathbf{D}_s\end{bmatrix}\boldsymbol{\alpha}\right)^\top \left(\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{B}\\ \sqrt{\lambda}\mathbf{D}_s\end{bmatrix}\boldsymbol{\alpha}\right) + \lambda\lambda_{rel}\boldsymbol{\alpha}^\top\mathbf{D}_t^\top\mathbf{D}_t\boldsymbol{\alpha}. \tag{4.8}$$

By denoting

$$\widetilde{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} \qquad \widetilde{\mathbf{B}} = \begin{bmatrix} \mathbf{B} \\ \sqrt{\lambda}\mathbf{D}_s, \end{bmatrix} \tag{4.9}$$

Equation 4.1 can be written as

$$\left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{B}}\boldsymbol{\alpha}\right)^\top \left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{B}}\boldsymbol{\alpha}\right) + \lambda_{rel}\boldsymbol{\alpha}^\top(\sqrt{\lambda}\mathbf{D}_t)^\top(\sqrt{\lambda}\mathbf{D}_t)\boldsymbol{\alpha}. \tag{4.10}$$

Given λ is fixed in step 3.2 of the algorithm, this alternative formulation of the PLS expression, shown in Equation 4.10, allows λ_{rel} to be tuned using the methodology set out for tuning λ.

4.4 Improving Computational Speed

The formulation of $f_{M_\lambda|\mathbf{y}}$ used by Evers et al. [2015], detailed in Equation 2.49, differs slightly from the formulation presented at the beginning of this chapter in Equation 4.2. Notice the addition of $|\mathbf{P}|^{1/2} = |\mathbf{D}^\top\mathbf{D}|^{1/2}$ in the numerator. In Evers' model, $|\mathbf{D}^\top\mathbf{D}|^{1/2}$, contained within $|\mathbf{V}(\lambda)|^{-1/2} = |\lambda\mathbf{D}^\top\mathbf{D}|^{1/2}$, was removed when specifying the expression proportional to $f_{M_\lambda|\mathbf{y}}$ since it remained constant with changing λ. In the case of the model with two smoothing parameters, $\mathbf{D}^\top\mathbf{D}$, now expressed as $\mathbf{P}$, does not remain constant for differing values of λ_{rel} and thus $\mathbf{P}$ and its determinant needs to be computed for every value of λ_{rel} that is considered. For step 1 of the algorithm, when many values of λ_{rel} are being considered, the computational expense due to this calculation increases rapidly. The matrix $\mathbf{P}$ and its determinant are also required to be calculated once for each substep of step 3 in the algorithm.

To overcome the computational complexity of calculating the determinant of the matrix, the penalty matrix $\mathbf{P}$ can be jointly diagonalised. By the theorem of spectral decomposition (Appendix B.1), $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ can be denoted as:

$$\mathbf{P}_1 = \boldsymbol{\Gamma}_1\boldsymbol{\Delta}_1\boldsymbol{\Gamma}_1^\top$$
$$\mathbf{P}_2 = \boldsymbol{\Gamma}_2\boldsymbol{\Delta}_2\boldsymbol{\Gamma}_2^\top$$
$$\mathbf{P}_3 = \boldsymbol{\Gamma}_3\boldsymbol{\Delta}_3\boldsymbol{\Gamma}_3^\top$$

where $\boldsymbol{\Gamma}_1$, $\boldsymbol{\Gamma}_2$ and $\boldsymbol{\Gamma}_3$ are orthogonal matrices and $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$ and $\boldsymbol{\Delta}_3$ are diagonal matrices. The full tensor product penalty can also be re-expressed using the aforementioned theorem as:

$$\begin{aligned}\mathbf{P} &= \mathbf{P}_1^{\mathrm{F}} + \mathbf{P}_2^{\mathrm{F}} + \lambda_{rel}\mathbf{P}_3^{\mathrm{F}} \\ &= (\mathbf{P}_1 \otimes \mathbf{I}_2 \otimes \mathbf{I}_3) + (\mathbf{I}_1 \otimes \mathbf{P}_2 \otimes \mathbf{I}_3) + \lambda_{rel}(\mathbf{I}_1 \otimes \mathbf{I}_2 \otimes \mathbf{P}_3) \\ &= \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}^\top.\end{aligned} \tag{4.11}$$

It can be shown that the orthogonal matrix, $\boldsymbol{\Gamma}$, obtained from the spectral decomposition of the full penalty is equivalent to the Kronecker product of $\boldsymbol{\Gamma}_1$, $\boldsymbol{\Gamma}_2$ and $\boldsymbol{\Gamma}_3$ i.e.

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_1 \otimes \boldsymbol{\Gamma}_2 \otimes \boldsymbol{\Gamma}_3. \tag{4.12}$$

$\boldsymbol{\Delta}$ can then be obtained by rearranging Equation 4.11:

$$\begin{aligned}\boldsymbol{\Delta} = \boldsymbol{\Gamma}^\top\mathbf{P}\boldsymbol{\Gamma} &= (\boldsymbol{\Gamma}_1^\top \otimes \boldsymbol{\Gamma}_2^\top \otimes \boldsymbol{\Gamma}_3^\top)\,[(\mathbf{P}_1 \otimes \mathbf{I}_2 \otimes \mathbf{I}_3) + (\mathbf{I}_1 \otimes \mathbf{P}_2 \otimes \mathbf{I}_3)\ + \\ &\qquad \lambda_{rel}(\mathbf{I}_1 \otimes \mathbf{I}_2 \otimes \mathbf{P}_3)]\,(\boldsymbol{\Gamma}_1 \otimes \boldsymbol{\Gamma}_2 \otimes \boldsymbol{\Gamma}_3) \\ &= \underbrace{(\boldsymbol{\Delta}_1 \otimes \mathbf{I}_2 \otimes \mathbf{I}_3)}_{\boldsymbol{\Delta}_1^F} + \underbrace{(\mathbf{I}_1 \otimes \boldsymbol{\Delta}_2 \otimes \mathbf{I}_3)}_{\boldsymbol{\Delta}_2^F} + \lambda_{rel}\underbrace{(\mathbf{I}_1 \otimes \mathbf{I}_2 \otimes \boldsymbol{\Delta}_3)}_{\boldsymbol{\Delta}_3^F}\end{aligned} \tag{4.13}$$

By definition:

$$|\mathbf{P}| = \prod_{i=1}^{n} \boldsymbol{\Delta}[i,i] \tag{4.14}$$

i.e. the determinant of $\mathbf{P}$ is equal to the product of the eigenvalues of $\mathbf{P}$.

However, since $\mathbf{P}$ is not of full rank it is also not invertible since its determinant is equal to 0. To by-pass this computational issue, a small ridge penalty, $\tau\mathbf{I}$, is added to $\boldsymbol{\Delta}$ where

$\mathbf{I}$ is the identity matrix and τ is a small value i.e. 1×10^{-10}. Adding this ridge penalty avoids zero eigenvalues in $\boldsymbol{\Delta}$ and so the determinant of $\mathbf{P}$ can be computed.

As $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$ and $\boldsymbol{\Delta}_3$ are of dimension, $m_k \times m_k$, where m_k is the number of 1-dimensional basis functions for component k, they are computationally inexpensive to obtain. $\boldsymbol{\Delta}_1^F$, $\boldsymbol{\Delta}_2^F$ and $\boldsymbol{\Delta}_3^F$ are each made up of the Kronecker product of a diagonal matrix with identity matrices and so they are also inexpensive to compute. This results in $\boldsymbol{\Delta}$ being obtained by spectral decomposition of 3 matrices of dimension $m_k \times m_k$, instead of one matrix of dimension $\left(\prod_{k=1}^3 m_k\right) \times \left(\prod_{k=1}^3 m_k\right)$ where m_k is the number of basis functions for dimension k.

Therefore, by adopting the proposed algorithm, the computational complexity of the optimisation of the two smoothing parameters is now

$$\mathcal{O}(\underbrace{q^2 \times \widetilde{f}^3}_{\text{Grid Search}} + \underbrace{(f^3 + (f \times l))}_{\text{Tuning } \lambda} + \underbrace{(f^3 + (f \times r))}_{\text{Tuning } \lambda_{rel}})$$

where q is the dimensions of the initial coarse grid search, $\widetilde{f} = \prod_{k=1}^3 \widetilde{m}_k$, where $\widetilde{m}_k$ is the reduced number of one dimensional basis functions for dimension k in step 1 of the algorithm, l is the number of candidate values of λ in step 3.1, r is the number of candidate values of λ_{rel} in step 3.2 and $f = \prod_{k=1}^3 m_k$ where m_k is the number of one-dimensional basis functions for dimension k.

4.5 Simulation Study

A study was conducted to compare the smoothing parameters obtained by the algorithm to the values obtained by the grid search. The predictive performance of models with these parameters were also compared. For this study, the data simulated for the comparison study in Chapter 3 were used along with a new dataset described in Section 4.5.1. For each dataset, two sampling scenarios were considered (a full design and a realistic design) and the smoothing parameters along with the MSPEs were compared for each method of obtaining the smoothing parameters. The ratio of the spatial dimensions were different for each dataset thus, for PDE1, 18 basis functions were used for the easting and time components with the northing component having its number of basis function scaled by the study regions dimensions. The model for PDE2 was similarly constructed using 15 basis functions for the easting and time components. A smaller number of functions was chosen due to the ratio of spatial dimensions being 1 and thus a larger number of basis functions would make the computational expense unmanageable.

4.5.1 Data Simulation

The second set of groundwater data were simulated from a variation of the PDE (Equation 3.1) used in Chapter 3. The second PDE is detailed below in Equation 4.15.

$$\frac{\partial y}{\partial t} = D \cdot \left(\frac{\partial^2 y}{\partial x_1^2} + \frac{\partial^2 y}{\partial x_2^2} + \frac{\partial^2 y}{\partial x_1 \partial x_2} \right) + \omega_1(x_1, x_2) \frac{\partial y}{\partial x_1} + \omega_2(x_1, x_2) \frac{\partial y}{\partial x_2} \tag{4.15}$$

where y are the contaminant concentrations, x_1 and x_2 are the spatial coordinates and $t \in [0, 1]$ denotes time. In the first term D is a constant controlling how quickly the solute spreads and is combined with the sum of the 2^{nd} partial derivatives to give a term which describes the spread by diffusion of the contaminant in the groundwater. The remaining two advection terms describe how the contaminant is affected by groundwater flow, where ω_1 and ω_2 describe its direction and velocity in each direction respectively. Theses functions were chosen based on observed groundwater levels at a current site.

Observed measurements were simulated by randomly generating a network made up of 22 monitoring wells (points in Figure 4.10) from a grid covering the vicinity of the contaminant plume. Measurement noise was added on the log scale to represent multiplicative error. The true concentrations (i.e. test data) were obtained by interpolating the numerical solution to the PDE, computed over a $100 \times 100 \times 200$ grid. Each panel of Figure 4.10 shows the spread and location of the contaminant plume at times $t \in \{0, 0.25, 0.75, 1\}$.

The datasets used for the study were made up of samples taken from 32 random sampling times. For the realistic design, 50% of the full set of observations (704 observations) were randomly removed to give a dataset containing 352 observations.

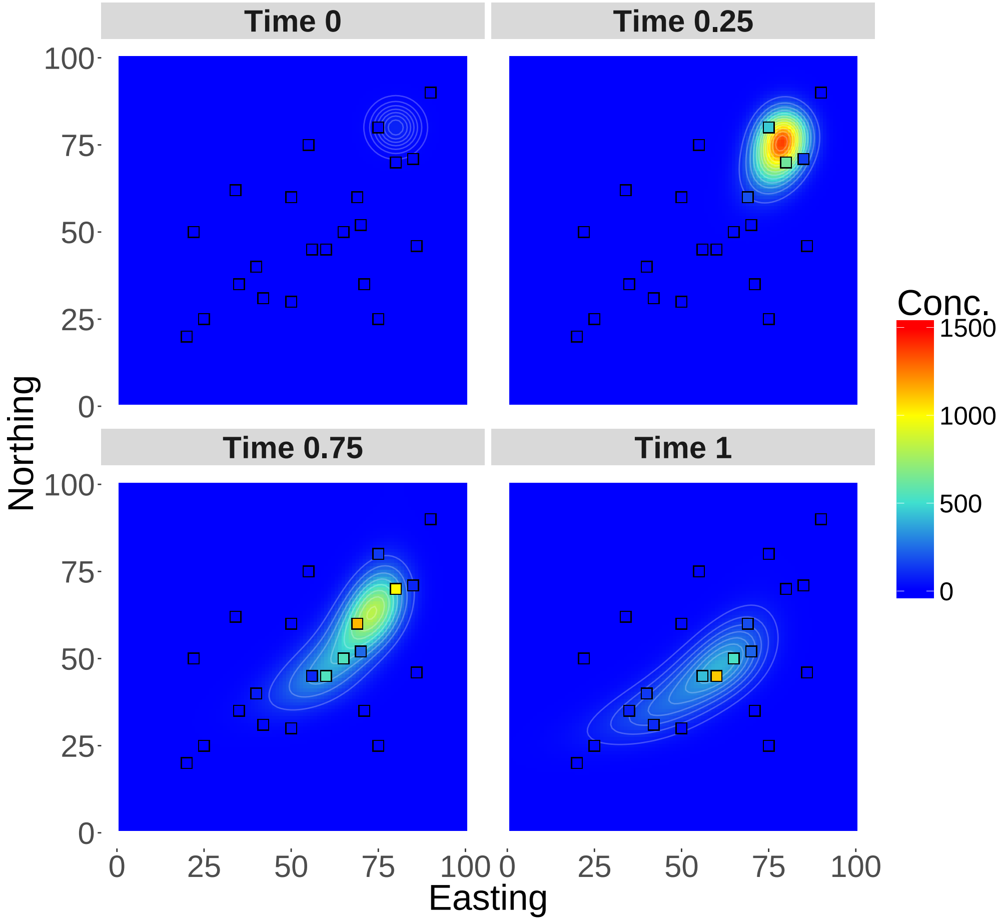

FIGURE 4.10: True underlying PDE described in Equation 4.15 (PDE2) at times $t \in \{0.25, 0.50, 0.75, 1\}$

4.5.2 Results

Table 4.2 shows the optimal values of λ and λ_{rel} obtained both by performing a grid search and by using the algorithm proposed in Section 4.3, for each dataset. It also shows the MSPEs for models fitted with these combinations of λ and λ_{rel}. These are the values used to produce the prediction surface shown in Figures 4.16, 4.18, 4.17 and 4.19.

Table 4.2: Optimal values of λ and λ_{rel} chosen by a classical grid search and the proposed algorithm along with MSPEs for predictions made with the resulting models for the PDE introduced in Chapter 3, described in Equation 3.1, (PDE1) and the PDE introduced in this chapter, described in Equation 4.15 (PDE2).

Data		Grid Search			Algorithm		
		λ	λ_{rel}	MSPE	λ	λ_{rel}	MSPE
PDE1	Real Design	2.2×10^{-5}	923	0.2524	1.8×10^{-4}	1168.3	0.2763
	Full Design	5.8×10^{-5}	2596.9	1.3637	1.1×10^{-4}	1168.3	0.6739
PDE2	Real Design	2.2×10^{-4}	41.4	0.2729	7.1×10^{-4}	10.1	0.1846
	Full Design	3.6×10^{-4}	24.7	0.3278	8.9×10^{-4}	10.1	0.2452

Table 4.2 shows that the value chosen for λ for each method of optimisation is very small, resulting in almost no penalty being applied to the spatial dimensions for both datasets and both designs. For λ_{rel} both optimisation methods select similar values for both designs and datasets, with the exception of the full design on the first dataset where the value of λ_{rel} is significantly larger for the grid search, which results in the temporal dimension being more heavily penalised.

Figures 4.11, 4.12, 4.13 and 4.14 assess how close the optimal values produced by the algorithm are to the optimal values derived from the log posterior surface through a grid search. From these figures, it is clear that the optimal locations differ slightly, with both methods selecting different points along the main ridge of the surface. Table 4.3 shows the value of the log posterior distribution using the optimal combination of smoothing parameters for each method, PDE and design. The values of the log posterior are not identical but they are very similar. An additional round of tuning could be incorporated to improve the match but given the very minor change this would give in the log posterior value in comparison to the very large computational expense, the current values seem satisfactory.

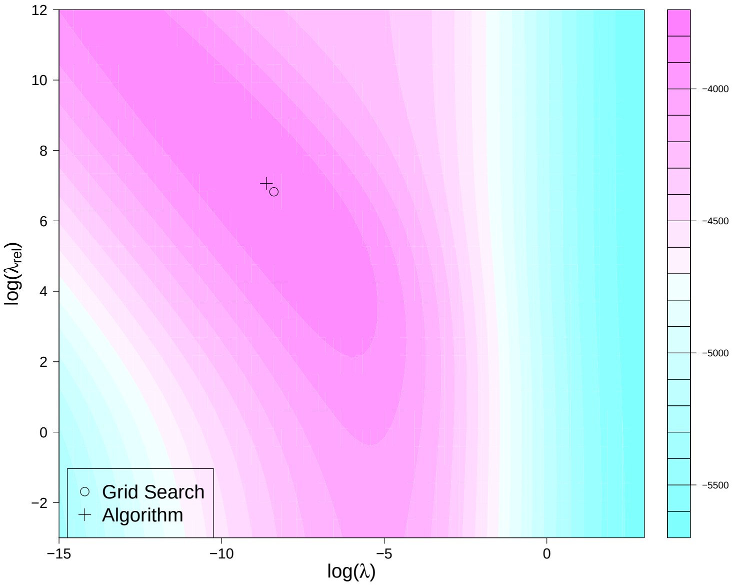

FIGURE 4.11: Log posterior surface of the two smoothing parameters for PDE1 with a realistic design over a dense 80 × 30 grid. The point indicates the location of the maximum i.e. the optimal combination, whilst the cross indicates the optimal combination obtained using the algorithm.

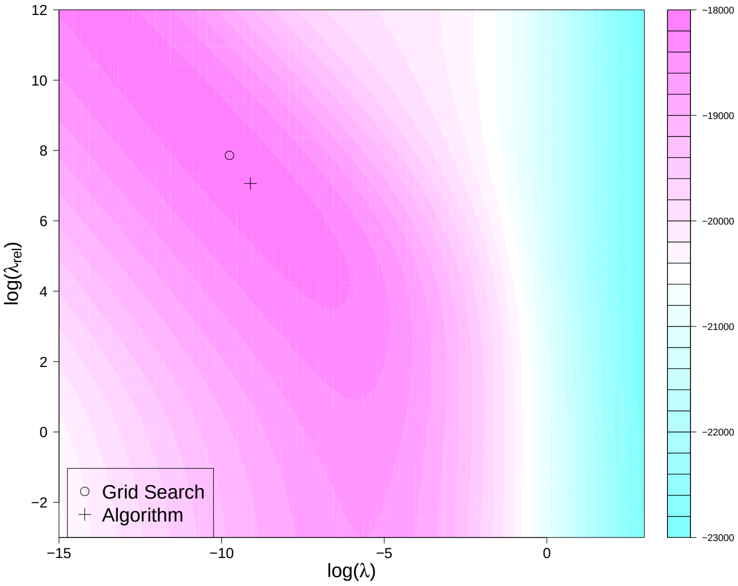

FIGURE 4.12: Log posterior surface of the two smoothing parameters for PDE1 with a full design over a dense 80 × 30 grid. The point indicates the location of the maximum i.e. the optimal combination, whilst the cross indicates the optimal combination obtained using the algorithm.

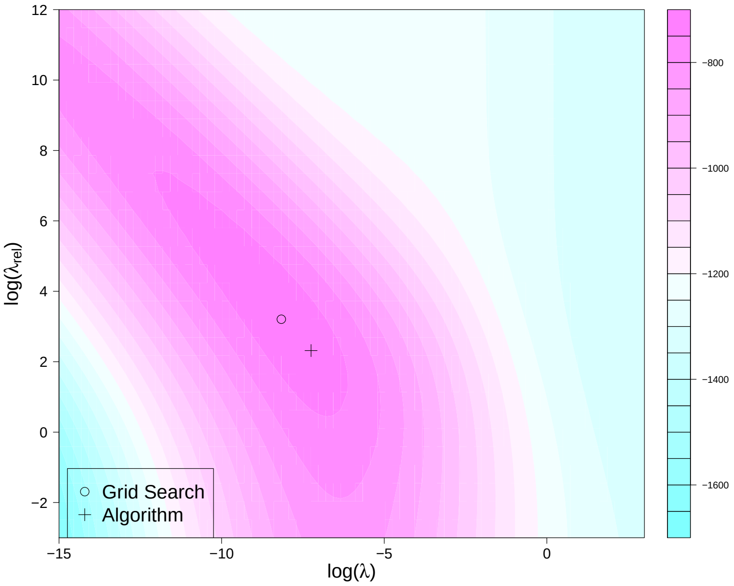

FIGURE 4.13: Log posterior surface of the two smoothing parameters for PDE2 with a realistic design over a dense 80 × 30 grid. The point indicates the location of the maximum i.e. the optimal combination, whilst the cross indicates the optimal combination obtained using the algorithm.

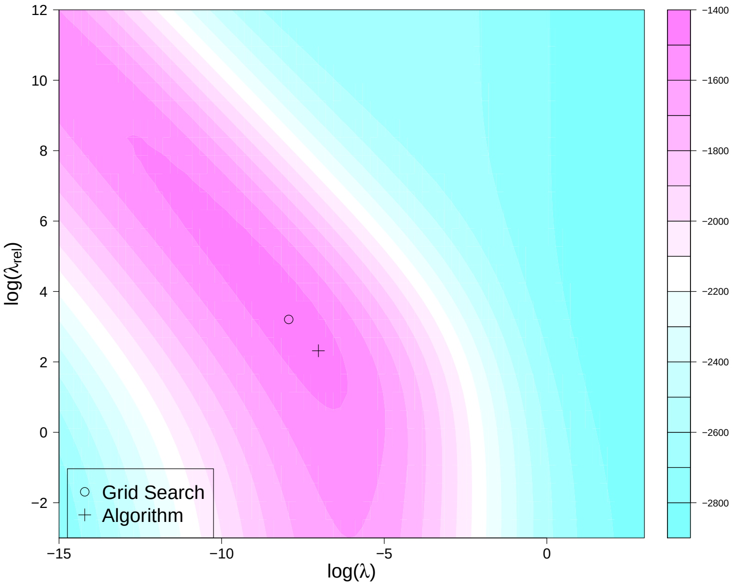

FIGURE 4.14: Log posterior surface of the two smoothing parameters for PDE2 with a full design over a dense 80 × 30 grid. The point indicates the location of the maximum i.e. the optimal combination, whilst the cross indicates the optimal combination obtained using the algorithm.

Table 4.3: Values of the maximum log posterior values for the optimal combinations of λ and λ_{rel} for each PDE, design and optimisation method.

Data		Grid Search	Algorithm
PDE1	Real Design	-3800	-3799.9
	Full Design	-18094.4	-18095.5
PDE2	Real Design	-717.9	-722.5
	Full Design	-1460	-1467

The predicted surfaces for the first PDE (Figures 4.16 and 4.17) show subtle differences due to the slight differences in the smoothing parameters chosen by each optimisation method. Ballooning, which was identified in the one smoothing parameter model in Chapter 3, is also present in the predictions of this two smoothing parameter model on the full design. This will be investigated further in Chapter 5. The balloons are less severe for the model chosen by the algorithm due to selection of a slightly higher value of λ which more heavily penalises the spatial component. The predicted surfaces for the second PDE are also very similar, with the algorithm predicting slightly smoother surfaces due to the selected value of λ being higher for both designs.

Figure 4.15 compares the optimisation times for the two smoothing parameter model using a grid search and the proposed algorithm for the full dataset from PDE1 which contains 4843 observations. From this plot it is clear that the proposed optimisation algorithm significantly reduces the optimisation time, by up to 90% compared with a grid search.

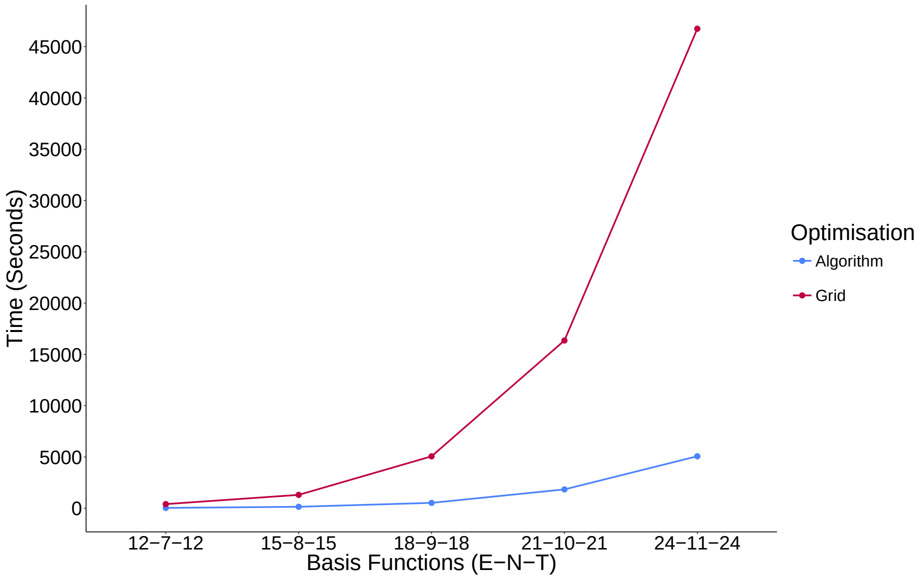

FIGURE 4.15: Comparison of optimisation times (in seconds) for tuning two smoothing parameters using a grid search and the proposed algorithm with increasing numbers of basis functions on a dataset containing $\sim$ 4800 observations.

4.6 Summary

A method for efficiently optimising two smoothing parameters in a spatio-temporal p-splines model has been proposed. The algorithm is able to reduce the computational time by up to 90% by exploiting trends in the optimal combination of smoothing parameters as the number of basis functions in the model increases. The algorithm does not obtain exactly the same values obtained by a grid search with the algorithm but the predictions are very similar and, given the large computational saving, this slight discrepancy is acceptable. To improve the accuracy, a further tuning step could be incorporated but, as previously mentioned, the computational expense of this would need to be weighed against the accuracy achieved.

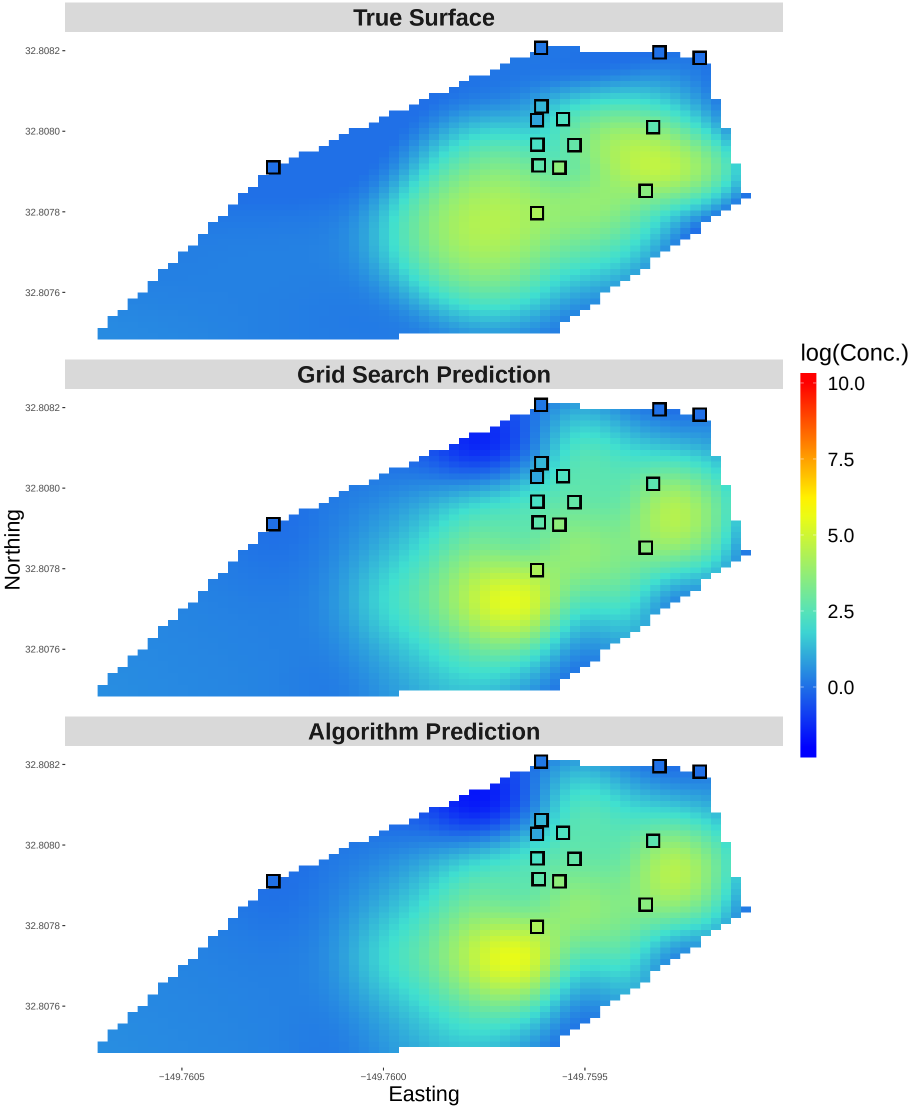

FIGURE 4.16: Predicted surfaces for spatio-temporal p-spline models built using the combination of λ and λ_{rel} chosen by a grid search and the proposed algorithm for PDE1 and a realistic design. The plot at the top is the true surface from PDE1.

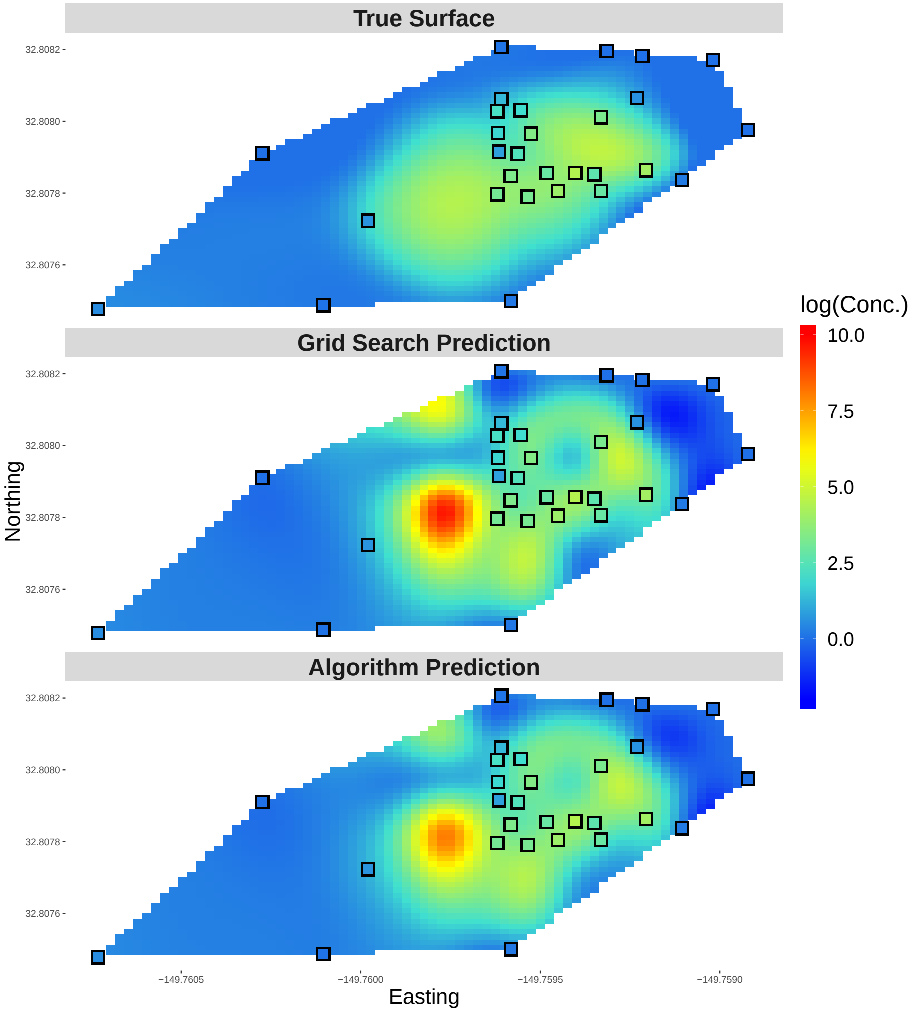

FIGURE 4.17: Predicted surfaces for spatio-temporal p-spline models built using the combination of λ and λ_{rel} chosen by a grid search and the proposed algorithm for PDE1 and a full design. The plot at the top is the true surface from PDE1. Ballooning can be observed in these predictions, this will be investigated in Chapter 5.

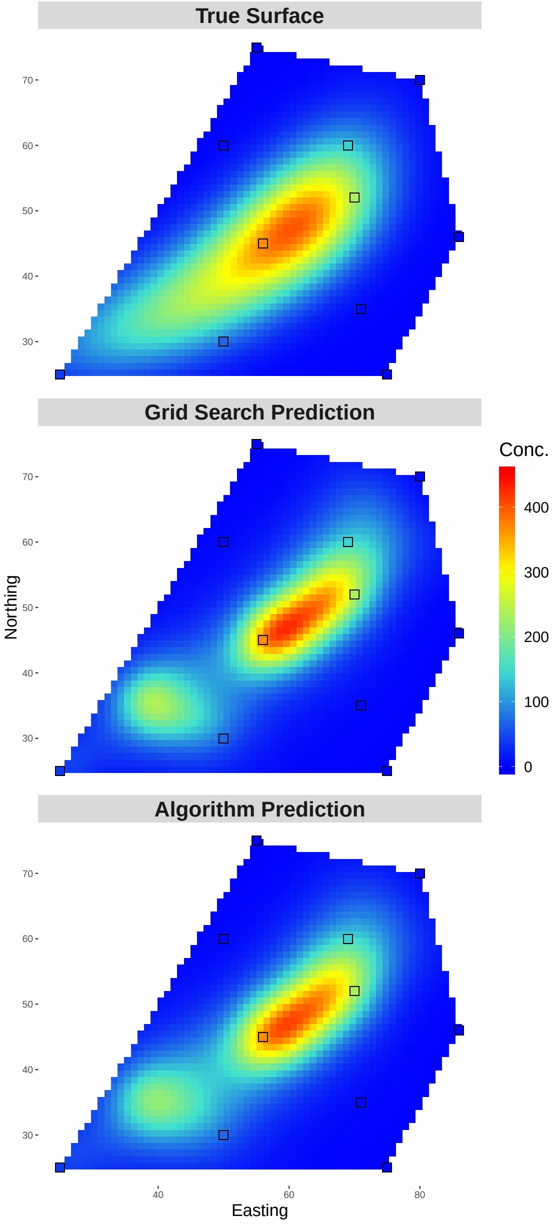

FIGURE 4.18: Predicted surfaces for spatio-temporal p-spline models built using the combination of λ and λ_{rel} chosen by a grid search and the proposed algorithm for PDE2 and a realistic design. The plot at the top is the true surface from PDE2.

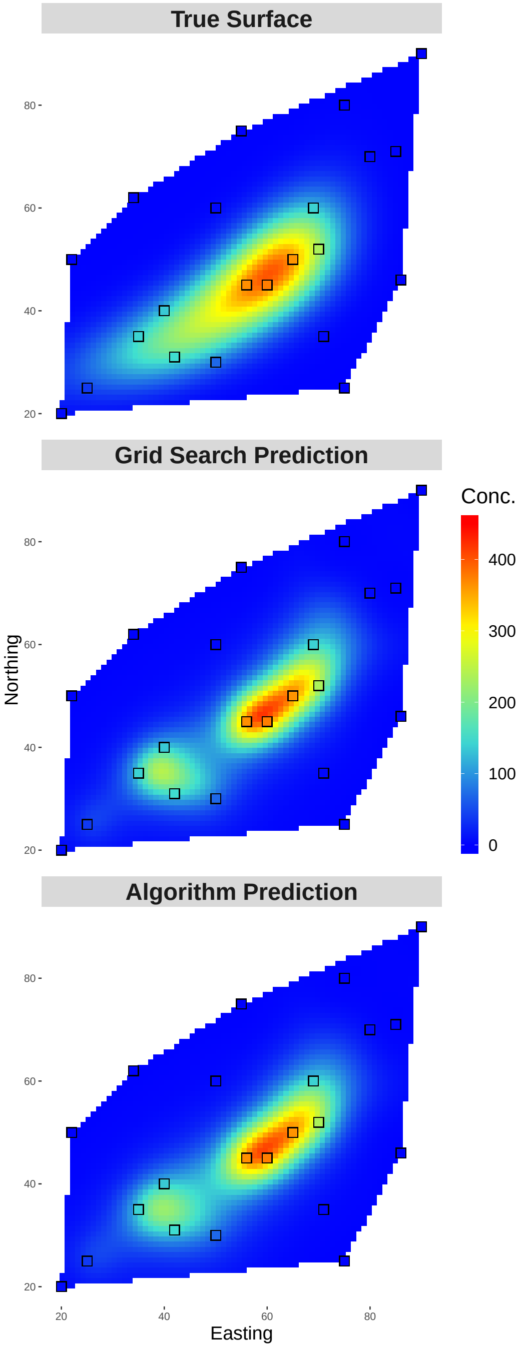

FIGURE 4.19: Predicted surfaces for spatio-temporal p-spline models built using the combination of λ and λ_{rel} chosen by a grid search and the proposed algorithm for PDE2 and a full design. The plot at the top is the true surface from PDE2.

Chapter 5

Ballooning

The spatio-temporal p-spline models used in Chapters 3 and 4, with one and two smoothing parameters respectively, indicated that their predictions were prone to an effect known as 'ballooning' for some models and data simulations (see Figures 3.7, 3.8, 3.9 and 4.17). Ballooning is a term used to describe the event where unusually high or low predictions are made in areas with little data support. This is caused by a steep gradient of observations just before a 'hole' in the predictor variables, as described by Molinari [2014].

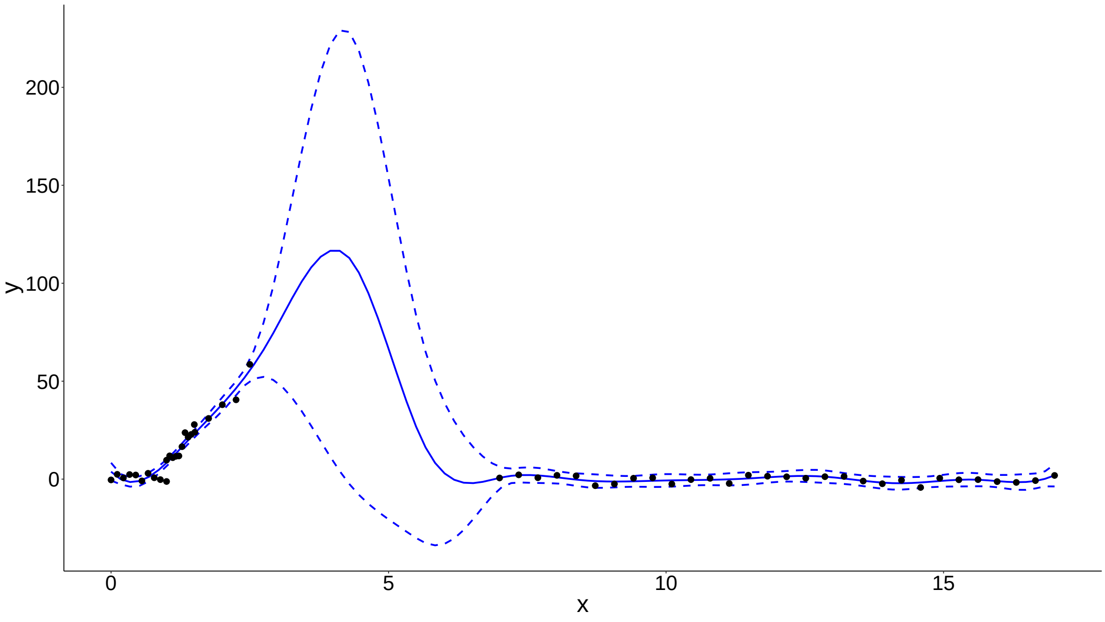

FIGURE 5.1: Prediction from a one-dimensional simulation where ballooning is evident, along with 95% confidence bands

Figure 5.1 shows ballooning occurring in a one dimensional spline model, the steep gradient of observations around $x = 2.5$ followed by an area with no data results in a spike in the predicted curve with no data to support this trend.

In a spatio-temporal groundwater setting, the distribution of the wells within the monitoring network has a large influence on whether ballooning is present in the predictions. The description of ballooning suggests a gridded network is not as likely to encounter ballooning as a randomly chosen network; which is more likely to see wells positioned closely together and regions with no well coverage. In the case of real world monitoring networks, transport infrastructure and housing heavily influence the locations of monitoring wells and thus the proximity of the wells can be sporadic and out with the control of the engineers who position them. Ballooning is not a problem that exclusively affects spline-based models, it has also been seen to occur in spatial and spatio-temporal Kriging models.

5.1 Basis Functions Simulation Study

To reduce the likelihood of ballooning occurring, Molinari [2014] suggest using an increased number of basis functions. However, their chosen number of functions is still relatively low with the maximum being (14, 8, 5) for each dimension respectively. To assess the effect of increasing the number of basis functions on the likelihood of ballooning occurring, two simulation studies were conducted using the PDEs and the designs presented in Chapters 3 and 4 and illustrated in Figures 3.1 and 4.10 respectively. These two designs differ in the sense that the first (from Chapter 3) contains many closely placed wells and also areas with no well coverage, whereas the second dataset (from Chapter 4) has wells which are much more evenly spread out across the study region.

The studies also aimed to determine the minimum resolution of basis functions required to prevent ballooning from occurring. For each simulated dataset six spatio-temporal p-splines models based on a single smoothing parameter and three spatio-temporal p-spline models based on two smoothing parameters were built with increasing numbers of basis functions. In both models the number of basis functions chosen for each spatial component reflected the spatial dimension ratio of the study region. For the temporal dimension of the one smoothing parameter model a lower number of basis functions was assigned to reflect the fact that the concentrations vary more over space than they do over time. In the two smoothing parameter model the same number of basis functions were used for time as for space, with the smoothness controlled by the additional smoothing parameter.

5.1.1 Simulation Study 1 - A dataset prone to ballooning

Tables 5.1 and 5.2 show the mean square prediction errors (MSPEs) for each of the one and two smoothing parameter models respectively on the first PDE. Using the simulation set up from Section 3.1, a full and a realistic design were considered with predictions being made at the 100^{th} and 167^{th} (last) time points.

The results indicate that for both sampling scenarios and models the predictive performance improves with increasing numbers of basis functions, which is unsurprising since extra basis functions allow for a more flexible model. They also suggest that ballooning may be present in some of the predictions of the lower resolution models due to the higher MSPEs and standard error (SE) values.

Focusing on the one smoothing parameter model and the models where the number of spatial basis functions are fixed, increasing the number of temporal basis functions does not appear to have much effect on the predictive performance of the models with (14, 8) and (25, 15) spatial basis functions. On the other hand, there is a significant improvement in predictive performance for the models with (8, 5) spatial basis functions when the temporal basis functions are increased, particularly in the case of the full design.

Table 5.1: Mean Square Prediction Errors (MSPEs) and standard errors (SEs) at the 100th and 167th (last) time points of PDE1, Equation 3.1, under realistic and full sampling scenarios (200 simulations) for one smoothing parameter spatio-temporal p-spline models with differing numbers of basis functions (northing (N), easting (E), time (T)).

# Basis Functions	Realistic Design		Full Design	
	Time 100	Time 167	Time 100	Time 167
(8, 5, 3)	4.5295 (0.13)	1.0682 (0.04)	90.3621 (11.71)	32.7145 (4.32)
(8, 5, 8)	3.0611 (0.10)	0.6346 (0.02)	16.5130 (1.33)	8.0879 (0.49)
(14, 8, 3)	0.4647 (0.01)	0.3728 (0.01)	1.2012 (0.09)	0.7090 (0.04)
(14, 8, 8)	0.4724 (0.01)	0.3656 (0.01)	0.4288 (0.02)	0.3948 (0.01)
(25, 15, 3)	0.3575 (0.01)	0.4874 (0.01)	0.3071 (0.01)	0.4827 (0.01)
(25, 15, 8)	0.3723 (0.01)	0.4722 (0.01)	0.2876 (0.01)	0.4329 (0.01)

Table 5.2: Mean Square Prediction Errors (MSPEs) and standard errors (SEs) at the 100th and 167th (last) time points of PDE1 under realistic and full sampling scenarios (200 simulations) for two smoothing parameter spatio-temporal p-spline models with differing numbers of basis functions (northing (N), easting (E), time (T)).

# Basis Functions	Realistic Design		Full Design	
	Time 100	Time 167	Time 100	Time 167
(8, 5, 8)	8.7635 (0.88)	7.0537 (0.96)	18.1504 (5.19)	9.8443 (3.37)
(14, 8, 14)	0.6423 (0.03)	0.5625 (0.03)	19.2282 (4.44)	19.8401 (4.54)
(18, 10, 18)	0.6909 (0.04)	0.7375 (0.04)	4.0425 (0.48)	4.4272 (0.49)

In the two smoothing parameter models results, although increasing the number of basis functions generally improves the predictive performance for the full design, it does not match the performance of the one parameter model.

5.1.2 Simulation Study 2 - A dataset with a more evenly spread design

Tables 5.3 and 5.4 show the mean MSPEs for both models on the second PDE simulation from Chapter 4. Again, two time points were used for prediction, namely, the 10^{th} and the 32^{nd} (last time) and a full and realistic (incomplete) design were considered. The results indicate that it is likely ballooning is not present in any of the model predictions since the mean MSPE across the models is relatively consistent and the SEs are low.

This is likely due to the wells in this design being more evenly spread across the study region.

Table 5.3: Mean Square Prediction Error (MSPE) and standard errors (SEs) at the 10th and 32nd (last) time points of PDE2 under realistic and full sampling scenarios (200 simulations) for the one smoothing parameter spatio-temporal p-spline models with differing numbers of basis functions (northing (N), easting (E), time (T)).

# Basis Functions	Realistic Design		Full Design	
	Time 10	Time 32	Time 10	Time 32
(8, 8, 3)	0.1934 (0.002)	0.3134 (0.006)	0.1499 (0.001)	0.1945 (0.003)
(8, 8, 8)	0.1301 (0.002)	0.3252 (0.006)	0.1121 (0.001)	0.2255 (0.003)
(14, 14, 3)	0.4852 (0.004)	0.3831 (0.005)	0.6609 (0.005)	0.3131 (0.004)
(14, 14, 8)	0.2226 (0.002)	0.3550 (0.005)	0.2089 (0.001)	0.2827 (0.002)
(20, 20, 3)	0.9422 (0.003)	0.7090 (0.007)	1.4235 (0.012)	0.6840 (0.008)
(20, 20, 8)	0.3972 (0.003)	0.3029 (0.006)	0.3962 (0.002)	0.4399 (0.003)

Table 5.4: Mean Square Prediction Error (MSPE) and standard errors (SEs) at the 10th and 32nd (last) time points of PDE2 under realistic and full sampling scenarios (200 simulations) for the two smoothing parameter spatio-temporal p-spline models with differing numbers of basis functions (northing (N), easting (E), time (T)).

# Basis Functions	Realistic Design		Full Design	
	Time 10	Time 32	Time 10	Time 32
(8, 8, 8)	0.1370 (0.00)	0.2879 (0.01)	0.1169 (0.00)	0.2028 (0.00)
(14, 14, 14)	0.3607 (0.01)	0.5302 (0.01)	0.4122 (0.01)	0.5605 (0.01)
(18, 18, 18)	0.7149 (0.01)	1.1230 (0.01)	0.7347 (0.00)	1.0976 (0.00)

Interestingly, as the number of basis functions increases, in both models, the MSPE also generally increases which is the converse to what would be expected and to what was seen in the results for the first PDE. This can be explained by looking at the predictions (Figures C.1, C.2, C.3 and C.4). The models with a higher number of basis functions produce a less smooth surface with much more localised contaminant plumes compared with the lower resolution models, suggesting that these models are overfitting. Looking at the true surface for the last time point, the strength of the contamination in the plume very gradually declines to the left. The model with a lower number of basis functions mimics this decline more accurately than a model with a higher number of basis functions.

5.2 A Measure for Detecting Ballooning

During this study, ballooning was observed in the predictions of the first PDE for both spatio-temporal p-spline models with a lower number of basis functions. Predictions from one simulation are shown in Figure 5.2; here positive and negative balloons can be seen. The well network in this dataset contained clusters of closely positioned wells along with regions with no well coverage i.e. conditions where ballooning is expected to occur.

Along with increasing the number of basis functions, Molinari [2014] also suggest other alterations that can be made to the model specification to try and reduce the occurrences of ballooning, for example, using a first order difference penalty. However this, along with an increased number of basis functions does not always stop ballooning occurring. Without visualising the data, which can be inconvenient when running several models, detection of ballooning can be difficult and thus a measure was developed to flag when ballooning may be present in the predictions.

For unusual predictions to be classed as ballooning two criteria (detailed below) must be satisfied. These criteria were designed to highlight when unusually high or low predictions were made in areas with little data support. Predictions which are:

1. one standard deviation above the maximum observed value or one standard deviation below the minimum observed value,

   and,

2. are at locations whose distance from the nearest well is greater than the median distance of any location in the sampling area from its nearest well.

are classed as ballooning.

In the case of the realistic sampling design, when not every well was sampled at every time, the distances calculated in step 2 were between all wells and pixels, not just the wells that had been sampled at the prediction time point.

Figure 5.3 highlights unusually high/low predictions from a prediction using a model with (8, 5, 3) basis functions under the full sampling scenario (shown in Figure 5.2).

Points highlighted in pink indicate predictions which are unusually low and points highlighted in black indicate predictions which are unusually high. All of these highlighted locations satisfy the first criterion.

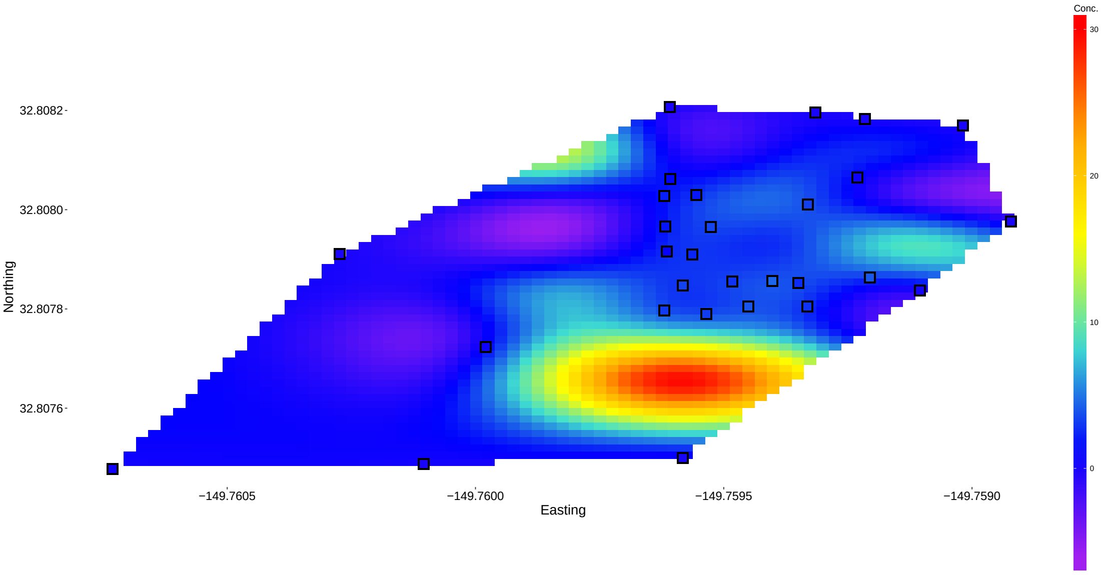

FIGURE 5.2: Predicted surface using a spatio-temporal p-splines model with ballooning evident. Wells are plotted as filled black squares.

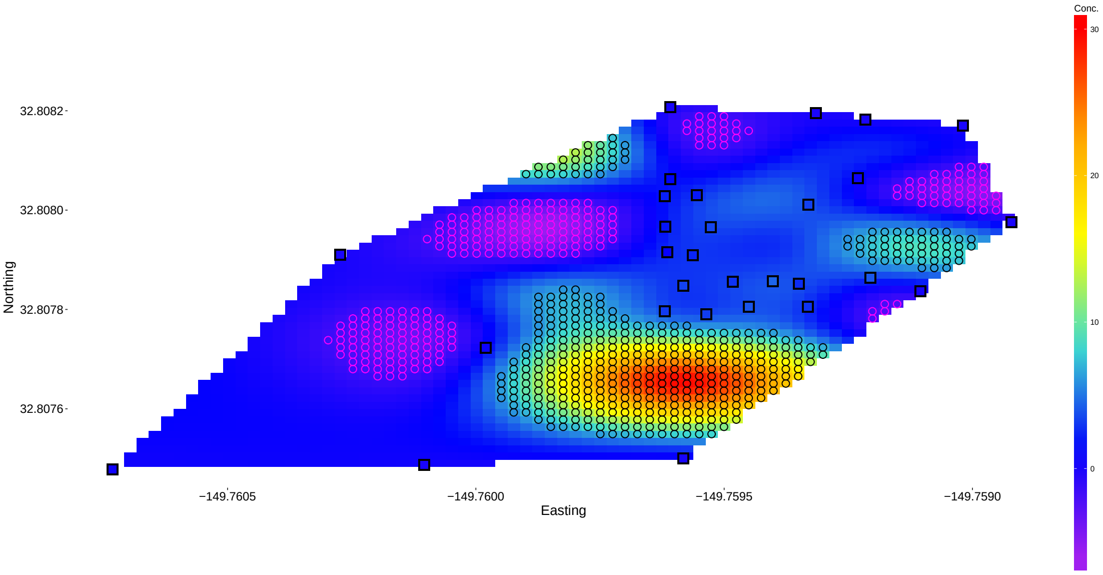

FIGURE 5.3: Balloons highlighted by the detection mechanism, pixels outlined in pink circles indicate unusually low predictions whilst black circles indicate unusually high predictions. Wells are plotted as filled black squares.

There are several areas highlighted in this prediction, some of which are not located near any wells, for example the region of high concentration in the lower right section of the surface. This area would be highlighted as ballooning. There are also several areas where unusually low predictions have been made, these regions would also be highlighted as ballooning.

5.2.1 Ballooning Simulation Study

The simulation study in Section 5.1.1, on the first PDE, was repeated with the ballooning detection measure incorporated after predictions were made for each simulation. The measure was also used on the second simulated dataset and, as anticipated, ballooning was not detected in any simulations. Only the results of the first PDE will therefore be discussed.

Figures 5.4 and 5.5 show the number of occurrences of ballooning from the 200 simulations of each p-splines model. As well as testing for the presence of ballooning, if ballooning was present, the proportion of pixels at which ballooning occurred was also computed. Table 5.5 contains the mean number of ballooned pixels over all simulations when ballooning was detected for the p-splines model with one smoothing parameter and Table 5.6 contains the proportions for the model with two smoothing parameters.

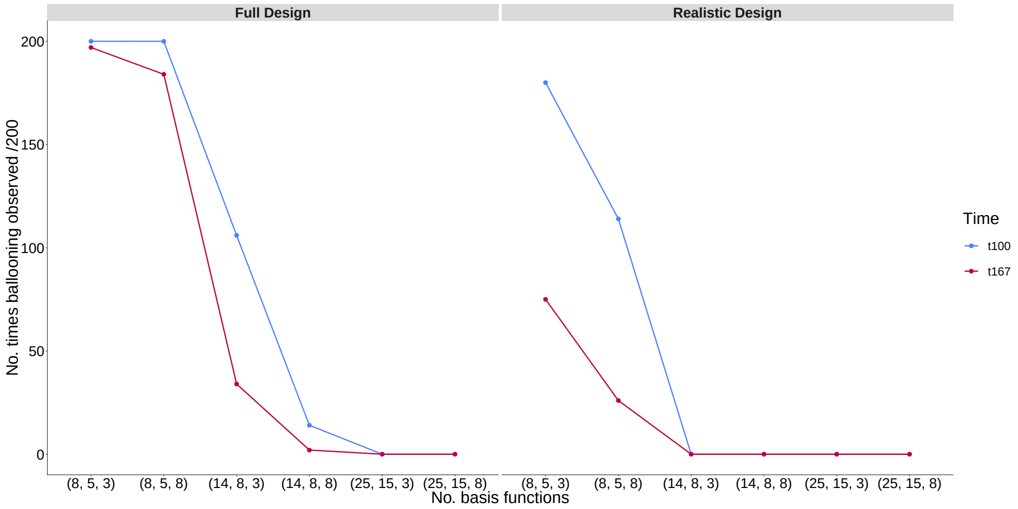

FIGURE 5.4: Counts of the number of times ballooning was detected for each combination of basis functions in the spatio-temporal p-splines model, with one smoothing parameter, for each dataset from PDE1

As suggested by Molinari [2014], in general as the number of basis functions increases the frequency of ballooning decreases for both designs, models and prediction times. For the full design ballooning seems to occur more frequently for the lower resolution models compared with the real design. This can be explained by the description of ballooning given at the beginning of the chapter. The well network contains several clusters of closely positioned wells. A full design implies data are collected from every well at every time point and thus there are more scenarios when a steep gradient is present followed by a 'hole' in the data, compared with the real design which is more likely to have less steep gradients in the predictions.

Table 5.5: Mean proportion of ballooned pixels for those simulations which detected ballooning at the 100^{th} and 167^{th} time points of PDE1 under realistic and full sampling scenarios (200 simulations) for one smoothing parameter spatio-temporal p-spline models with differing numbers of basis functions (northing (N), easting (E), time (T)).

# Basis Functions	Realistic Design		Full Design	
	Time 100	Time 167	Time 100	Time 167
(8, 5, 3)	0.0672	0.0200	0.2601	0.1687
(8, 5, 8)	0.0436	0.0119	0.1478	0.0839
(14, 8, 3)	0	0	0.0358	0.0190
(14, 8, 8)	0	0	0.0187	0.0032
(25, 15, 3)	0	0	0	0
(25, 15, 8)	0	0	0	0

Prediction time also has an effect on the likelihood of ballooning. Figures 5.4 and 5.5 show the count lines for time 167 (the last time point) below the lines for time 100 (in the middle of the data) for both models and sampling scenarios. This suggests that ballooning is less likely to occur when predictions are made at sampling times located at the ends of the data range. From the count lines for the two smoothing parameter model (Figure 5.5), the effect of different sampling scenarios on the frequency of ballooning being observed is very evident, with the full design counts not dropping off as rapidly as the realistic design. For this model, increasing the number of basis functions does not completely eliminate ballooning.

In both models, as the total number of basis functions increases, the proportion of ballooned pixels decreases. This is as expected, since ballooning generally only occurs

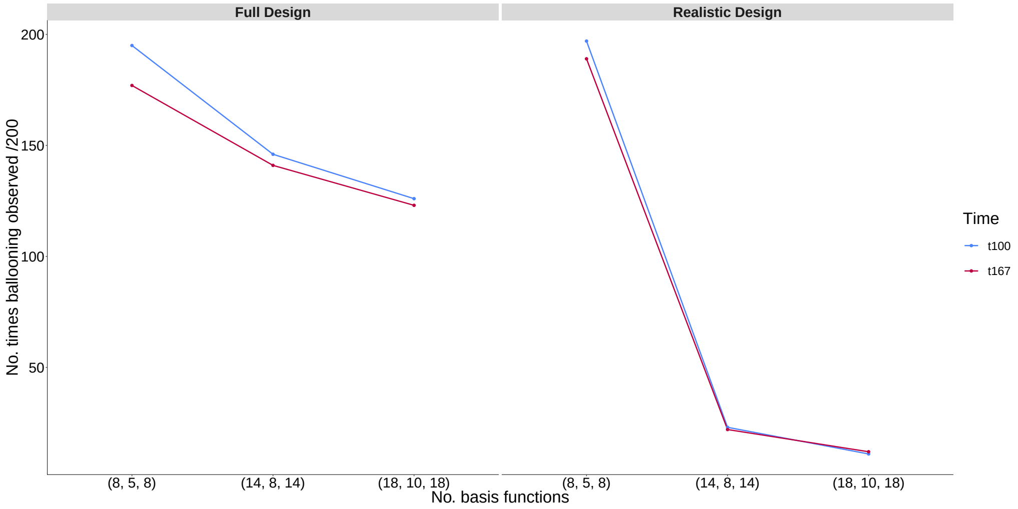

Figure 5.5: Counts of the number of times ballooning was detected for each combination of basis functions in the spatio-temporal p-splines model, with two smoothing parameters, for each dataset from PDE1

with a few basis functions. With more basis functions the width of each is narrower and thus a smaller number of pixels is covered.

Table 5.6: Mean proportion of ballooned pixels for those simulations which detected ballooning at the 100^{th} and 167^{th} time points of PDE1 under realistic and full sampling scenarios (200 simulations) for two smoothing parameter spatio-temporal p-spline models with differing numbers of basis functions (northing (N), easting (E), time (T)).

# Basis Functions	Realistic Design		Full Design	
	Time 100	Time 167	Time 100	Time 167
(8, 5, 8)	0.1206	0.0810	0.1201	0.1040
(14, 8, 14)	0.0123	0.0111	0.1348	0.1259
(18, 10, 18)	0.0301	0.0254	0.0697	0.0635

5.3 Summary of Ballooning Properties

From the simulation studies conducted in Sections 5.1 and 5.2.1 the following conclusions can be made about ballooning.

- The likelihood of ballooning occurring decreases as the number of basis functions increases for both one and two smoothing parameter p-spline models. However, care should still be taken, particularly when using a two smoothing parameter model.

- Well positioning has a large influence on whether ballooning is present or not. A well network with clusters of closely positioned wells is more likely to encounter ballooning when a full design is observed compared with an incomplete design, on the same network, which is more likely to mimic a regular grid. This was evident in the basis function simulation studies presented in Section 5.1.1. The simulation study in Section 5.1.2 showed that ballooning is unlikely to occur if the well network is evenly spaced across the study region.
- Predictions for times positioned at the end of the data are less likely to exhibit ballooning compared with times positioned in the middle of the data.

5.4 Conservation of Plume Mass Penalty

During the comparison study conducted in Section 5.1, ballooning was detected when using spatio-temporal p-spline models with a low number of basis functions. Section 5.2 went on to suggest a method for detecting when ballooning might be present. As mentioned earlier, increasing the number of basis functions reduces the chance of ballooning being present. This ‘fix’ is satisfactory; however, with spatio-temporal data, increasing the number of basis functions in each dimension dramatically increases the computation time for the model.

In order to use a model with a lower number of basis functions but still suppress ballooning and obtain reliable and robust predictions, a penalty based on the change in contaminant plume mass over time was proposed.

5.4.1 Penalty Formulation

The penalty was motivated by the idea that the contaminant cloud would move across the study region over time, however, its mass should not change significantly. Thus the penalty is designed to penalise changes in the predicted plume mass over time. In theory this should control unusual fluctuations in the total mass which is the primary characteristic of ballooning. The change in mass is computed by first integrating the function over the spatial domain to give an estimate of the plume mass. The derivative of this is then taken with respect to time to give the change in mass over time. Finally, this term is squared and integrated over time to give the total change in mass over the time frame of interest.

The penalised least squares expression using the proposed penalty is denoted in Equation 5.1, where **PEN** denotes the new penalty.

$$\sum_{i=1}^{n}(y_i - m(x_{1i}, x_{2i}, t_i)))^2 + \lambda \underbrace{\int_t \left\{ \frac{d}{dt}\left[\iint m(x_1, x_2, t)\ dx_1 dx_2 \right] \right\}^2 dt}_{\textbf{PEN}} \tag{5.1}$$

as before, $m(x_{1i}, x_{2i}, t_i) = \sum_{jkl} \alpha_{jkl} B_j(x_{1i}) B_k(x_{2i}) B_l(x_{ti}$ and $B_j(x_1)$, $B_k(x_2)$ and $B_l(x_t)$ are basis functions for each dimension with corresponding basis coefficients α_{jkl}. Here a B-spline basis is used.

The change in mass penalty term, λ**PEN**, can be decomposed as follows to allow for more efficient computation,

$$\begin{aligned}
&\lambda \int_t \left\{ \frac{d}{dt}\left[\iint m(x_1, x_2, t)\ dx_1 dx_2 \right] \right\}^2\ dt \\
&= \lambda \int_t \left\{ \frac{d}{dt}\left[\iint \sum_{jkl} \alpha_{jkl} B_j(x_1) B_k(x_2) B_l(t)\ dx_1 dx_2 \right] \right\}^2\ dt \\
&= \lambda \int_t \left\{ \sum_{jkl} \left[\alpha_{jkl} \int_{x_1} B_j(x_1)\ dx_1 \int_{x_2} B_k(x_2)\ dx_2\ B_l'(t) \right] \right\}^2\ dt \\
&= \lambda \int_t \left[\sum_{jkl} \sum_{mno} \alpha_{jkl} \alpha_{mno} \int_{x_1} B_j(x_1)\ dx_1 \int_{x_1} B_m(x_1)\ dx_1 \cdot \right. \\
&\qquad \left. \int_{x_2} B_k(x_2)\ dx_2 \int_{x_2} B_n(x_2)\ dx_2\ B_l'(t) B_o'(t) \right]\ dt \\
&= \lambda \sum_{jkl} \sum_{mno} \alpha_{jkl} \alpha_{mno} \int_{x_1} B_j(x_1)\ dx_1 \int_{x_1} B_m(x_1)\ dx_1 \cdot \\
&\qquad \int_{x_2} B_k(x_2)\ dx_2 \int_{x_2} B_n(x_2)\ dx_2 \int_t B_l'(t) B_o'(t)\ dt
\end{aligned} \tag{5.2}$$

The integral of each B-spline basis function, $\int B_i(x)\ dx$, will be the same in each dimension with the exception of the first and last p functions, where p is the degree of the basis function. The value of these integrals will be smaller due to there not being full basis functions at the beginning and end of the basis. Figure 5.6 illustrates this idea, with the functions whose integral is the same taking the same colour. Computing the integral of a B-spline is done by computing the integral of the polynomial pieces used in its construction and adding them together.

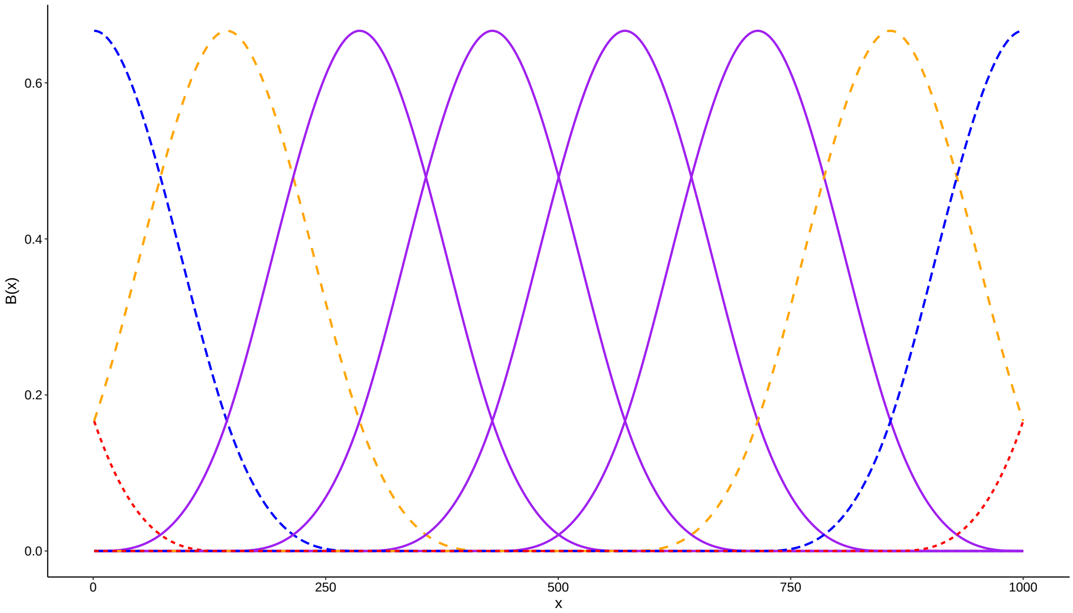

FIGURE 5.6: Basis functions of degree 3. Functions plotted in the same colour have the same integral.

Following on from the decomposition in Equation 5.2, the penalty can be denoted as a quadratic form. By denoting the penalty in this form the penalised least squares expression for penalised regression splines, detailed in Equation 2.19, can be utilised. Thus, in vector-matrix notation, the expression to be minimised for a model using the conservation of mass penalty is:

$$\mathrm{PLS}(\boldsymbol{\alpha}) = (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha})^{\top} (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^{\top} \mathbf{K} \boldsymbol{\alpha} \tag{5.3}$$

where

$$\mathbf{K} = \mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C} \tag{5.4}$$

and

$$\begin{aligned}
\mathbf{A}[j,m] &= \int_{x_1} B_j(x_1)\ dx_1 \cdot \int_{x_1} B_m(x_1)\ dx_1, \\
\mathbf{B}[k,n] &= \int_{x_2} B_k(x_2)\ dx_2 \cdot \int_{x_2} B_n(x_2)\ dx_2, \\
\mathbf{C}[l,o] &= \int_t B_l'(t)B_o'(t)\ dt.
\end{aligned}$$

To assess the effectiveness of the proposed penalty, two simulated datasets were considered. The findings of these studies are presented in the following sections.

5.4.2 Two-Dimensional Simulation Study - a toy example

The first simulated dataset consisted of one spatial dimension and one temporal dimension. The data were simulated from a normal density function with increasing standard deviation over time, shown in Figure 5.7. These particular data were used since the area under the density curve remains constant irrespective of the standard deviation and so, if the penalty has been correctly formulated, regardless of the severity of the penalty parameter λ, the predicted mass should not change over time. The dataset was made up of 10 monitoring wells each sampled at 10 time points.

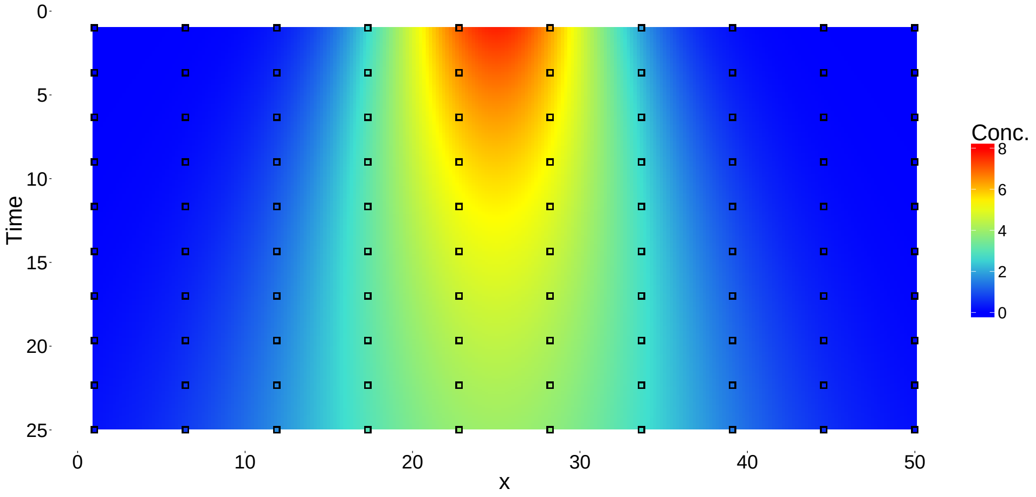

FIGURE 5.7: True simulated data for the two dimensional study. The data consists of a normal density function over space with increasing standard deviation over time. Points indicate locations where observations were recorded.

Two-Dimensional Penalty Formulation

The mass penalty term for data with one spatial and one temporal dimension can be formulated as follows:

$$\begin{aligned}
&\lambda \int_t \left\{ \frac{d}{dt} \left[\iint f(x,t)\ dx \right] \right\}^2\ dt \\
&= \lambda \int_t \left\{ \left[\iint \sum_{jk} \alpha_{jk} B_j(x) B_k'(t)\ dx \right] \right\}^2\ dt \\
&\vdots \\
&\vdots \quad \textit{(derived in a similar manner to Equation 5.2)} \\
&\vdots \\
&= \lambda \sum_{jk} \sum_{mn} \alpha_{jk} \alpha_{mn} \int_{x_1} B_j(x)\ dx \int_{x_1} B_m(x)\ dx \int_t B_k'(t) B_n'(t)\ dt.
\end{aligned} \tag{5.5}$$

Here, $B_j(x)$ and $B_k(t)$ are basis functions for the space and time dimensions respectively, with corresponding basis coefficients α_{jk}.

Results

To assess whether the proposed penalty correctly penalised changes in the plume mass over time, five penalty parameter values were considered $\lambda = \{1 \times 10^{-6}, 1 \times 10^{-3}, 1 \times 10^{0}, 1 \times 10^{3}, 1 \times 10^{6}\}$. Penalised regression spline models were built with the proposed penalty for each λ and compared to the true data shown in Figure 5.7. P-spline models with the same values of λ were also built for additional comparison.

Figure 5.8 shows the predicted surfaces for each value of λ (rows) along with the true underlying surface. As expected, as the value of λ increases the p-splines model forces the prediction to a constant surface. On the contrary, increasing λ has little or no effect on the predicted surface of the model with the conservation of mass penalty, with the surfaces for the five considered models being almost identical. This indicates that regardless of the value of λ no penalty is applied because no change in mass is present between prediction times. Table 5.7 shows the MSPEs for each prediction in Figure 5.8. They further back up that the mass penalty is working in the intended manner since the MSPE for each model is the same to four significant figures. When λ is very small i.e. 1×10^{-6} and 1×10^{-3}, the model with the difference penalty performs as well as the model with the mass penalty.

Table 5.7: Mean Square Prediction Errors (MSPEs) for increasing values of $\lambda \in \{1 \times 10^{-6}, 1 \times 10^{-3}, 1 \times 10^{0}, 1 \times 10^{3}, 1 \times 10^{6}\}$ using the mass penalty and a difference penalty.

λ	MSPE	
	Mass Penalty	Difference Penalty
1×10^{-6}	0.0081	0.0093
1×10^{-3}	0.0081	0.0041
1×10^{0}	0.0081	0.7814
1×10^{3}	0.0081	3.7333
1×10^{6}	0.0081	3.7450

In addition to looking at the MSPEs and predicted surfaces, the plume mass at each time was also calculated through numerical integration of the predicted values over the spatial domain at each time point. The plume mass of the predictions for each model at each time are presented in Figure 5.9.

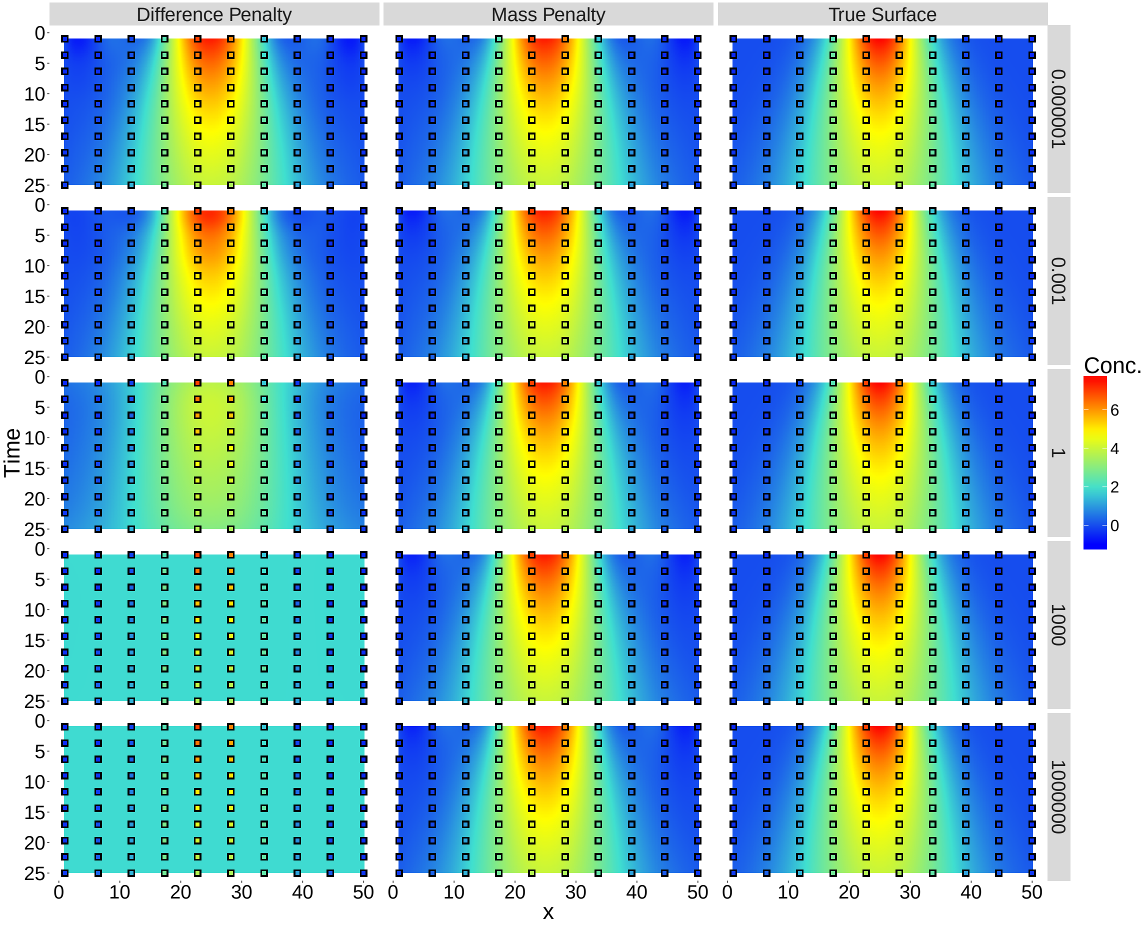

FIGURE 5.8: Predicted surfaces using spline models with the difference penalty (left) and the conservation of mass penalty (middle). Five values of the penalty parameters $\lambda = \{1 \times 10^{-6}, 1 \times 10^{-3}, 1 \times 10^{0}, 1 \times 10^{3}, 1 \times 10^{6}\}$ (rows) are presented along with the true underlying surface (right).

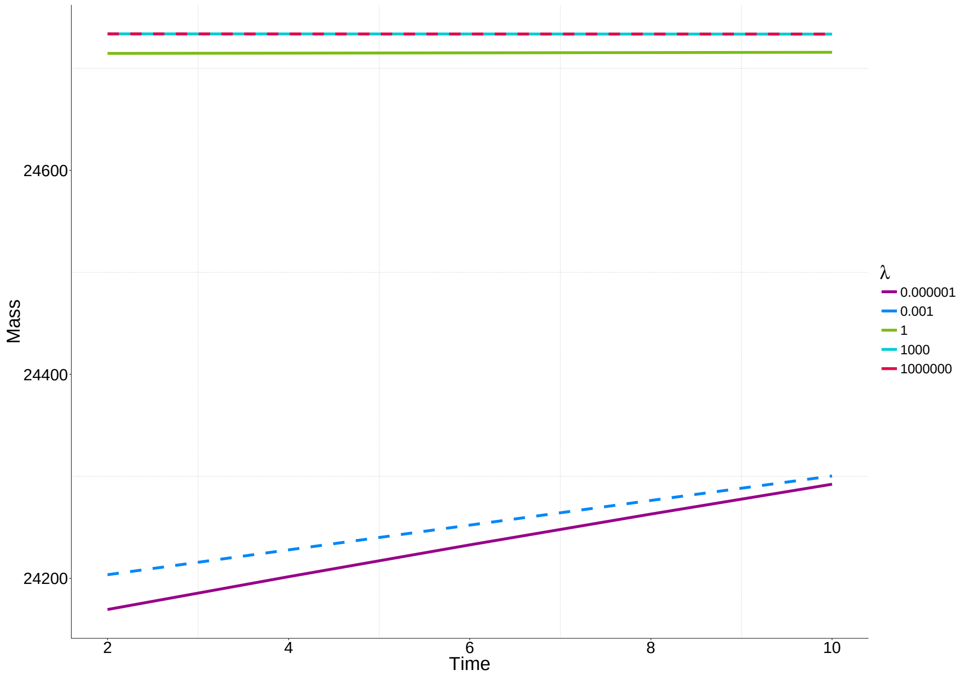

FIGURE 5.9: Predicted plume mass at each time for each of the five considered penalised regression spline models with the conservation of mass penalty.

The predicted plume masses at each time further back up that the penalty is working as hoped. When the penalty is essentially 'turned off', i.e. $\lambda = 1 \times 10^{-6}$, the mass does not remain completely constant, it gradually increases by about 30 units at each prediction time. Increasing the penalty parameter to $\lambda = 1 \times 10^{-3}$ results in a smaller increase in plume mass, with the change in mass between the first and last time points being smaller compared with $\lambda = 1 \times 10^{-6}$. This trend of the change in mass reducing as λ increases continues for $\lambda = 1$. As λ becomes increasingly large, say $\lambda = 1 \times 10^{3}$ or $\lambda = 1 \times 10^{6}$, the change in mass over time is negligible, illustrated by the constant horizontal lines at the top of Figure 5.9. This indicates that changes in the mass have been heavily penalised, forcing a model which predicts a constant mass across all prediction times to be selected.

5.4.3 Three-Dimensional Simulation Study

The second simulation study was performed over two spatial dimensions and one temporal dimension i.e. spatio-temporal data. The data were made up of a two-dimensional normal density in space moving in a downward spiral motion over time, the temporal trajectory of the spatial density is shown in Figure 5.10 along with the well network that was used in this particular dataset, the true plume mass does not change significantly over time. However, the contamination would not be detected by the wells and thus the model, until later time points, since most of the wells are located in the lower half of the study region. Based on this idea, there should be a point in time where the predicted contaminant mass will suddenly increase and the proposed penalty then add contaminant mass into the predictions at the earlier time points.

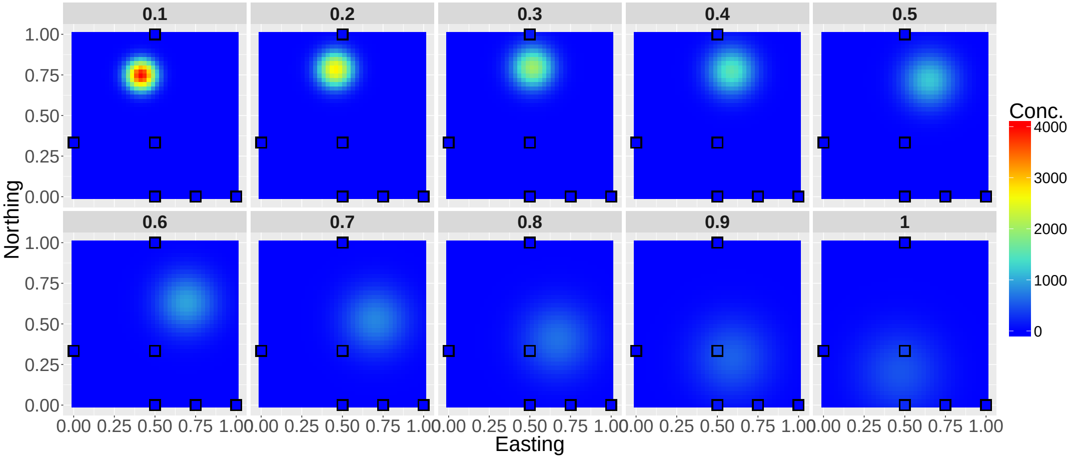

FIGURE 5.10: True simulated data for the three dimensional study. The data consist of a normal density function over space with a downward spiral motion over time. Points indicate locations where observations were recorded.

Model Formulation

In the spatio-temporal setting, the mass penalty was used in addition to a spatial difference penalty. The mass penalty was designed to penalise changes over time but this does not control the spatial variation. Therefore, by retaining the spatial component of the difference penalty, a control for the spatial smoothness of the model is provided. Thus the penalised least squares expression to be minimised is

$$\mathrm{PLS}(\boldsymbol{\alpha}) = (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha})^\top (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) + \lambda_1 \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} + \lambda_2 \boldsymbol{\alpha}^\top \mathbf{D}_s^\top \mathbf{D}_s \boldsymbol{\alpha} \tag{5.6}$$

where $\mathbf{K}$ takes the form given in the penalty description in Equation 5.4 and $\mathbf{D}_s$ is a difference matrix of first order penalising only the spatial component.

Results

To assess the performance of the penalty, various values of λ_1 were used to build penalised spatio-temporal regression spline models, along with $\lambda_2 = 1 \times 10^{-4}$. This value of λ_2 was chosen as it was the mean optimal value, using the log posterior criterion for selection (see Section 2.3.2), when independent spatial p-splines model were built at each time point.

Figure 5.11 shows the predicted surfaces at each time point for each value of λ_1. When the mass penalty is 'turned off', i.e. $\lambda_1 = 0$, the model does not detect any contamination in the first few time points. This is due to the contamination not spreading over any wells. It then detects a small amount of contamination at the 3^{rd} and 4^{th} time points at the top boundary of the study region. This contamination then disappears for two prediction times and reappears in the last four times, when the contaminant cloud moves over monitoring wells. Under these modelling conditions, the mass of the prediction at each time fluctuates as the contaminant plume moves across the study region and is picked up by monitoring wells.

As λ_1 increases, the predictions at the first six time points gradually change. This can be seen by looking down the columns in Figure 5.11. From $\lambda_1 = 0.01$ upwards, masses of contaminant begin to appear in regions where there is no well coverage at these first six time points. This is evidence of the mass penalty forcing the model to add the mass that is present in the later prediction times into the earlier predictions. The predictions for the final four time points are relatively similar for all values of λ_2, the shape of the contaminant plume does however become more defined as λ_2 increases.

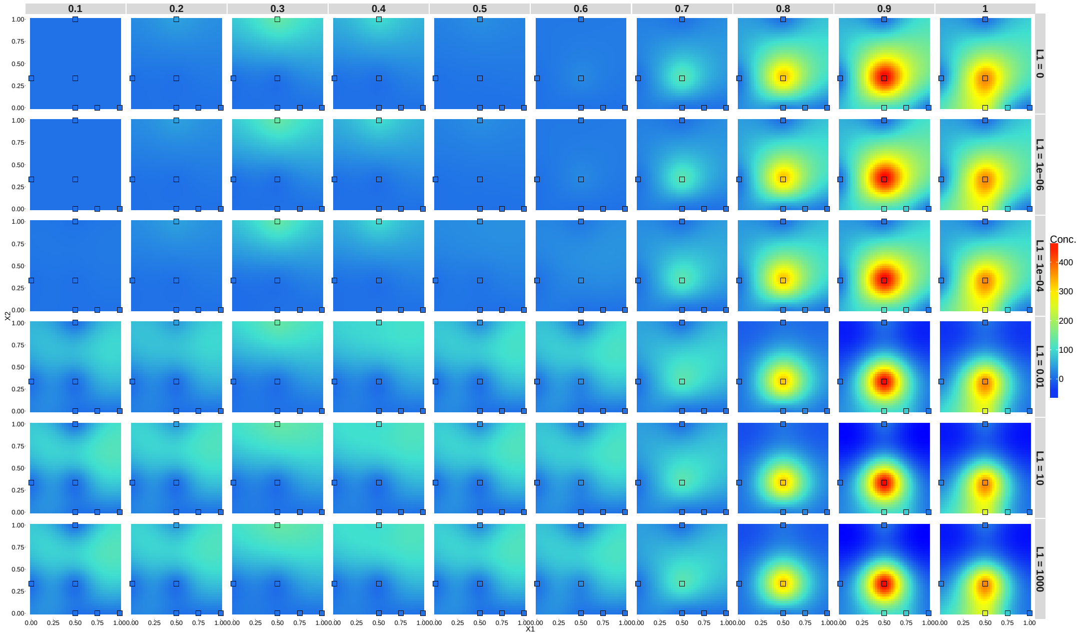

FIGURE 5.11: Predicted surfaces at each time point for penalised spatio-temporal spline models built with $\lambda_1 \in \{0, 1 \times 10^{-6}, 1 \times 10^{-4}, 1 \times 10^{-2}, 10, 1000\}$ and $\lambda_2 = 1 \times 10^{-4}$ using the conservation of mass penalty and a spatial difference penalty. Columns represent prediction times and rows correspond to each values of λ_1 for the mass penalty.

Figure 5.12 shows the total predicted plume mass at each time point for each considered model. The plume mass was computed by numerical integration over the predictions at each time point. When the mass penalty is 'turned off' i.e. $\lambda_1 = 0$ the mass of the plume initially increases, then decreases before again increasing. This follows the trend seen in the predictions in the top row of Figure 5.11. In a similar manner to the 2D simulation study, as λ_1 increases the fluctuations in the mass decrease until a constant mass at all times is reached when $\lambda_1 = 10$. In order to achieve this constant contaminant mass, the model forces increasingly more mass into regions of the first 6 time points where there is no well coverage rather than suppressing the sudden increase as had been hoped.

The plots of the predicted surfaces and total mass over time, from this simulation study, suggest that the conservation of mass penalty has been computed as desired and is conserving the contaminant mass over time. However, the plots highlight issues with the model formulation in practice.

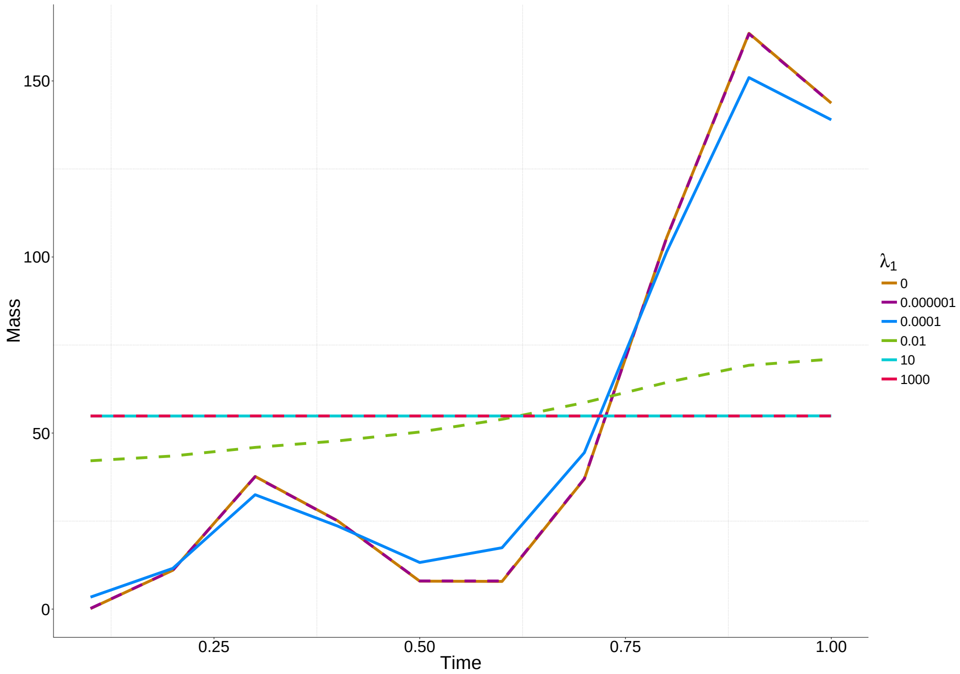

FIGURE 5.12: Predicted plume mass at each time for each of the five considered spline models with the conservation of mass penalty.

5.4.4 Summary and Issues

In the two-dimensional study, presented in Section 5.4.2, the proposed conservation of mass penalty appeared to be correctly conserving the predicted mass over time. The primary aim of this first study was to test the proposed methodology, hence the sampling design and dataset were chosen to ensure the mass remained constant over time. Increasing values of the penalty parameter λ, were used for model building and prediction. The results indicated that regardless of the value of the penalty parameter, the predicted mass did not significantly change over time suggesting the methodology and code were working in their intended manner and changes in the mass were being penalised correctly.

The second simulation study, presented in Section 5.4.3, was designed to assess how well the penalty could control and suppress a sudden change in the plume mass. The dataset and well network were chosen such that the contamination would not be picked up by the monitoring wells until the later time points in the simulation. It was hoped that the penalty would suppress this sudden increase in contaminant mass. However, in practice this is not what happened and in hindsight it is unsurprising. The high concentrations present in the observed data indicated that, at later times, contamination was present in the study region. Therefore, rather than compress this new additional mass, the penalty added extra mass into the predictions at earlier time points. Essentially, rather than controlling ballooning, the penalty caused the model to balloon more in earlier time points to compensate for the known mass that was observed in the later times. It is also worth noting that the simulated dataset is not mimicing exactly what happens when ballooning occurs since there are data to support the sudden change in mass, this is not the case when ballooning occurs.

Based on these results it would be expected that the penalty would add additional mass into the prediction times which do not have any balloons rather than squash the balloons that are present. Thus, the results suggest that the penalty may not be a suitable solution for preventing ballooning.

The penalty did however predict mass at time points where the observed data had no record of it due to the plume not passing over any monitoring wells at these time points. Monitoring wells missing contamination at some sampling events, due to the sporadic nature of some monitoring networks, is relatively common and thus, with improvements, the penalty could be used to inform models of the location of the contaminant mass when it is not picked up by the wells until later times. To use the penalty for this purpose, further developments are required. In its current state, the penalty conserves the mass as desired, but the model has no information on where to store this extra mass so it is forced into regions of the study area where there is no information from monitoring wells.

To overcome this issue of where this unaccounted for contaminant mass is located, one potential extension could be to incorporate a further penalty based on a PDE that describes the motion of the groundwater; see Frasso et al. [2016*b*]. This would then give the current model an idea of where the contaminant plume has travelled and hence where the mass is located when the observed data does not have any record of it until later. Determining the PDE that accurately describes the groundwater system is difficult and there is evidence of variation between sites. However, if an accurate representation were to be found and subsequently a penalty developed, this would give the model an idea of where contaminant mass could be present and hence where additional mass may be stored.

Chapter 6

Optimal Sampling Design of Monitoring Networks with Spatial and Spatio-temporal Models

Groundwater contamination poses a potential threat to human and environmental health, and the contamination can occur very easily. Whenever waste or chemicals are released into the environment, there is potential for the groundwater to become polluted. The clean-up operation can be difficult and expensive, with groundwater contamination most commonly occurring in densely populated regions where the land is exhaustively used.

Given that groundwater is located under the surface, determining the speed and direction of its flow is complex and subsequently determining the current and future location of a contaminant plume is also complex. To obtain accurate predictions, a high density of data is required, but, collecting groundwater data is expensive. Samples need to be collected from a set of wells at different locations and from sites of varying size, ranging from fuel stations to large refineries. Normally the data are collected by staff with expertise in engineering or science who are often ill informed on where these samples should be taken and when. Due to the constantly moving nature of a contaminant plume, both time and location are factors which influence the estimation of the contaminant plume and so these should be considered as decision variables when choosing sampling locations.

The following section presents a review of the methodology used to determine sampling designs for groundwater contamination data, both in a spatial and spatio-temporal context. Two objective functions, based on minimising the variance of the plume mass (VM) and the integrated prediction variance (IV), will then be presented and derived in Sections 6.3.1, 6.3.2, 6.4.1 and 6.4.2 for spatial and spatio-temporal Kriging and p-spline models. Several simulation studies will then be conducted to assess: the effect sampling frequency has on the optimal design; which wells can potentially be removed from the network i.e. well redundancy analysis and if a well were to be added to the network, where would it be positioned.

6.1 Review of Current Literature

Groundwater quality monitoring network design can be divided into three main subgroups: hydro-geological approaches, statistical approaches and model-based approaches; see Herrera and Pinder [2005], Loaiciga et al. [1992].

Before recent advances, sampling procedures most commonly fell into the category of hydro-geological approaches. Designs resulting from these approaches are based solely on the qualitative and quantitative hydrological information from the site of interest. These methods require large amounts of data and are best suited to studies where early contaminant detection is the main objective; see Herrera and Pinder [2005]. There are several obvious drawbacks to these approaches. Primarily, they rely on large volumes of data. Due to cost and time constraints, obtaining a sufficient number of observations to reduce uncertainty around the estimated plume is not always feasible.

In recent years, statistical and model-based approaches have become more frequently used. The aim in the statistical framework is to build on the hydro-geological approaches through developing designs based on inferences from the data. Geostatistics are used to explain the complex spatial and temporal variations in the data through covariance structures and the methods can further be used for prediction of the contaminant plume. In contrast, model based approaches use a combination of mathematical models and physical knowledge of ground water movement to anticipate the contaminant plume's location and the uncertainty associated with this estimation.

The type of design approach is generally determined by the methodology used to estimate the location of the contaminant plume; as described by Loaiciga et al. [1992]. Attempts have also been made to combine the statistical and model-based approaches, see Loaiciga [1989], Reed et al. [2000]. This work will primarily focus on statistical based approaches, with applications of the model-based approach being discussed where appropriate.

The main goal of optimal network design, in any application associated with data collected over space and time, is to determine the most appropriate location and frequency of observations in order to meet a predefined objective or set of objectives; see Fretwell et al. [2006], van Geer et al. [2008]. A comprehensive review of sampling design is provided by Maher et al. [1994], who highlight the importance of defining the problem that is to be solved and emphasise the importance of clearly defining the objectives prior to optimising the design. The objective function that is to be optimised should be chosen such that it reflects the objectives that are to be met. A prevalent theme amongst publications on environmental design is the use of an objective function based on minimising either the estimation or the prediction variance, combined with several design constraints. The enforced constraints vary across designs, with the most common constraints being the cost and time associated with obtaining samples; see Yeh [2015]. Cost constraints can also be imposed through capping the number of samples that can be taken. Section 6.1.1 explores commonly used objective functions and criteria in more detail.

The chosen objective function greatly influences the resulting design. Focussing on groundwater quality monitoring, Loaiciga et al. [1992] and later McPhee and Yeh [2005] classify groundwater monitoring network design objectives into four main categories:

- **Ambient Monitoring** aims to understand temporal trends in regional groundwater quality variations. This is achieved by regularly sampling wells on a regional basis. For this type of monitoring, generally water supply wells are sampled over monitoring wells.

- **Detection Monitoring** seeks to identify the presence of contaminants as soon as their concentration exceeds pre-determined levels. This type of monitoring is needed around sources of contamination, for example in toxic waste sites; see Environmental Protection Agency [2017].

- **Compliance Monitoring** adheres to groundwater quality monitoring guidelines after the existence of chemical compounds are detected. This is achieved by setting out a series of strict monitoring requirements for specific compounds, thus allowing the re-mediation process to be monitored.
- **Research Monitoring** involves sampling over space and time to meet predefined targets and research aims.

In the following review, the focus will be on detection and compliance groundwater monitoring, which aim to identify groundwater contamination as soon as it is released and monitor the contaminant plumes movements after detection. Detection monitoring aims to satisfy three contradicting objectives, as detailed by Meyer et al. [1994] and Angulo and Tang [1999], namely (i) maximisation of the probability of detecting contaminants, (ii) minimisation of plume mass when detected and (iii) minimisation of the total cost.

When constructing optimal designs for groundwater quality monitoring networks there are several choices within the design that need to be made prior to the optimisation. The first choice relates to where and how many samples are taken. Are they from predetermined locations such as a current network of wells, or are the locations of the wells chosen as part of the design to create a new network? Following on from the first option, if the wells are already positioned, can wells be added into the network, or is the aim to remove redundant wells? Cost constraints can be incorporated by limiting the number of wells that can be sampled in a single event or, in the spatio-temporal case, over a series of events. Alternatively, minimisation of a cost function can be adopted as an additional objective function, where the cost function incorporates the financial expense of sampling from specific wells or the price of adding a well into the network; see Angulo and Tang [1999].

Spatial sampling designs for monitoring networks have been widely investigated in the groundwater field. These design optimisations depend solely on spatial measurements and no temporal correlations are considered. Most commonly, Kriging is used for estimation and prediction while genetic algorithms and simulated annealing combined with other methods are widely used for optimisation of monitoring networks; see Brus and Heuvelink [2007], Cameron and Hunter [2000], Reed et al. [2001, 2000], Romary et al. [2014], Yeh et al. [2006], Zhu and Stein [2006].

A methodology adopted by several authors for spatial network optimisation is to incorporate minimising uncertainty about the covariance parameters into the optimisation along with minimising prediction uncertainty. This has been tackled for spatial design in a general sense, rather than with a particular application, by Romary et al. [2014], Zhu and Stein [2005, 2006].

In the Bayesian framework, Nowak et al. [2010] build on the idea of Bayesian geostatistical design, introduced by Diggle and Lophaven [2006], by transferring the concept into geostatistical inverse problems. They incorporate reducing the uncertainty of the covariance parameters in the geostatistical model into the design as a secondary objective. The primary objective is to minimise the expected Bayesian prediction variance. By using a Matérn covariance function, the covariance shape uncertainty is accounted for with the additional shape parameter contained in the Matérn functions structure.

More recently, methodologies developed for spatial design optimisation have been extended into the spatio-temporal setting. The extension of these methods can be implemented with relative ease but there are some important distinctions. Variation between observations in space is likely to be different to variation in time. Also, the cost of sampling at a single site several times can be less than sampling from the same number of wells but in different locations; see Heuvelink et al. [2012]. In the statistical framework, several authors have proposed optimal design methodologies for spatio-temporal processes in general, rather than for specific studies of interest; see Mateu and Muller [2013]. In line with spatial designs, Kriging methods are frequently used to estimate variances and are also used to estimate the state of the groundwater; see Nunes, Paralta, Cunha and Ribeiro [2004].

Incorporating the uncertainty around the covariance parameters into the design is utilised in the spatio-temporal setting by Bohorquez et al. [2016] who propose a dynamic procedure for optimising spatial networks by exploiting the idea that the spatial covariance structure varies through time. Their approach is attractive as it only depends on the covariance structure of the data and not an underlying spatio-temporal process. Using historical sampling data from a site, the spatial covariance parameters are estimated. A multivariate time series is then fitted to the parameter estimates, and in turn the design used in the next period or periods can be adapted. Forecasts of the covariance parameters at a future time are then made. They consider two design objectives; optimal mean estimation using generalised least squares, and optimal prediction using ordinary Kriging (see Section 2.4). The designs are constructed with the aim of minimising the variance associated with the chosen objective. An example using air pollution monitoring networks is presented.

Cameron and Hunter [2002] propose an optimisation consisting of two algorithms for groundwater monitoring networks, through reducing spatial and temporal redundancy. The first algorithm combines time series of data from wells to construct a composite temporal variogram which in turn is used to determine sampling frequencies. In the second, global Kriging weights are assigned to well locations in the monitoring network to ascertain their relative contribution to the contaminant plume map. The least influential wells are then removed. However, the paper does not account for spacetime correlation of the contaminant. Nunes, Cunha and Ribeiro [2004] also utilise minimising temporal redundancy whilst simultaneously minimising the spatial estimation error variance in a space-time approach.

Combining the statistical and model-based approaches was briefly discussed earlier. This can be done by first using a partial differential equations model to estimate the location of the contaminant plume. In practice, Darcy's Law (Whitaker [1986]) is regularly used to model the groundwater field. Statistical methods are then used to optimise a sampling strategy. Generally, these methodologies aim to minimise the uncertainty associated with the PDE model parameters as part of their objective function; see Helle and Pebesma [2012].

Reed et al. [2000] identify cost-effective sampling plans along with the total contaminant mass through simulations from a transport model combined with a genetic algorithm to optimise spatial designs. They interpolate the contaminant cloud using a combination of inverse distance weighting along with ordinary Kriging to obtain the contaminant mass. For their study, the objective function to be minimised is the cost combined with estimated mass error. Wu et al. [2005] extend the methodology by Reed et al. [2000], by introducing additional constraints on the optimisation based on second and third moments of a three-dimensional contaminant plume.

Meyer et al. [1994], also working with spatial designs, consider the three objectives: minimise the number of wells, maximise the probability of detecting a leak and minimise the projected area of contamination at the time of detection. They conduct uncertainty analysis using Monte Carlo simulations with random hydraulic conductivity values and contaminant source locations and then solve the multi-objective integer programming problem using simulated annealing.

Spatio-temporal designs of this nature are illustrated in two pieces of work by Herrera, Pinder and Zhang (Herrera and Pinder [2005], Zhang et al. [2005]). They develop a methodology similar to that of Montas et al. [2000] who also use simulations of the contaminant plume to optimise networks over space and time with the aim of minimising plume characterisation errors whilst limiting the number of wells that can be sampled.

In their first paper, Herrera and Pinder [2005] use a Kalman filter with a space-time covariance matrix obtained from Monte Carlo simulations of a stochastic transport model. To determine the location and timings of wells to be sampled, a function of the predicted estimate and its error variance is used. A sequential procedure is then used to select wells that minimise the value of this function until a predetermined value is reached. In their examples, they use the total variance of the estimate error and the coefficient of variation as their functions to be minimised. This is similar to the variance reduction analysis approach of Rouhani [1985]. They finish with a post-processing step where the Kalman filter is used to update the contaminant concentration estimate and its uncertainty. Although it is assumed in their work that the monitoring network is fixed, the network that is adopted is made up of a grid of equally spaced wells. In reality, this is an unlikely configuration due to transport infrastructure.

A similar approach is adopted in the second paper by Zhang et al. [2005] where they consider the uncertainty associated with the contaminant concentration field. The primary objective of this study is to minimise the cost. A Kalman filter is again used with a space-time covariance matrix obtained from simulations from the groundwater flow and transport model. To obtain the sampling times and locations, a genetic algorithm is used, combined with a Kalman filter for updating the covariance matrix. Andricevic [1990] use a similar methodology in their paper, with the Kalman filter also being used to update their covariance matrix however, they use a branch and bound algorithm for design optimisation.

Several authors have used sequential simulations of the flow field, in a similar manner to Herrera and Pinder [2005], in an attempt to estimate the uncertainty associated with the parameters of the flow field; see Bierkens [2006], Chadalavada [2008], Nunes et al. [2013], Nunes, Paralta, Cunha and Ribeiro [2004].

Optimisation of the sampling design can either be done in advance of the study, examples of which are presented in Herrera and Pinder [2005] described above, or alternatively a dynamic approach can be utilised; see Chadalavada [2008], Zhang et al. [2005].

The majority of this research aims to select wells from a dense or regularly spaced grid of potential locations. Here, networks which are already in place will be considered in a similar manner to Bohorquez et al. [2016], Cameron and Hunter [2002], Nunes, Cunha and Ribeiro [2004], Nunes, Paralta, Cunha and Ribeiro [2004].

6.1.1 Generic design optimalities extended to geostatistics

From the review given in Section 6.1, it is apparent that, most commonly, environmental design optimality objective functions are stated as a minimisation task of some scalar function, ϕ, of the sampling locations or model parameters such as the prediction covariance matrix; see Mateu and Muller [2013].

Reverting back to design in a single dimension, frequently the objective scalar function to be minimised is applied to the Fisher information matrix,

$$\mathbf{M}(Y, \mathbf{s}, \boldsymbol{\alpha}) = \mathbb{E}\left[\left\{\frac{\partial}{\partial \alpha} \log p(Y|\boldsymbol{\alpha})\right\}^2\right] \tag{6.1}$$

where $\log p(Y|\boldsymbol{\alpha})$ is the log likelihood function, $\mathbf{s}$ are the locations of the observations and $\boldsymbol{\alpha}$ are the model parameters. Through the Cramer-Rao inequality, it can be shown that the inverse of $\mathbf{M}$ is a lower bound for the conditional covariance matrix of the model parameters, $\boldsymbol{\alpha}$. In the case of normality, $\mathbf{M}$ is also the precision matrix i.e. the inverse of the covariance matrix of the model parameters conditional on the data; see Nowak [2010]. Therefore, while maximising a function of the information matrix is a popular choice, minimisation of a function of the covariance matrix associated with parameter accuracy can also be used.

Geostatistical design is primarily focused on accurate estimation at unmeasured locations rather than on model parameter accuracy. Thus, most commonly the covariance matrix associated with prediction accuracy, namely the Kriging variance matrix, $\mathbf{C}_{\mathbf{s}_0|\mathbf{y}}$, (see Chapter 2) is used; see Zimmerman and Li [2013].

Building on the modelling methodology and notation for spatial and spatio-temporal data for spline-based models (see Chapter 2), the covariance matrix of the basis coefficients, $\mathbf{C}_{\hat{\alpha}|y}$, can be used. Nowak [2010] discusses the relationship between design optimalities for classical regression design problems and geostatistical design problems. In the majority of cases, the parameter covariance matrix can be replaced by the prediction covariance matrix. Thus, for notational simplicity, the covariance matrix of interest will be referred to collectively as $\mathbf{C}$ hereafter; its dimensions are $m \times m$.

Outlined below is a summary of some of the more commonly used design criteria, extended from the traditional regression-like context into spatial and spatio-temporal domains, based on the work of Nowak [2010].

Design Objective Functions based on the Covariance Matrix

In 1959, Kiefer and Wolfowitz [1959] presented the concept of alphabetic optimalities with the introduction of D- and E- optimal designs for regression estimation problems. Their work has been significantly added to, with optimalities such as the A-, C- and T- now being used widely within the statistical literature. Traditional regression based designs assume observations are collected with independent errors. This assumption is clearly violated for spatial and spatio-temporal data and thus adjustments need to be made; see Mateu and Muller [2013]. Classical regression based designs focus primarily on minimising the uncertainty around the parameters being estimated in the models. In contrast, designs for spatial and spatio-temporal data mainly focus on minimising the uncertainty associated with predictions at unmeasured locations; see Le and Zidek [2006].

- **A - Optimality** aims to minimise the quadratic penalty function:

$$(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_{\text{true}})^{\top}\mathbf{A}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_{\text{true}})$$

  which is equivalent to minimising the average parameter estimation variance. The resulting function to be minimised is:

$$\phi_A = \frac{1}{m}\mathrm{tr}[\mathbf{AC}] \tag{6.2}$$

where **A** a non-negative definite matrix and m is the dimension of **C**.

- **D - Optimality** aims to minimise:

$$\phi_D = \det[\mathbf{C}]^{1/m} = \prod_{j=1}^{m} \mathrm{eig}_j(\mathbf{C})^{1/m} \tag{6.3}$$

  The logarithm of this function is also widely used. This is the most common objective function for classical regression design problems; however, in dense sampling networks the computational expense is significantly large. This measure is also very sensitive to a single highly informative observation; see Nowak [2010].

- **E - Optimality** minimises:

$$\phi_E = \max(\mathrm{eig}(\mathbf{C})) \tag{6.4}$$

  The primary role of the E- criterion is to assess whether large-scale variability has been removed. The computational cost of this measure is not as extreme as that of D- due to the computation time of the largest eigenvalue being significantly less than the computation time of the full set.

- **G - Optimality** aims to minimise the maximum prediction variance over the design region. The criterion to be minimised is:

$$\phi_G = \max[\mathbf{C}] \tag{6.5}$$

- **P - Optimality** provides a generalisation of the A-, D- and E- optimalities, with the criteria to be minimised being:

$$\phi_P = \left[\sum_{j=1}^{m} \mathrm{eig}_j\left(C^P\right)\right]^{1/P} \tag{6.6}$$

  When $P = 1, 0$ and ∞, P-optimality becomes the A-, D- and E- optimalities respectively. However, for large datasets this measure is extremely expensive to compute due to it requiring all eigenvalues to be computed.

- **Average estimation variance** exploits the fact that estimation variance is the value of the conditional covariance when the distance between locations is 0. This corresponds to the diagonal elements of $\mathbf{C}$. The average estimation variance which is to be minimised is then given as:

$$\overline{\sigma_E^2} = \frac{1}{m}\text{tr}(\mathbf{C}). \tag{6.7}$$

  This is equivalent to the A- criterion when the matrix $\mathbf{A}$ is the identity matrix $(\mathbf{A} = \mathbf{I})$. This is also known as the AI optimality, denoted as ϕ_{AI}.

Other Design Objective Functions

In the measures for optimisation discussed earlier, the main aim was to achieve the best predictive accuracy possible. Studies whose aims are primarily exploratory may want to choose an objective function that gives good coverage of the study region and that are space filling in nature; see Mateu and Muller [2013].

- **Minimax Distance Design** aims to minimise the maximum distance between a given unsampled point $\mathbf{s}_0$ and its closest point in a given design $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. The criterion to be minimised over the study region, D, is therefore:

$$\phi_{mM} = \max_{\mathbf{s}_0 \in D} \min_{\mathbf{s}_i} ||\mathbf{s}_0 - \mathbf{s}_i||. \tag{6.8}$$

6.2 Two alternative objective functions

The remainder of this chapter will focus on the derivation and application of two alternative objective functions for optimising groundwater sampling designs, namely the Variance of the estimated plume Mass (VM) and the Integrated Variance of the prediction (IV).

Spatial moments of the contaminant plume, for which variance is the second, have been widely used in the groundwater design context. Many studies aim to minimise errors or uncertainty associated with contaminant plume characteristics, such as the mass, as part of their design optimisation. Reed et al. [2000] use the global mass estimation error as a constraint in their optimisation of a cost function. In a later paper Reed and Minsker [2004] again make use of the global mass estimate; with maximisation of the accuracy of the contaminant mass estimate being used as one of several objective functions. Montas et al. [2000] minimise the error of three plume characteristics as their design objective function, one of which is the contaminant mass error. Wu et al. [2005] look to estimate global three-dimensional plumes using the first three moments and Chadalavada [2008] calculate the pollution mass estimate error in their paper, however this is not used explicitly as a design objective function. Often the errors associated with these plume characteristics are computed by comparing the spatial moments to the true moments taken from stochastic groundwater flow simulations. Estimation of these errors are however, constrained by the need for a flow and transport model of the study site, this can be complex to estimate.

Objective functions founded on minimising the variance or error of the estimated state of the groundwater in a region i.e. minimise the uncertainty of the prediction, are also very popular in the groundwater design context. Wagner [1995] seek to minimise the trace of the prediction covariance matrix as their design objective function. Herrera and Pinder [2005] minimise a function of the error variance of the concentration estimate, whereas Bohorquez et al. [2016] minimise prediction error. Dhar [2013] aim to minimise the maximum normalised absolute deviation between the estimated and observed concentrations at unmeasured locations. See Zimmerman and Li [2013] for more design objective function based on the prediction variance.

Minimisation of the integrated prediction variance, which is to be used here, is an already well established criterion in the classical linear model framework, often referred to as IV-optimal. In a geostatistical setting, Diggle and Lophaven [2006] use a similar criterion in a Bayesian paradigm for assessing whether sampling locations can be removed or added to a network. Their methodology is based on Kriging and is applied solely to spatial models, a variation of this criterion is also used for considering a new sampling network.

As discussed, variations of both of the proposed objective functions have been widely used in monitoring network design for groundwater data. However almost all of these pieces of work use Kriging-based models to interpolate the plume and obtain the variances associated with the errors to compute the desired objective functions, for example

Bohorquez et al. [2016], Cameron and Hunter [2000], Chadalavada [2008], Dhar [2013], Diggle and Lophaven [2006], Reed and Minsker [2004], Reed et al. [2000]). In contrast, here, spatial and spatio-temporal p-spline models will be used to calculate the objective functions.

The following sections will derive each objective function for Kriging and p-spline models in a spatial and spatio-temporal framework. The results section will then go onto apply both objective functions using spatial and spatio-temporal p-spline models to a groundwater monitoring network, with spatial Kriging also being used to optimise the functions to give a comparison to the currently most widely used spatial and spatio-temporal interpolation method.

6.3 Variance of the Plume Mass (VM) Objective Function

For the VM criterion, the optimal sampling design is found by minimising,

$$\phi_{VM} = \text{var}\left(\int_{x_1}\int_{x_2} \hat{y}(x_1,x_2)\ dx_2dx_1\right) \tag{6.9}$$

where $\int_{x_1}\int_{x_2} \hat{y}(x_1,x_2)\ dx_2dx_1$ is the estimated plume mass, computed by integrating over the spatial study region. From here in, $\hat{y}(x_1,x_2)$ will be denoted $\hat{y}$.

6.3.1 VM Objective Function - Spatial Models

Kriging

In section 2.4.3 the Kriging predictor for a spatial process at new location, $\mathbf{s}_0$, given observed data, $\mathbf{y} = (y(\mathbf{s}_1), \cdots y(\mathbf{s}_n))$, was shown as

$$\mathbb{E}[y(\mathbf{s}_0)|\mathbf{y}] = \hat{\mu}_y + \mathbf{c}_0^\top \mathbf{K}^{-1}(\mathbf{y} - \hat{\mu}_y\mathbf{1}), \tag{6.10}$$

where, $\mathbf{c}_0 = (C_y(||\mathbf{s}_0 - \mathbf{s}_1||;\boldsymbol{\theta}), \ldots, C_y(||\mathbf{s}_0 - \mathbf{s}_n||;\boldsymbol{\theta}))$ is the covariance between the new location and observed locations; $\mathbf{K}_{ij} = C_y(||\mathbf{s}_i - \mathbf{s}_j||;\boldsymbol{\theta})$ is the covariance between the observed locations; $\hat{\mu}_y$ is the estimated constant mean; $\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2)$ are the covariance model parameters, which can be estimated by maximum likelihood, and $C_y()$ is a covariance function (see Section 2.4.2). This predictor function can be denoted alternatively as

$$\hat{y}(\mathbf{s}_0) = \hat{\mu}_y + \sum_{i=1}^{n} \alpha_i (y_i - \hat{\mu}_y), \tag{6.11}$$

where $\alpha_i = (\mathbf{K}^{-1})_{i\bullet}^{\top}\mathbf{c}_0$, known as the Kriging weight of the i^{th} observation, is the weight of contribution of the i^{th} observation on the prediction at new location, $\mathbf{s}_0$. Using this formulation the mass of the prediction can be computed by integrating over the spatial region i.e.

$$\begin{aligned}\int_{x_1}\int_{x_2} \hat{y}\ dx_2 dx_1 =& \int_{x_1}\int_{x_2} \left(\hat{\mu}_y + \sum_{i=1}^{n} \alpha_i (y_i - \hat{\mu}_y)\right)\ dx_2 dx_1 \\ =& \int_{x_1}\int_{x_2} \hat{\mu}_y\ dx_2 dx_1 + \sum_{i=1}^{n} (y_i - \hat{\mu}_y) \int_{x_1}\int_{x_2} \alpha_i\ dx_2 dx_1 \\ =& M + \sum_{i=1}^{n} (y_i - \hat{\mu}_y) a_i\end{aligned} \tag{6.12}$$

where $M = (\mathrm{mx}_1\mathrm{mx}_2 - \mathrm{mx}_2\mathrm{mn}_1 - \mathrm{mn}_2\mathrm{mx}_1 + \mathrm{mn}_1\mathrm{mn}_2)\hat{\mu}_y$ is a constant with:

$$\begin{aligned} \mathrm{mx}_1 &= \max(x_1) \qquad & \mathrm{mn}_1 &= \min(x_1) \\ \mathrm{mx}_2 &= \max(x_2) \qquad & \mathrm{mn}_2 &= \min(x_2) \end{aligned}$$

and

$$\begin{aligned} a_i =& \int_{x_1}\int_{x_2} \alpha_i\ dx_2 dx_1 \\ =& (\mathbf{K}^{-1})_{i\bullet}^{\top} \left[\int_{x_1}\int_{x_2} \mathbf{c}_0^1\ dx_2 dx_1 \cdots \int_{x_1}\int_{x_2} \mathbf{c}_0^n\ dx_2 dx_1\right]. \end{aligned} \tag{6.13}$$

Computation of the integral, $\int_{x_1}\int_{x_2} \mathbf{c}_0^i\ dx_2 dx_1$, is dependent on the chosen covariance function. A closed form expression is not always available, for example in the case of a Matérn function, and thus numerical integration is needed.

The objective function using a spatial Kriging model can then be computed by making use of the rule for the variance of a linear combination of variables detailed in Appendix A.2, as

$$\begin{aligned}
\phi_{VM} =&\text{var}\left(\int_{x_1}\int_{x_2}\hat{y}\ dx_2dx_1\right)\\
=&\text{var}\left(M+\sum_{i=1}^{w}y_ia_i-\sum_{i=1}^{w}\hat{\mu}_ya_i\right)\\
=&\text{var}\left(\sum_{i=1}^{w}y_ia_i\right)\\
=&\sum_{i=1}^{w}\sum_{j=1}^{w}a_ia_j\text{cov}(y_i,y_j)\\
=&\sum_{i=1}^{w}\sum_{j=1}^{w}a_ia_j\mathbf{K}_{ij}\\
=&\boldsymbol{a}^\top\mathbf{K}\boldsymbol{a}=\text{tr}\left(\boldsymbol{a}\boldsymbol{a}^\top\mathbf{K}\right).
\end{aligned}\tag{6.14}$$

This is similar to the A-optimality objective function shown in Equation 6.2, with $\mathbf{A}=\boldsymbol{a}\boldsymbol{a}^\top$.

The optimisation procedure seeks to find, for a fixed set size w, the set of wells which minimise the expression in Equation 6.14. Historical data are used to estimate the covariance function parameters $\boldsymbol{\theta}$ and $\hat{\mu}_y$.

P-splines

The objective function using a p-splines model for spatial prediction can be similarly derived. Using the formula for the fitted value at location i,

$$\hat{y}_i=\sum_{j=1}^{m_1}\sum_{k=1}^{m_2}\hat{\alpha}_{jk}B_j(x_{1i})B_k(x_{2i}),\tag{6.15}$$

where m_1 and m_2 are the number of basis functions, $B_j(x_{1i})$ and $B_k(x_{2i})$, in each dimension respectively, with corresponding coefficient $\hat{\alpha}_{jk}$. The mass of the predicted surface can then be computed as

$$\begin{aligned}\int_{x_1}\int_{x_2}\hat{y}\ dx_2dx_1 &= \int_{x_1}\int_{x_2}\sum_{j=1}^{m_1}\sum_{k=1}^{m_2}\hat{\alpha}_{jk}B_j(x_1)B_k(x_2)\ dx_2dx_1 \\ &= \sum_{j=1}^{m_1}\sum_{k=1}^{m_2}\hat{\alpha}_{jk}\underbrace{\int_{x_1}B_j(x_1)\ dx_1}_{a_{1j}}\underbrace{\int_{x_2}B_k(x_2)\ dx_2}_{a_{2k}}\end{aligned} \tag{6.16}$$

where vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ contain the area under each basis function in each dimension respectively. Each entry in the vector will be the same with the exception of the first p and last p entries, where p is the degree of the basis function, this is illustrated in Figure 5.6 of Chapter 5. These vectors will also be identical for models with the same number of basis functions in each dimension.

In the model construction, the Kronecker product is used to obtain the structure required to model the spatial surface. This methodology can also be applied here. The Kronecker product of the two vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ is denoted as $\boldsymbol{a} = \boldsymbol{a}_1 \otimes \boldsymbol{a}_2$. Thus Equation 6.16 can be re-expressed as:

$$\begin{aligned}\int_{x_1}\int_{x_2}\hat{y}\ dx_2dx_1 &= \sum_{j=1}^{m_1}\sum_{k=1}^{m_2}\hat{\alpha}_{jk}\underbrace{\int_{x_1}B_j(x_1)\ dx_1}_{a_{1j}}\underbrace{\int_{x_2}B_k(x_2)\ dx_2}_{a_{2k}} \\ &= \sum_{l=1}^{M}\hat{\alpha}_l a_l\end{aligned} \tag{6.17}$$

where $M = m_1 \times m_2$.

Thus, utilising the variance of the sum of a linear combination of variables detailed in Appendix A.2, the objective function can be expressed as:

$$\begin{aligned}\phi_{VM} &= \text{var}\left(\int_{x_1}\int_{x_2}\hat{y}\ dx_2dx_1\right) \\ &= \text{var}\left(\sum_{l=1}^{M}\hat{\alpha}_l a_l\right) \\ &= \sum_{p=1}^{M}\sum_{q=1}^{M}a_p a_q \text{cov}(\hat{\alpha}_p, \hat{\alpha}_q) \\ &= \boldsymbol{a}^\top \mathbf{C}_{\hat{\alpha}|y}\boldsymbol{a} = \text{tr}\left(\boldsymbol{a}\boldsymbol{a}^\top \mathbf{C}_{\hat{\alpha}|y}\right)\end{aligned} \tag{6.18}$$

where $\mathbf{C}_{\hat{\alpha}|y}[p,q] = \text{cov}(\hat{\alpha}_p, \hat{\alpha}_q)$. Again, this is similar to the A-optimality objective function with $\mathbf{A} = \boldsymbol{a}\boldsymbol{a}^\top$. The optimisation procedure seeks to find, for a fixed set size w, the set of wells which minimise the expression in Equation 6.18.

6.3.2 VM Objective Function - Spatio-temporal Models

The objective function can be computed for spatio-temporal Kriging and p-spline models in a similar manner to that of the spatial models. In the spatio-temporal setting the next sampling time is assumed to be known and fixed. The mass is only computed over the spatial domain with the temporal dimension being used to add weight from earlier observations depending on their proximity to the proposed next sampling time.

Kriging

In Section 2.4.5 the Kriging predictor for a spatio-temporal process at new spatial location, $\mathbf{s}_0$, at time, t_0, given observed data, $\mathbf{y} = (y(\mathbf{s}_1, t_1), ..., y(\mathbf{s}_n, t_m))^\top$, was shown to be:

$$\mathbb{E}[y(\mathbf{s}_0, t_0)|\mathbf{y}] = \hat{\mu}_y + \mathbf{c}_0^\top \mathbf{K}^{-1}(\mathbf{y} - \hat{\mu}_y \mathbf{1}), \tag{6.19}$$

where, assuming a separable covariance structure, $\mathbf{c}_0 = (C_y(t_0 - t_1; \boldsymbol{\theta}_t) \otimes C_y(\mathbf{s}_0 - \mathbf{s}_1; \boldsymbol{\theta}_s), \ldots, C_y(t_0 - t_m; \boldsymbol{\theta}_t) \otimes C_y(\mathbf{s}_0 - \mathbf{s}_n; \boldsymbol{\theta}_s))$ is the covariance between the new location and the observed locations; $\mathbf{K} = (\boldsymbol{\Sigma}^{(t)}) \otimes (\boldsymbol{\Sigma}^{(s)})$ is the covariance between the observed locations; $\hat{\mu}_y$ is the estimated constant mean; $\boldsymbol{\Sigma}^{(s)}_{ij} = C_y(||\mathbf{s}_i - \mathbf{s}_j||; \boldsymbol{\theta}_s)$ is the spatial covariance matrix; $\boldsymbol{\theta}_s = (\sigma^2_s, \tau^2_s, \phi_s)$ are the spatial covariance parameters; $\boldsymbol{\Sigma}^{(t)}_{ij} = C_y(||t_i - t_j||; \boldsymbol{\theta}_t)$ is the temporal covariance matrix; $\boldsymbol{\theta}_t = (\sigma^2_t, \tau^2_t, \phi_t)$ are the temporal covariance parameters and $C_y()$ is a covariance function (see Section 2.4.2). For notational ease, here the case where every spatial location is sampled at every time point is presented. When this is not the case exploitation of the Kronecker products cannot be used and element wise matrix multiplication is required. In a similar way to the spatial Kriging model, the prediction function can be denoted alternatively as

$$\hat{y}(\mathbf{s}_0, t_0) = \hat{\mu}_y + \sum_{i=1}^{n} \alpha_i (y_i - \hat{\mu}_y) \tag{6.20}$$

where $\alpha_i = (\mathbf{K}^{-1})_{i\bullet}^{\top}\mathbf{c}_0$, known as the Kriging weight of the i^{th} observation, is the weight of contribution of the i^{th} observation to the prediction. Using this formulation the mass of the prediction can be computed by integrating over the spatial region.

$$\begin{aligned}
\int_{x_1}\int_{x_2} \hat{y}\ dx_2 dx_1 =& \int_{x_1}\int_{x_2} \left(\hat{\mu}_y + \sum_{i=1}^{n} \alpha_i (y_i - \hat{\mu}_y)\right)\ dx_2 dx_1 \\
=& \int_{x_1}\int_{x_2} \hat{\mu}_y\ dx_2 dx_1 + \sum_{i=1}^{n} (y_i - \hat{\mu}_y) \int_{x_1}\int_{x_2} \alpha_i\ dx_2 dx_1 \\
=& M + \sum_{i=1}^{n} (y_i - \hat{\mu}_y) a_i
\end{aligned} \tag{6.21}$$

where $M = (\mathrm{mx}_1\mathrm{mx}_2 - \mathrm{mx}_2\mathrm{mn}_1 - \mathrm{mn}_2\mathrm{mx}_1 + \mathrm{mn}_1\mathrm{mn}_2)\hat{\mu}_y$ is a constant with:

$$\begin{aligned}
\mathrm{mx}_1 &= \max(x_1) & \mathrm{mn}_1 &= \min(x_1) \\
\mathrm{mx}_2 &= \max(x_2) & \mathrm{mn}_2 &= \min(x_2)
\end{aligned}$$

and

$$\begin{aligned}
a_i =& \int_{x_1}\int_{x_2} \alpha_i\ dx_2 dx_1 \\
=& (\mathbf{K}^{-1})_{i\bullet}^{\top} \left[\int_{x_1}\int_{x_2} \mathbf{c}_0^1\ dx_2 dx_1 \cdots \int_{x_1}\int_{x_2} \mathbf{c}_0^n\ dx_2 dx_1 \right].
\end{aligned} \tag{6.22}$$

Each entry of the vector of integrals can be computed as:

$$\int_{x_1}\int_{x_2} \mathbf{c}_0^i\ dx_2 dx_1 = C_y(t_0 - t_i; \boldsymbol{\theta}_t) \int_{x_1}\int_{x_2} C_y(\mathbf{s}_0 - \mathbf{s}_i; \boldsymbol{\theta}_s)\ dx_2 dx_1. \tag{6.23}$$

Computation of this integral depends on the specified covariance function. A closed form expression cannot always be achieved, for example in the case of a Matérn covariance function, and thus numerical integration is needed to obtain an estimate.

The objective function to be minimised using a spatio-temporal Kriging model is then:

$$\begin{aligned}
\phi_{VM} =& \text{var}\left(\int_{x_1}\int_{x_2} \hat{y}\ dx_2 dx_1\right) \\
=& \text{var}\left(M + \sum_{i=1}^{w} y_i a_i - \sum_{i=1}^{w} \mu_y a_i\right) \\
=& \text{var}\left(\sum_{i=1}^{w} y_i a_i\right) \\
=& \sum_{i=1}^{w}\sum_{j=1}^{w} a_i a_j \text{cov}(y_i, y_j) \\
=& \sum_{i=1}^{w}\sum_{j=1}^{w} a_i a_j \mathbf{K}_{ij} \\
=& \boldsymbol{a}^\top \mathbf{K} \boldsymbol{a} = \text{tr}\left(\boldsymbol{a}\boldsymbol{a}^\top \mathbf{K}\right).
\end{aligned} \tag{6.24}$$

The optimisation procedure seeks to find, for a fixed set size, w, and new sampling time, t_0, the set of wells which minimise the expression in Equation 6.24. Historical data are used to estimate the covariance function parameters, $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_t$ and $\hat{\mu}_y$.

P-splines

The VM objective function using a spatio-temporal p-splines model can be derived using the formula for a fitted value at location i,

$$\hat{y}_i = \sum_{j,k,l=1}^{m_1,m_2,m_3} \hat{\alpha}_{jkl} B_j(x_{1i}) B_k(x_{2i}) B_l(t_i), \tag{6.25}$$

where m_1, m_2 and m_3 are the number of basis functions, $B_j(x_{1i})$, $B_k(x_{2i})$ and $B_l(t_i)$, in each dimension respectively, with corresponding basis coefficient $\hat{\alpha}_{jkl}$. The spatial mass of the predicted surface can subsequently be computed as

$$\begin{aligned}
\int_{x_1}\int_{x_2} \hat{y}\ dx_2 dx_1 =& \int_{x_1}\int_{x_2} \sum_{j,k,l} \hat{\alpha}_{jkl} B_j(x_1) B_k(x_2) B_l(t)\ dx_2 dx_1 \\
=& \sum_{j,k,l} \hat{\alpha}_{jkl}\, \underbrace{B_l(t)}_{b_l} \underbrace{\int_{x_1} B_j(x_1)\ dx_1}_{a_{1j}} \underbrace{\int_{x_2} B_k(x_2)\ dx_2}_{a_{2k}}
\end{aligned} \tag{6.26}$$

where vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ contain the area under each basis function in each spatial dimension respectively and vector $\mathbf{b}$ contains the result of evaluating each temporal basis function at time t i.e. the time at which the mass is to be computed.

In the model construction, Kronecker products are used to obtain the structure required to model the spatio-temporal surface. This methodology can also be applied here. The Kronecker product of the three vectors; $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\mathbf{b}$ can be denoted as $\boldsymbol{a} = \boldsymbol{a}_1 \otimes \boldsymbol{a}_2 \otimes \mathbf{b}$; thus, Equation 6.26 can be re-expressed as

$$\begin{aligned} \int_{x_1}\int_{x_2} \hat{y}\ dx_2 dx_1 =& \sum_{j,k,l} \hat{\alpha}_{jkl} \underbrace{B_l(t)}_{b_l} \underbrace{\int_{x_1} B_j(x_1)\ dx_1}_{a_{1j}} \underbrace{\int_{x_2} B_k(x_2)\ dx_2}_{a_{2k}} \\ =& \sum_{m=1}^{M} \hat{\alpha}_m a_m \end{aligned} \tag{6.27}$$

where $M = m_1 \times m_2 \times m_3$.

Therefore, the objective function to be minimised is:

$$\begin{aligned} \phi_{VM} =& \text{var}\left(\int_{x_1}\int_{x_2} \hat{y}\ dx_2 dx_1\right) \\ =& \text{var}\left(\sum_{m=1}^{M} \hat{\alpha}_m a_m\right) \\ =& \sum_{p=1}^{M}\sum_{q=1}^{M} a_p a_q \text{cov}\left(\hat{\alpha}_p, \hat{\alpha}_q\right) \\ =& \boldsymbol{a}^\top \mathbf{C}_{\hat{\alpha}|y} \boldsymbol{a} = \text{tr}\left(\boldsymbol{a}\boldsymbol{a}^\top \mathbf{C}_{\hat{\alpha}|y}\right) \end{aligned} \tag{6.28}$$

where $\mathbf{C}_{\hat{\alpha}|y}[p,q] = \text{cov}\left(\hat{\alpha}_p, \hat{\alpha}_q\right)$. The optimisation procedure seeks to find, for a fixed set size w and new sampling time t_0, the set of wells which minimise the expression in Equation 6.28.

6.4 Integrated Prediction Variance (IV) Objective Function

For the IV criterion, the optimal design is found by minimising,

$$\phi_{IV} = \int_{x_1}\int_{x_2} \text{var}(\hat{y}(x_1,x_2))\ dx_2dx_1, \tag{6.29}$$

where again, the integral is computed only over the spatial dimensions. From here in $\hat{y}(x_1,x_2)$ is denoted by $\hat{y}$.

6.4.1 IV Objective Function - Spatial Models

Kriging

Using the spatial Kriging model notation in Section 6.3.1, the integrated variance can be expressed as follows:

$$\begin{aligned}
\int_{x_1}\int_{x_2} \text{var}(\hat{y})\ dx_2dx_1 =& \int_{x_1}\int_{x_2} \text{var}\left(\hat{\mu}_y + \sum_{i=1}^{n}\alpha_i(y_i-\hat{\mu}_y)\right)\ dx_2dx_1 \\
=& \int_{x_1}\int_{x_2} \text{var}\left(\sum_{i=1}^{n}\alpha_i y_i\right)\ dx_2dx_1 \\
=& \int_{x_1}\int_{x_2} \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j \text{cov}(y_i,y_j)\ dx_2dx_1 \\
=& \sum_{i=1}^{n}\sum_{j=1}^{n}\text{cov}(y_i,y_j)\int_{x_1}\int_{x_2}\alpha_i\alpha_j\ dx_2dx_1.
\end{aligned} \tag{6.30}$$

Denoting $(\mathbf{K}^{-1})_{i\bullet} = \mathbf{m}_i^\top$, the integral of the product of the Kriging weights can be decomposed as follows:

$$\begin{aligned}
\int_{x_1}\int_{x_2}\alpha_i\alpha_j\ dx_2dx_1 =& \int_{x_1}\int_{x_2}(\mathbf{K}^{-1})_{i\bullet}^\top\mathbf{c}_0\cdot(\mathbf{K}^{-1})_{j\bullet}^\top\mathbf{c}_0\ dx_2dx_1 \\
=& \int_{x_1}\int_{x_2}\mathbf{m}_i^\top\mathbf{c}_0\cdot\mathbf{m}_i^\top\mathbf{c}_0\ dx_2dx_1 \\
=& \int_{x_1}\int_{x_2}\sum_k \mathbf{m}_{ik}\mathbf{c}_{0k}\cdot\sum_l \mathbf{m}_{jl}\mathbf{c}_{0l}\ dx_2dx_1 \\
=& \sum_k\sum_l \mathbf{m}_{ik}\mathbf{m}_{jl}\int_{x_1}\int_{x_2}\mathbf{c}_{0k}\mathbf{c}_{0l}\ dx_2dx_1 \\
=& \sum_k\sum_l \mathbf{m}_{ik}\mathbf{m}_{jl}\mathbf{C}[k,l] \\
=& \mathbf{m}_i^\top\mathbf{C}\mathbf{m}_j,
\end{aligned} \tag{6.31}$$

where $\mathbf{C}[k,l] = \int_{x_1}\int_{x_2} \mathbf{c}_{0k}\mathbf{c}_{0l}\ dx_2dx_1$. Computation of this double integral is dependent on the chosen covariance function. A closed form expression is not always available, for example in the case of a Matérn covariance function, and thus numerical integration is needed.

The objective function to be minimised, using a spatial Kriging model, is then:

$$\begin{aligned}\phi_{IV} &= \int_{x_1}\int_{x_2} \text{var}(\hat{y})\ dx_2dx_1 \\ &= \sum_{i=1}^{w}\sum_{j=1}^{w} \text{cov}(y_i, y_j)\mathbf{m}_i^\top \mathbf{C}\mathbf{m}_j.\end{aligned} \tag{6.32}$$

The optimisation procedure seeks to find, for a fixed set size, w, the set of wells which minimise the expression in Equation 6.32. Historical data are used to estimate the covariance function parameters, $\boldsymbol{\theta}$.

P-splines

Using the spatial p-spline model discussed in Section 6.3.1, the integrated variance of the prediction can be derived as:

$$\begin{aligned}\int_{x_1}\int_{x_2} \text{var}(\hat{y})\ dx_2dx_1 &= \int_{x_1}\int_{x_2} \text{var}\left(\sum_{j=1}^{m_1}\sum_{k=1}^{m_2} \hat{\alpha}_{jk}B_j(x_1)B_k(x_2)\right)\ dx_2dx_1 \\ &= \int_{x_1}\int_{x_2} \sum_{j,l=1}^{m_1}\sum_{k,m=1}^{m_2} B_j(x_1)B_k(x_2)B_l(x_1)B_m(x_2)\ \text{cov}(\hat{\alpha}_{jk}, \hat{\alpha}_{lm})\ dx_2dx_1 \\ &= \sum_{j,l=1}^{m_1}\sum_{k,m=1}^{m_2} \text{cov}(\hat{\alpha}_{jk}, \hat{\alpha}_{lm}) \int_{x_1} B_j(x_1)B_l(x_1)\ dx_1 \int_{x_2} B_k(x_2)B_m(x_2)\ dx_2 \\ &= \text{tr}\left[\mathbf{C}_{\hat{\alpha}|y}(\widetilde{\mathbf{B}}_1 \otimes \widetilde{\mathbf{B}}_2)\right]\end{aligned} \tag{6.33}$$

where,

$$\widetilde{\mathbf{B}}_1[j,l] = \int_{x_1} B_j(x_1)B_l(x_1)\ dx_1 \qquad \text{and} \qquad \widetilde{\mathbf{B}}_2[k,m] = \int_{x_2} B_k(x_2)B_m(x_2)\ dx_2.$$

Thus the objective function to be minimised is,

$$\begin{aligned}\phi_{IV} =& \int_{x_1}\int_{x_2} \mathrm{var}\left(\hat{y}\right) \;\; dx_2 dx_1 \\ =& \mathrm{tr}\left[\mathbf{C}_{\hat{\alpha}|y}(\widetilde{\mathbf{B}}_1 \otimes \widetilde{\mathbf{B}}_2)\right],\end{aligned} \tag{6.34}$$

where $\mathbf{C}_{\hat{\alpha}|y}[p,q] = \mathrm{cov}(\hat{\alpha}_p, \hat{\alpha}_q)$. The optimisation procedure seeks to find, for a fixed set size w, the set of wells which minimise the expression in Equation 6.34.

6.4.2 IV Objective Function - Spatio-temporal Models

Kriging

Using the spatio-temporal Kriging model notation from Section 6.3.2, the integrated variance of the prediction can be computed:

$$\begin{aligned}\int_{x_1}\int_{x_2} \mathrm{var}(\hat{y}) \;\; dx_2 dx_1 =& \int_{x_1}\int_{x_2} \mathrm{var}\left(\hat{\mu}_y + \sum_{i=1}^{n} \alpha_i (y_i - \hat{\mu}_y)\right) \;\; dx_2 dx_1 \\ =& \int_{x_1}\int_{x_2} \mathrm{var}\left(\sum_{i=1}^{n} \alpha_i y_i\right) \;\; dx_2 dx_1 \\ =& \int_{x_1}\int_{x_2} \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \mathrm{cov}(y_i, y_j) \;\; dx_2 dx_1 \\ =& \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{cov}(y_i, y_j) \int_{x_1}\int_{x_2} \alpha_i \alpha_j \;\; dx_2 dx_1.\end{aligned} \tag{6.35}$$

Denoting $(\mathbf{K}^{-1})_{i\bullet} = \mathbf{m}_i^\top$, the integral of the product of the Kriging weights can be decomposed as follows:

$$\begin{aligned}\int_{x_1}\int_{x_2} \alpha_i \alpha_j \;\; dx_2 dx_1 =& \int_{x_1}\int_{x_2} (\mathbf{K}^{-1})_{i\bullet}^\top \mathbf{c}_0 \cdot (\mathbf{K}^{-1})_{j\bullet}^\top \mathbf{c}_0 \;\; dx_2 dx_1 \\ =& \int_{x_1}\int_{x_2} \mathbf{m}_i^\top \mathbf{c}_0 \cdot \mathbf{m}_j^\top \mathbf{c}_0 \;\; dx_2 dx_1 \\ =& \int_{x_1}\int_{x_2} \sum_k \mathbf{m}_{ik}\mathbf{c}_{0k} \cdot \sum_l \mathbf{m}_{jl}\mathbf{c}_{0l} \;\; dx_2 dx_1 \\ =& \sum_k \sum_l \mathbf{m}_{ik}\mathbf{m}_{jl} \int_{x_1}\int_{x_2} \mathbf{c}_{0k}\mathbf{c}_{0l} \;\; dx_2 dx_1 \\ =& \sum_k \sum_l \mathbf{m}_{ik}\mathbf{m}_{jl}\mathbf{C}[k,l] \\ =& \mathbf{m}_i^\top \mathbf{C} \mathbf{m}_j\end{aligned} \tag{6.36}$$

where

$$\begin{aligned}\mathbf{C}[k,l] =& \int_{x_1}\int_{x_2} \mathbf{c}_{0k}\mathbf{c}_{0l}\ dx_2dx_1 \\ =& C_y(t_0 - t_k;\boldsymbol{\theta}_t)C_y(t_0 - t_l;\boldsymbol{\theta}_t)\int_{x_1}\int_{x_2} C_y(\mathbf{s}_0 - \mathbf{s}_k;\boldsymbol{\theta}_s)C_y(\mathbf{s}_0 - \mathbf{s}_l;\boldsymbol{\theta}_s)\ dx_2dx_1.\end{aligned} \tag{6.37}$$

Computation of this double integral is dependent on the chosen covariance function. A closed form expression is not always available, for example in the case of a Matérn function, and thus numerical integration is needed.

The objective function to be minimised, using a spatio-temporal Kriging model, is then:

$$\begin{aligned}\phi_{IV} =& \int_{x_1}\int_{x_2} \text{var}(\hat{y})\ dx_2dx_1 \\ =& \sum_{i=1}^{w}\sum_{j=1}^{w} \text{cov}(y_i, y_j)\mathbf{m}_i^\top\mathbf{C}\mathbf{m}_j.\end{aligned} \tag{6.38}$$

The optimisation procedure seeks to find, for a fixed new sampling time t_0 and set size w, the set of wells which minimise the expression in Equation 6.38. Historical data are used to estimate the covariance function parameters, $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_t$, and $\hat{\mu}_y$.

P-splines

Using the formula for the fitted values of a spatio-temporal p-spline model from Section 6.3.2, the integrated variance of the prediction can be derived as:

$$\begin{aligned}
\int_{x_1}\int_{x_2} \text{var}(\hat{y})\ dx_2 dx_1 &= \int_{x_1}\int_{x_2} \text{var}\left(\sum_{j,k,l} \hat{\alpha}_{jkl} B_j(x_1) B_k(x_2) B_l(t)\right)\ dx_2 dx_1 \\
&= \int_{x_1}\int_{x_2} \sum_{j,m}^{m_1}\sum_{k,n}^{m_2}\sum_{l,o}^{m_3} B_j(x_1)B_k(x_2)B_l(t)\ \cdot \\
&\qquad\qquad \cdot B_m(x_1)B_n(x_2)B_o(t)\ \text{cov}(\hat{\alpha}_{jkl}, \hat{\alpha}_{mno})\ dx_2 dx_1 \qquad (6.39) \\
&= \sum_{j,m}^{m_1}\sum_{k,n}^{m_2}\sum_{l,o}^{m_3} \text{cov}(\hat{\alpha}_{jkl}, \hat{\alpha}_{mno}) \cdot B_l(t)B_o(t) \int_{x_1} B_j(x_1)B_m(x_1)\ dx_1\ \cdot \\
&\qquad\qquad \cdot \int_{x_2} B_k(x_2)B_n(x_2)\ dx_2 \\
&= \text{tr}\left[\mathbf{C}_{\hat{\alpha}|y}(\widetilde{\mathbf{B}}_1 \otimes \widetilde{\mathbf{B}}_2 \otimes \widetilde{\mathbf{B}}_3)\right]
\end{aligned}$$

where

$$\widetilde{\mathbf{B}}_1[j,m] = \int_{x_1} B_j(x_1)B_m(x_1)\ dx_1 \qquad\qquad \widetilde{\mathbf{B}}_2[k,n] = \int_{x_2} B_k(x_2)B_n(x_2)\ dx_2$$

$$\widetilde{\mathbf{B}}_3[l,o] = B_l(t)B_o(t).$$

Thus the objective function is

$$\begin{aligned}
\phi_{IV} &= \int_{x_1}\int_{x_2} \text{var}\,(\hat{y})\ \ dx_2 dx_1 \\
&=\text{tr}\left[\mathbf{C}_{\hat{\alpha}|y}(\widetilde{\mathbf{B}}_1 \otimes \widetilde{\mathbf{B}}_2 \otimes \widetilde{\mathbf{B}}_3)\right]
\end{aligned} \qquad (6.40)$$

where $\mathbf{C}_{\hat{\alpha}|y}[p,q] = \text{cov}(\hat{\alpha}_p, \hat{\alpha}_q)$. The optimisation procedure seeks to find, for a fixed set size w and new sampling time t_0, the set of wells which minimise the expression in Equation 6.40.

For the spatial and spatio-temporal p-spline models, due to low estimates of the smoothing parameter λ, numerical instabilities were experienced when using the frequentist formulation of $\mathbf{C}_{\hat{\alpha}|y}$ detailed in Equation 2.21. To ensure robust and reliable results, the Bayesian formulation of this covariance was used instead. The Bayesian formulation of $\mathbf{C}_{\hat{\alpha}|y}$ is shown up to a multiplicative constant in Equation 6.41; more detail is given in Section 2.3.2:

$$\mathbf{C}_{\hat{\alpha}|y} \propto \left(\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{D}^\top \mathbf{D}\right)^{-1}. \tag{6.41}$$

For the spatio-temporal model the two smoothing parameter p-spline model, introduced in Chapter 4, was used.

From the derivations of both objective functions it is apparent that neither of them depend on the values observed in earlier sampling events; only the spatio-temporal models depend on the locations at which earlier samples were taken. However, the previously observed data are used to estimate the parameters contained in each of the models. During the design optimisation process, the uncertainty associated with the estimated parameters is not accounted for and it is assumed that the models used do not exhibit ballooning in their predictions.

6.5 Simulation Studies

The second half of this chapter presents the results of four simulation studies that were conducted using the objective functions discussed in the previous sections. The first study seeks to identify general trends in the sampling designs chosen by each objective function. The second study investigates the effect of well-specific sampling frequency on the sampling designs chosen for the next event. The final two studies use the objective functions and the prediction variance to alter the well network by considering the addition of a new well and the removal of an existing well. In these studies, the sampling design optimisations are constrained by the following three restrictions:

1. It is assumed that the monitoring well network is in place and fixed,
2. The number of wells to be sampled in the next event is also fixed,
3. The time of the next sampling event is predetermined.

Obviously, by the nature of the spatial models, knowing the time of the next sampling event is redundant and hence restriction (3) is irrelevant for these models.

The primary aim of the following studies is to try and determine the most suitable locations to take samples, given the restrictions, that will result in gaining the most knowledge about the study region by minimising the objective functions.

6.5.1 Data Simulation

To assess the optimal sampling designs chosen by each model and objective function, groundwater data were simulated from a PDE using a groundwater flow and contaminant transport model. Two network designs were used for this study, each made up of 14 monitoring wells. These will be referred to as 'Design 1' and ' Design 2' respectively. These designs were chosen as Design 1 contains clusters of closely positioned wells whereas Design 2 contains large areas with no well coverage. Figure 6.1 illustrates the simulated PDE at four time points, along with the two well networks.

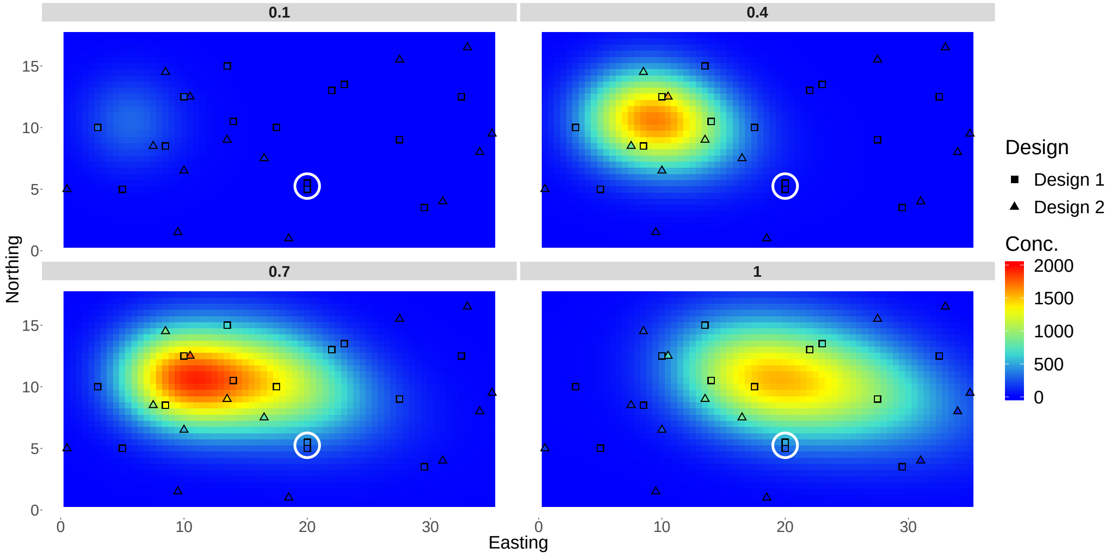

FIGURE 6.1: True surface simulated from the PDE at times $t \in \{0.1, 0.4, 0.7, 1\}$. The wells, from Design 1, contained within the white circle are the cluster that are discussed in Section 6.5.3.

Once the PDE had been simulated it was interpolated at 40 time points and the two sets of well locations to give observed data. Additive noise was then added. Since most data obtained from groundwater monitoring sites contains several missing values, a random 25% of the observed data were removed before the study was carried out. The data were then subsetted to only include observations from the first 20 time points. Results for Design 1 will be presented here, with the results for Design 2 being presented in Appendix D.

6.5.2 Objective Functions in Action

To compare the designs chosen by each model (spatial p-splines, spatial Kriging and spatio-temporal p-splines) and objective function combination, three proportions (25%, 50% and 75%) of the total number of wells were optimised for the next sampling event. For the spatio-temporal model the time of the next sampling event was chosen to be 5% of the temporal range of the current data into the future.

Designs for Spatial Models

Spatial p-splines and Kriging were used for optimising sampling designs using the two proposed objective functions. When using the p-splines model degree three basis functions were used with a first order difference penalty. For the Kriging model, a Matérn covariance function was used with $\kappa = 2$. Figures 6.2 and 6.3 show the optimal designs for each objective function for the spatial p-splines and Kriging models respectively.

For the spatial p-splines model, as the proportion of wells being optimised increases the designs chosen by each objective function become more similar, with the designs for 75% of the wells being the same for both objective functions. When optimising for 25% of the wells, both designs select wells spaced out across the study region whereas for 50% optimisation, the VM design selects more wells near the centre of the study region compared with the IV design. When wells are closely positioned i.e. those circled in Figure 6.1, both objective functions only select at most one of these wells in their optimal designs. This is also evident in the cluster located directly above the circled cluster.

The designs optimised using spatial Kriging differ significantly from those of the spatial p-splines model. When optimising for 25% of the wells, both objective functions select the same 4 wells positioned around the perimeter of the region. Similar results are seen when optimising for 50% of the wells using the VM objective function. On the contrary, for this optimisation, the IV objective function favours wells in the centre. When using

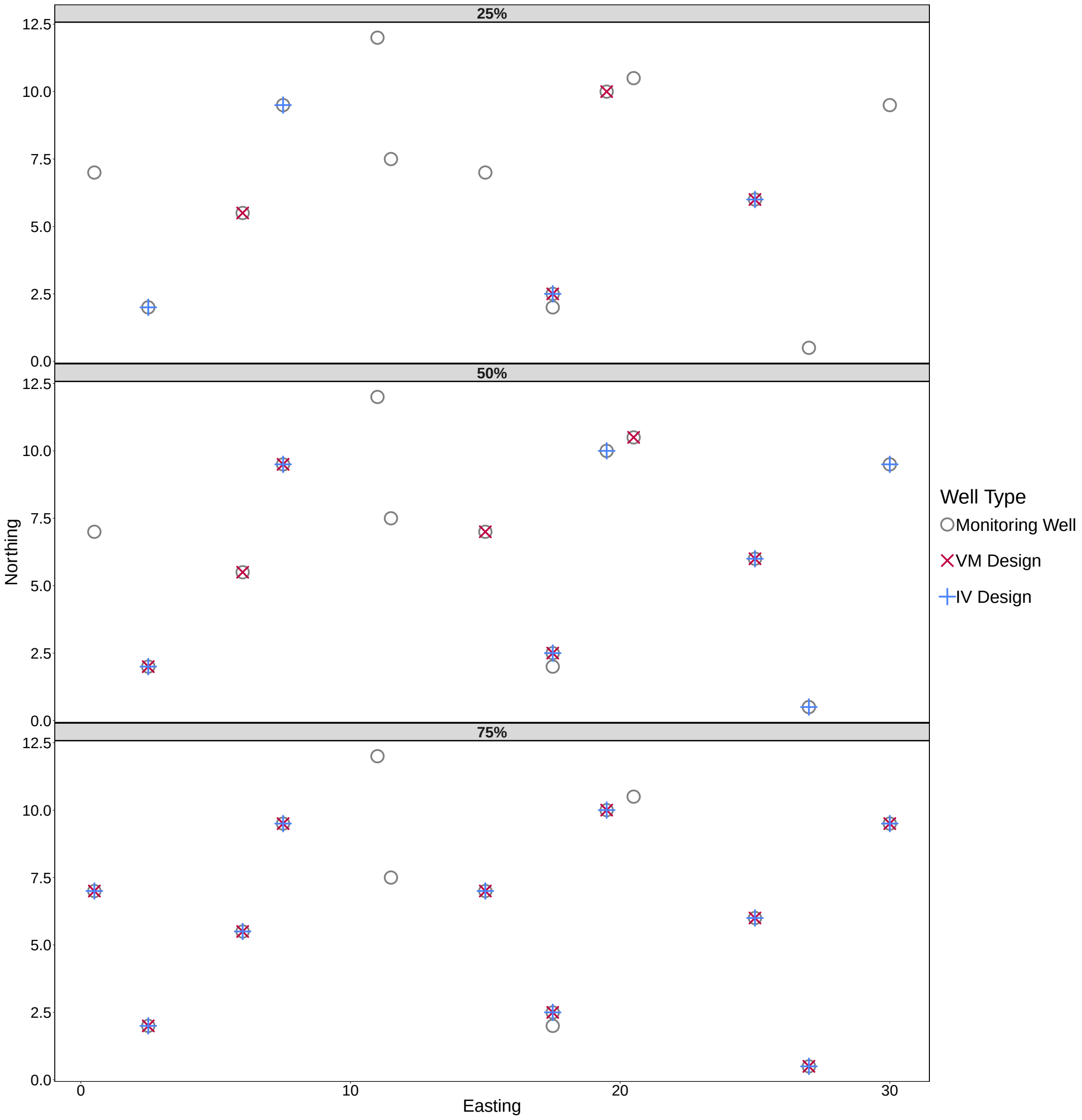

Figure 6.2: Optimal designs for the integrated variance (IV) and variance of the mass (VM) objective functions using a spatial p-splines model when optimising for three proportions of the total number of wells (25%, 50% and 75%) using Design 1.

75% of the wells, both objective functions optimal wells are located primarily to the left of the study region.

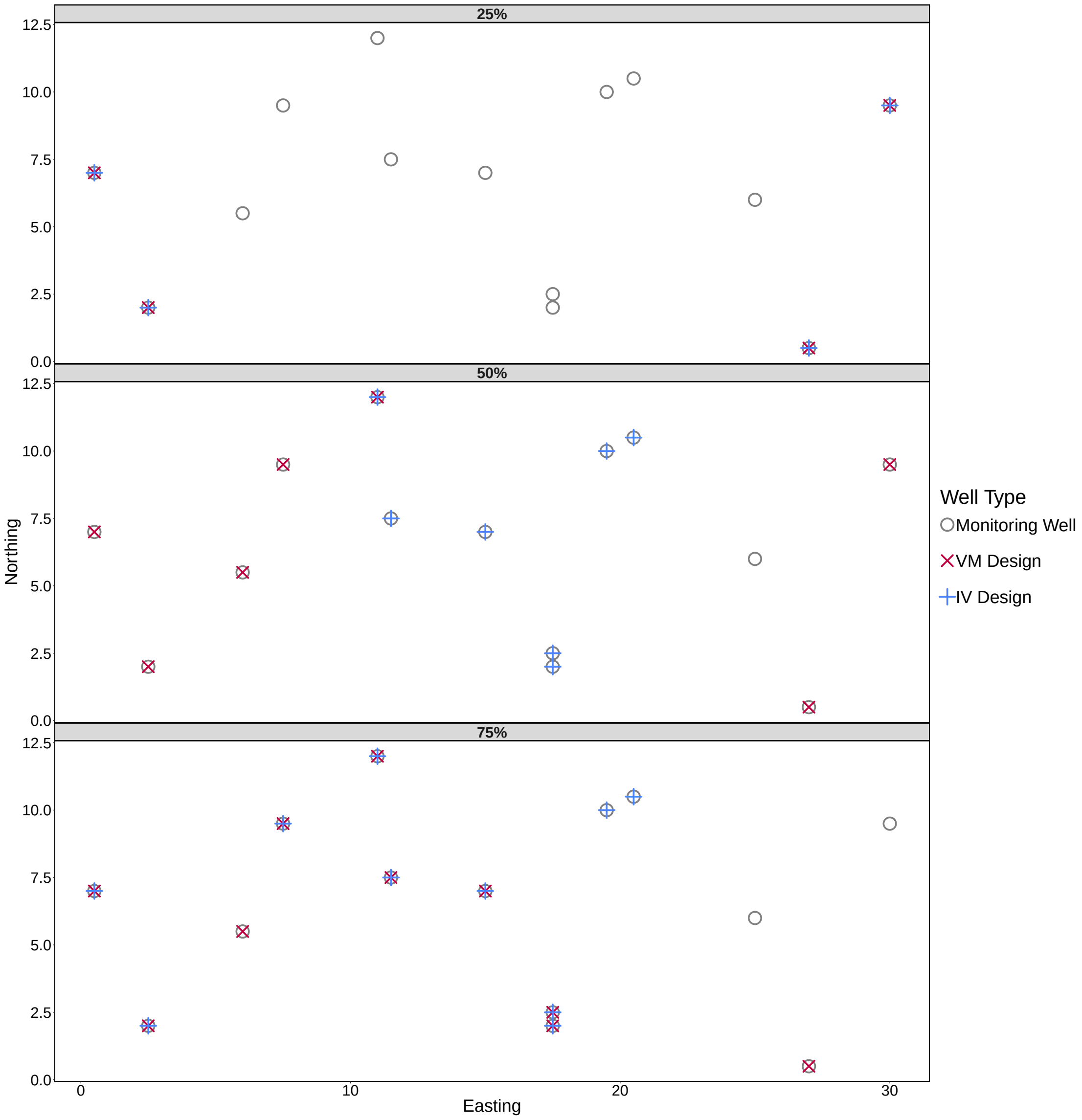

FIGURE 6.3: Optimal designs for the integrated variance (IV) and variance of the mass (VM) objective functions using a spatial Kriging model when optimising for three proportions of the total number of wells (25%, 50% and 75%) using Design 1.

Tables 6.1 and 6.2 cross-tabulate the objective function values for each optimal design for the spatial p-splines and Kriging models respectively. Bold values indicate the design with the minimum of each objective function i.e. across each row in each subsection.

Table 6.1: Cross-tabulation of the objective function values for each objective functions optimal design when optimising for 25%, 50% and 75% of the total number of wells using a spatial p-splines model using Design 1. Bold entries indicate the design with the minimum of each objective function.

		Optimal Design					
		25%		50%		75%	
		VM	IV	VM	IV	VM	IV
Value	VM	**416011**	464250	**180109**	234187	106966	106966
	IV	10164	**9980.6**	7388.3	**7210.9**	5597	5597

Table 6.2: Cross-tabulation of the value of the objective functions for each objective functions optimal design when optimising for 25%, 50% and 75% of the total number of wells using a spatial Kriging model using Design 1. Bold entries indicate the design with the minimum of each objective function.

		Optimal Design					
		25%		50%		75%	
		VM	IV	VM	IV	VM	IV
Value	VM	188980	188980	**372796**	442309	**547669**	550379
	IV	2491.9	2491.9	4388	**4268**	6057.7	**5991.1**

For both spatial models, the fact that all of the bold values lie on the main diagonal of each subsection indicates that the minimum value of the objective function is seen in the design optimised using that objective function, when the optimal designs differ between objective functions. This seems intuitive, but it also highlights that the choice of objective function matters and should be based on what the practitioner wants to learn from the study. If the interest is in minimising the prediction variance i.e. accurately estimating the state of the whole study region, then the design should be optimised using the IV objective function. Whereas if the primary interest is in accurately quantifying the total contaminant mass then the VM objective function should be used. Looking at optimising 50% of the wells, the VM objective function is approximately 20% lower for the design optimised using the VM objective function compared with the design optimised using the IV objective function. Similar results can be seen for the Kriging optimal design for 50% of the wells. By using a design optimised using the VM objective function, the variance of the mass is 15% lower than that of the design optimised using the IV objective function. As the proportion of wells being optimised increases, the difference in the objective function values between designs reduces due to the designs covering more of the study region and thus providing more information. The differences between designs are also not as extreme for the IV objective function, with the values for each design only differing at most by $\sim 2.5\%$ for both spatial models.

Designs for a Spatio-temporal Model

In a similar manner to the spatial models, a spatio-temporal p-splines model was used to optimise sampling designs using the two objective functions. As previously mentioned the time of the next sampling event was predefined ($\sim 5\%$ of the current temporal range into the future). The two smoothing parameter spatio-temporal p-splines model was used with 15 basis functions for the easting and time component and the northing component had its number of basis functions scaled by the dimensions of the study region. Additional basis functions should be added into the temporal component proportionally to how far into the future the next sampling event will occur. If this is not done, by their construction, the temporal basis will just be stretched out over the extended temporal domain, reducing the flexibility of the model and potentially over-smoothing the temporal dimension.

Figure 6.4 shows the optimal designs for each objective function when optimising for three proportions of the total number of wells (25%, 50% and 75%).

The majority of the wells chosen by each objective function are the same for each proportion of wells being optimised using the spatio-temporal model. The designs using 25% and 50% of the wells differ by only one location, whilst the designs using 75% of the wells are identical for both objective functions. Looking at the designs using 50% of the wells, similar to the trends seen in the spatial p-splines designs, the optimal design chosen using the VM objective function appears to favour wells located in the centre of the study region over those located around the boundary, whereas the IV objective function seems to choose wells that cover the study region. The VM function favouring wells near the centre may also be due to the perimeter wells being sampled in the previous sampling event.

For all three optimisation scenarios only one of the two wells present in the cluster in the lower half of the network (circled in white in Figure 6.1) are chosen. This was also

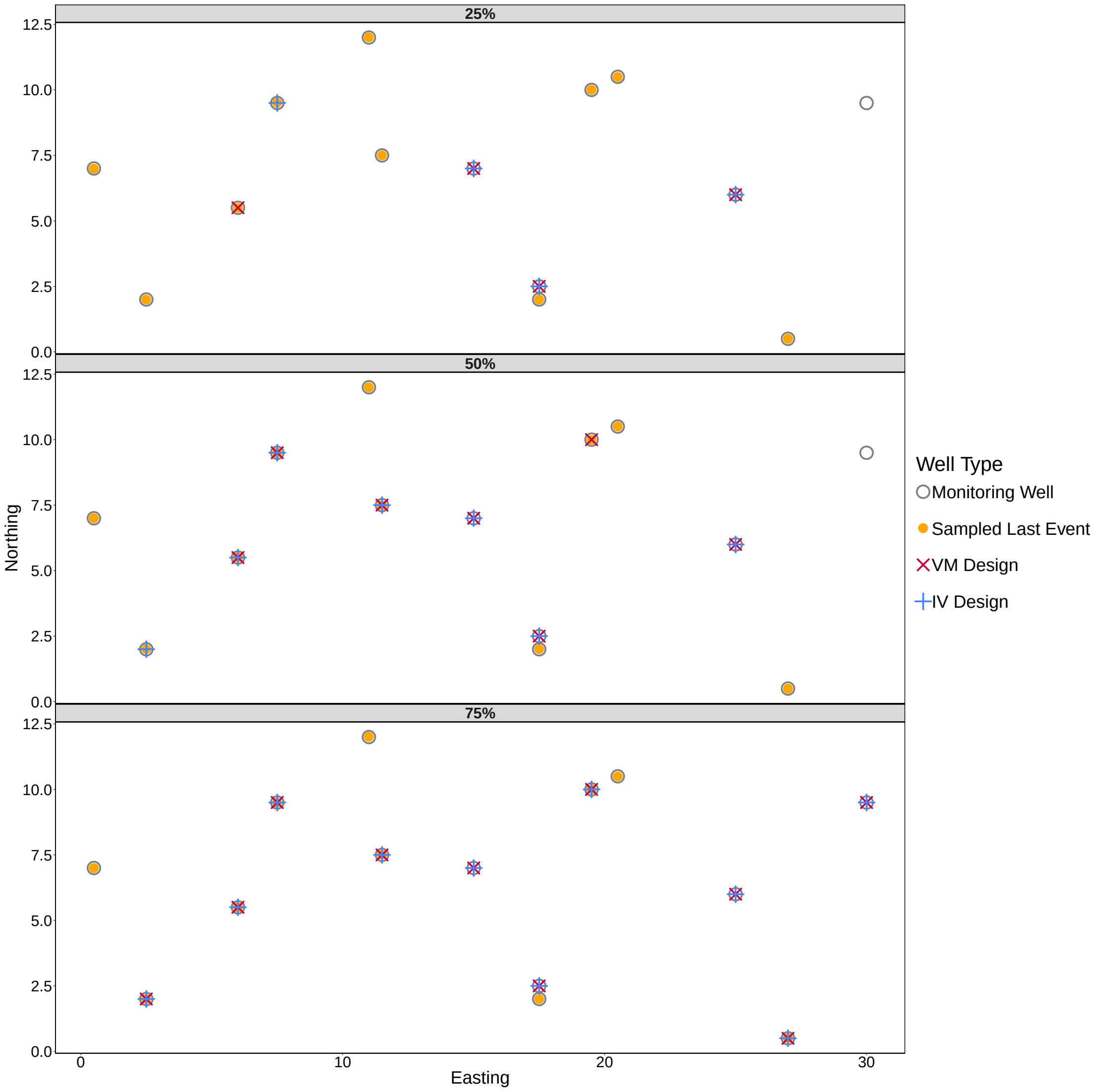

Figure 6.4: Optimal designs for the integrated variance (IV) and variance of the mass (VM) objective functions using a spatio-temporal p-splines model when optimising for three proportions of the total number of wells (25%, 50% and 75%) using Design 1.

the case when using a spatial p-splines model, suggesting that using both wells in the design provides redundant information.

Comparing the newly selected sampling locations to the 10 wells sampled in the last event (filled orange wells), when optimising for 25% of the wells, both objective functions select 3 of the 4 unsampled wells from the previous event resulting in only one well being left unsampled across the two sampling events. Similar results are seen when optimising for 50% of the wells. Finally, when optimising for 75% of the wells, the design chosen by both objective functions ensures that all 14 wells are sampled across the two sampling events.

TABLE 6.3: Cross-tabulation of the value of the objective functions for each objective functions optimal design when optimising for 25%, 50% and 75% of the wells using a spatio-temporal p-splines model for Design 1. Bold entries indicate the design with the minimum of each objective function.

		Optimal Design					
		25%		50%		75%	
		VM	IV	VM	IV	VM	IV
Value	VM	**40200**	40322	**39358**	36416	38837	38837
	IV	2396.7	**2395.9**	2374.5	**2373.6**	2355.5	2355.5

In a similar manner to the spatial models, the objective functions were cross-tabulated for each design for the spatio-temporal p-splines model. Again, the value of the objective function was lowest for the design optimised using the corresponding objective function. The increase in the objective function by using the converse design is not as substantial as it was for the spatial models, with the largest increase being $\leq 0.3\%$ (optimising for 25% of the wells using the VM objective function). This is likely due to the spatio-temporal model being able to carry information forward from earlier sampling events.

6.5.3 The Effect of the Previous Sampling Frequency

Focusing only on the spatio-temporal p-splines model, since the time to the next sampling event has no influence on the sampling design chosen using a spatial model, it was of interest to determine whether the previous well-specific sampling frequency had any influence on the optimal design chosen for the next sampling time.

To assess this, three new datasets were simulated from Design 1. Each dataset contained two samples in the first 6 time points from 7 of the 14 wells ('low frequency wells') and the other 7 wells were randomly sampled 14 times over the 20 sampling periods ('frequently sampled wells'). The first dataset contained 7 randomly selected wells in the 'low frequency well' set. This dataset was used to initially determine whether the 'low frequency wells' would be favoured over the previously more frequently sampled wells for the next sampling event. The remaining two datasets assessed the effect the cluster of wells in the lower half of the plot (circled in white in Figure 6.1) had on the chosen design based on their sampling frequency. The second dataset contained both wells from the cluster in the 'low frequency well' set. This dataset was designed to determine whether both wells, which provide very similar information, would be chosen for the next sampling event or if only one would be chosen. Finally, the third dataset included only one of the two wells from the cluster in the 'low frequency well' set. This dataset was used to determine whether the close spatial proximity of the wells had any influence on the well in the 'low frequency well' set being chosen in the design for the next sampling event, given there was already recent information in this region from the well in the 'frequently sampled well' set.

Each of the datasets were used along with spatio-temporal p-spline models and the two objective functions to optimise the sampling design for the next time point. Figures 6.5, 6.6 and 6.7 show the optimal designs for each objective function and dataset respectively when optimising for seven wells.

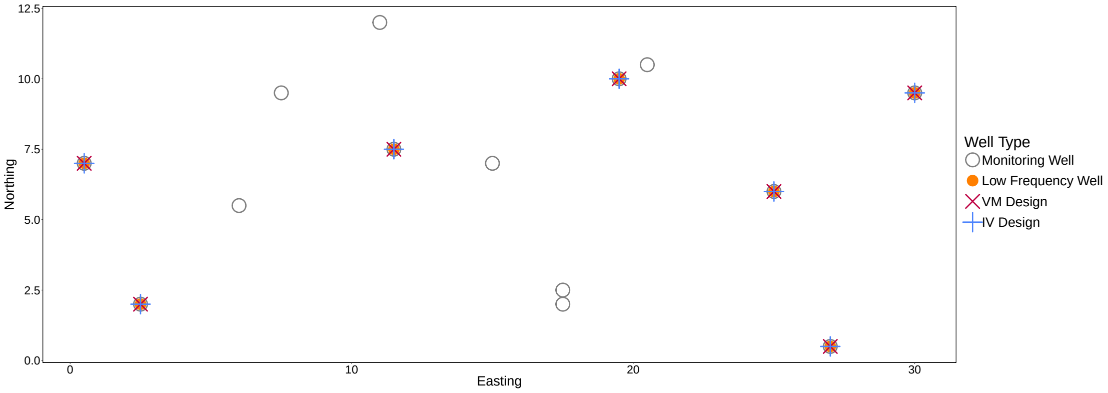

FIGURE 6.5: Optimal design of seven wells using the dataset containing seven randomly selected wells in the 'low frequency well' set i.e. first dataset.

From Figure 6.5 it is apparent that both objective functions favour wells for the next sampling event which have not been sampled recently when the first dataset has been observed. This is evident since all seven wells in the 'low frequency well' set of this dataset were selected as the optimal design for the next sampling period by both objective functions.

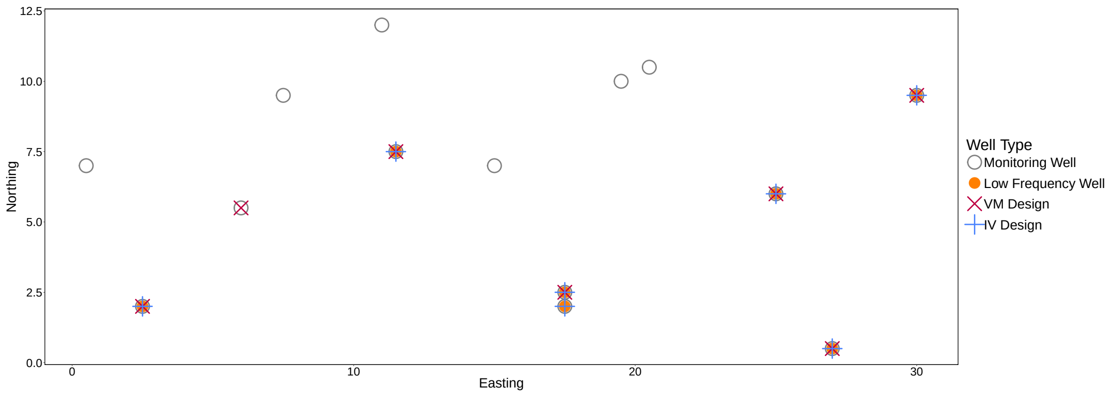

FIGURE 6.6: Optimal design of seven wells using the dataset containing both of the clustered wells in the 'low frequency well' set i.e. second dataset.

Optimal designs for the second dataset, where both of the clustered wells were in the 'low frequency well' set, differed slightly for each objective function. The IV objective function included all of the 'low frequency wells' in its design, whereas, the VM objective function selected six wells from the 'low frequency well' set, with only one from the cluster, and instead of selecting the other well in the cluster, selected one of the 'frequently sampled wells' in its optimal design.

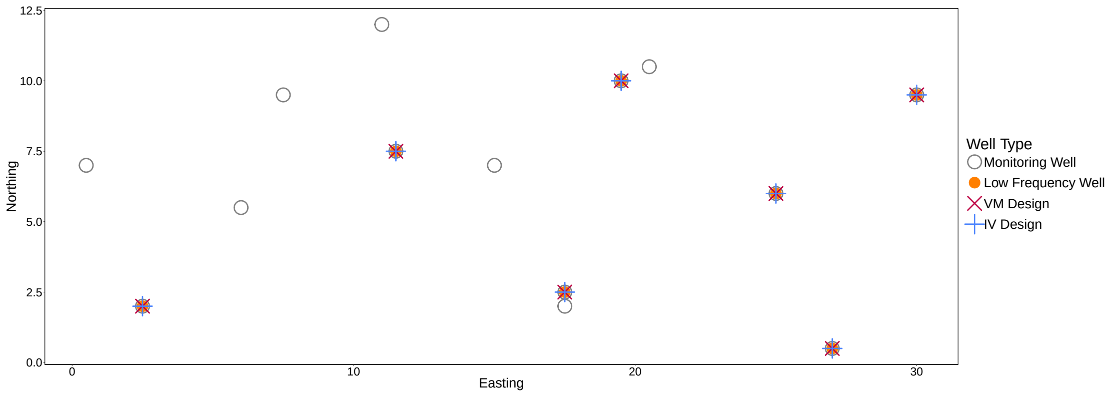

FIGURE 6.7: Optimal design of seven wells using the dataset containing one of the clustered wells in the 'low frequency well' set i.e. third dataset.

For the final dataset, where one of the two wells in the cluster is included in the 'low frequency well' set and the other is not, both objective functions again select all of the wells in the 'low frequency well' set as their optimal design for the next sampling event.

The results of this study indicate that previous well-specific sampling frequency does have an influence on the design chosen for the next sampling event by both the VM and IV objective functions. This is evident from almost all combinations of dataset and objective function choosing the wells in the 'low frequency well' set as the optimal design for the next sampling event. The only discrepancy from this is the VM objective function only selecting one of the two clustered wells in its next design when using the second dataset which contains both wells in the cluster in its 'low frequency well' set. Instead, this objective function opts for a well that provides more coverage of the study region.

6.5.4 Increasing the Time to the Next Sampling Event

The differences between using a spatio-temporal model and a spatial model to optimise the design for the next sampling event become negligible as the time to the next event increases. This can be shown theoretically for a Kriging model with relative ease.

Spatio-temporal Kriging Model

In Section 2.4.5 of Chapter 2, the covariance of the joint distribution of a new observation at location, $\mathbf{s}_0$, and time, t_0, and the current observations was defined as,

$$\mathbf{C} = \begin{pmatrix} k & \mathbf{c}_0^\top \\ \mathbf{c}_0 & \mathbf{K} \end{pmatrix} \tag{6.42}$$

where

$$k = C_y(t_0 - t_0; \boldsymbol{\theta}_t) \cdot C_y(\mathbf{s}_0 - \mathbf{s}_0; \boldsymbol{\theta}_s),$$

$$\mathbf{c}_0 = (C_y(t_0 - t_1; \boldsymbol{\theta}_t) \cdot C_y(\mathbf{s}_0 - \mathbf{s}_1; \boldsymbol{\theta}_s), \ldots, C_y(t_0 - t_m; \boldsymbol{\theta}_t) \cdot C_y(\mathbf{s}_0 - \mathbf{s}_n; \boldsymbol{\theta}_s)),$$

$$\mathbf{K} = \boldsymbol{\Sigma}^{(t)} \otimes \boldsymbol{\Sigma}^{(s)}.$$

Here, $\boldsymbol{\Sigma}_{ij}^{(s)} = C_y(||\mathbf{s}_i - \mathbf{s}_j||; \boldsymbol{\theta}_s)$ is the spatial covariance matrix, $\boldsymbol{\theta}_s = (\sigma_s^2, \tau_s^2, \phi_s)$ are the spatial covariance parameters, $\boldsymbol{\Sigma}_{ij}^{(t)} = C_y(||t_i - t_j||; \boldsymbol{\theta}_t)$ is the temporal covariance matrix, $\boldsymbol{\theta}_t = (\sigma_t^2, \tau_t^2, \phi_t)$ are the temporal covariance parameters and $C_y()$ is a covariance function from Section 2.4.2.

For ease of understanding, consider the simple example where there has been one sampling event at time t_1 and the next sampling event is to happen at time t_2. In this example, let $\mathbf{C}$ take the form,

$$\mathbf{C} = \begin{pmatrix} \mathbf{K}_{22} & \mathbf{K}_{12} \\ \mathbf{K}_{12}^\top & \mathbf{K}_{11} \end{pmatrix} \tag{6.43}$$

where

$$\mathbf{K}_{22} = \boldsymbol{\Sigma}_2^{(t)} \otimes \boldsymbol{\Sigma}_2^{(s)} = \sigma_t^2 \mathbf{I} \otimes \boldsymbol{\Sigma}_2^{(s)},$$

$$\mathbf{K}_{12} = \boldsymbol{\Sigma}_{12}^{(t)} \otimes \boldsymbol{\Sigma}_{12}^{(s)},$$

$$\mathbf{K}_{11} = \boldsymbol{\Sigma}_1^{(t)} \otimes \boldsymbol{\Sigma}_1^{(s)} = \sigma_t^2 \mathbf{I} \otimes \boldsymbol{\Sigma}_1^{(s)}.$$

Here, $\boldsymbol{\Sigma}_1^{(s)}$ is the spatial covariance matrix for the samples taken at time t_1 and similarly $\boldsymbol{\Sigma}_2^{(s)}$ is the spatial covariance matrix for the samples taken at time t_2. The corresponding temporal covariance matrices, $\boldsymbol{\Sigma}_1^{(t)}$ and $\boldsymbol{\Sigma}_2^{(t)}$, are equal to $\sigma_t^2 \mathbf{I}$ since $||t_1 - t_1|| = 0$ and

$||t_2 - t_2|| = 0$. Finally, $\mathbf{\Sigma}_{12}^{(t)}$ and $\mathbf{\Sigma}_{12}^{(s)}$ are the temporal and spatial covariances between the observations at times t_1 and t_2 respectively.

As the time between sampling events increases i.e. as $||t_1 - t_2|| \to \infty$, the entries in $\mathbf{\Sigma}_{12}^{(t)}$ tend towards 0, resulting in the entries of $\mathbf{K}_{12}$ tending towards 0 and thus

$$\mathbf{C} \to \begin{pmatrix} \mathbf{K}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{11} \end{pmatrix} \tag{6.44}$$

i.e. the sampling events becoming temporally independent. This is proportionally equivalent to two spatial Kriging arguments at times t_1 and t_2 if the spatial covariance parameters are the same for the spatial and spatio-temporal models.

Spatio-temporal P-splines Model

Establishing independence for large time separation in a spatio-temporal p-splines model is slightly more challenging due to the difference penalty. To begin with, consider the unpenalised spatio-temporal regression spline model made up of m B-spline basis functions for each spatial and temporal component i.e. the basis functions matrix, $\mathbf{B}$, is of dimension $n \times m^3$. To demonstrate this case, consider a spatio-temporal dataset that contains two temporal clusters of observations, with the vectors of time points within these clusters are denoted by $\mathbf{t}_1$ and $\mathbf{t}_2$. For now, assume a large time window between $\max \mathbf{t}_1$ and $\min \mathbf{t}_2$, this suggests there will be a block of several temporal basis functions between these two time point that are ‘inactive’ i.e. the corresponding columns of $\mathbf{B}$ will be equal to 0, because there are no data in this region. This idea is shown in Figure 6.8.

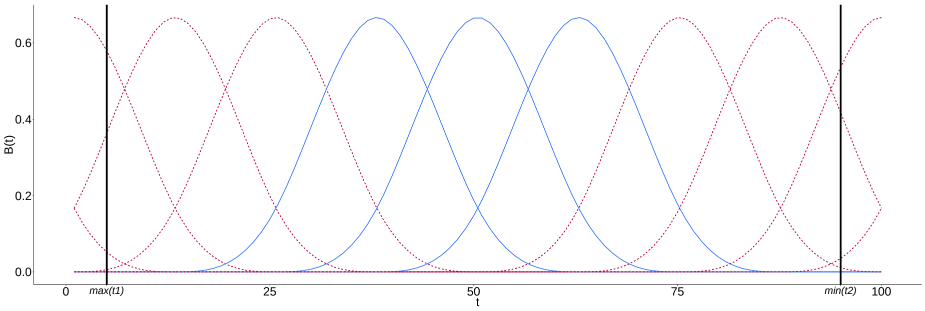

FIGURE 6.8: Temporal B-spline basis functions with two observed time points. Dashed red lines indicate 'active' basis functions (i.e. those positioned over observations).

Given $\mathbf{B}$ is constructed by row-wise Kronecker products of the marginal bases, $\mathbf{B}$ will be made up of blocks containing the Kronecker product of each temporal basis function and the spatial basis functions i.e.

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{t_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_{t_2} \end{pmatrix}$$

where $\mathbf{B}_{t_1}$ is made up of the Kronecker product of the four[1] 'active' temporal basis functions, over the observations at $\max \mathbf{t}_1$, and the spatial basis. $\mathbf{B}_{t_2}$ will be similarly constructed but with the 'active' temporal basis functions over the observation at $\min \mathbf{t}_2$. The block of $\mathbf{0}$s between $\mathbf{B}_{t_1}$ and $\mathbf{B}_{t_2}$ are the 'inactive' basis functions where there are no data i.e. are equal to 0.

To obtain the basis coefficients, the least squares estimator is used i.e.,

$$\hat{\alpha} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{y} = \begin{pmatrix} \left(\mathbf{B}_{t_1}^\top \mathbf{B}_{t_1}\right)^{-1} \mathbf{B}_{t_1}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{B}_{t_2}^\top \mathbf{B}_{t_2}\right)^{-1} \mathbf{B}_{t_2}^\top \end{pmatrix} \mathbf{y} \tag{6.45}$$

Since there is a block on 'inactive' temporal basis functions positioned between the 'active' temporal basis functions at the two sampling events, the coefficients for each block of 'active' basis functions can be separated and determined independently as shown in Equation 6.45. Thus indicating that as $\min \mathbf{t}_2 - \max \mathbf{t}_1 \to \infty$ the sampling design

[1] Here degree 3 B-splines are being used, so four B-splines are non-zero at each observation, see the beginning of Section 2.2.1

optimised for time $\min \mathbf{t}_2$ will become independent of the observations taken at $\max \mathbf{t}_1$ i.e. a spatial design.

With p-splines, the construction of the penalty term plays an important role in the design for the spatio-temporal p-splines model tending towards the design for a spatial model. The penalty will determine how the model will interpolate the 'gap' in the temporal observations. Assuming the two smoothing parameter p-splines model is used, if there is a first order difference penalty the model will interpolate at a constant mean value whereas if there is a second order penalty the model will interpolate in a linear trend. This indicates that there will be a dependence between the two time points and thus the designs will not be independent. As the temporal smoothing parameter, $\lambda_{rel} \to 0$, the design from the spatio-temporal model will converge on the spatial models design. The overall smoothing parameter, λ, will however have a different interpretation due to the temporal basis functions being included in the estimation of $\hat{\alpha}$.

6.6 Using the Objective Functions to Change the Well Network

6.6.1 Adding a New Well to the Network

Expanding the monitoring well network to increase knowledge of a study region is often of interest but determining the most appropriate location for a new well is difficult. Given the expense associated with digging and installing a new well, it is of paramount importance to ensure that the potential knowledge gained from the new well installation is maximised.

The IV and VM objective functions can be used to determine where a new monitoring well could potentially be positioned by determining the location that gives the largest decrease in the current minimum of the objective function. For this study, individually each pixel was proposed as a location for a new well and each objective function was calculated using the optimal 10 wells and this 'new' location. The difference in the current minimum of the objective function using the 10 optimal wells and the value achieved with the addition of this new location was then calculated. For this study only spatial and spatio-temporal p-splines were used.

The colour of each pixel in Figures 6.9, 6.10, 6.11 and 6.12 shows the change in value of each objective function, using the spatial and spatio-temporal p-spline models, when this pixel is proposed as a new well location along with the optimal 10 monitoring wells.

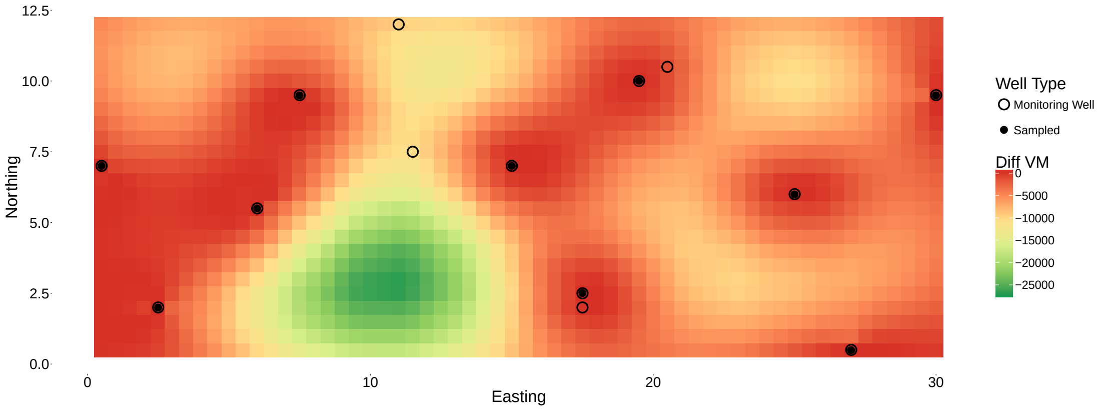

FIGURE 6.9: Change in the VM objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatial p-splines model for Design 1.

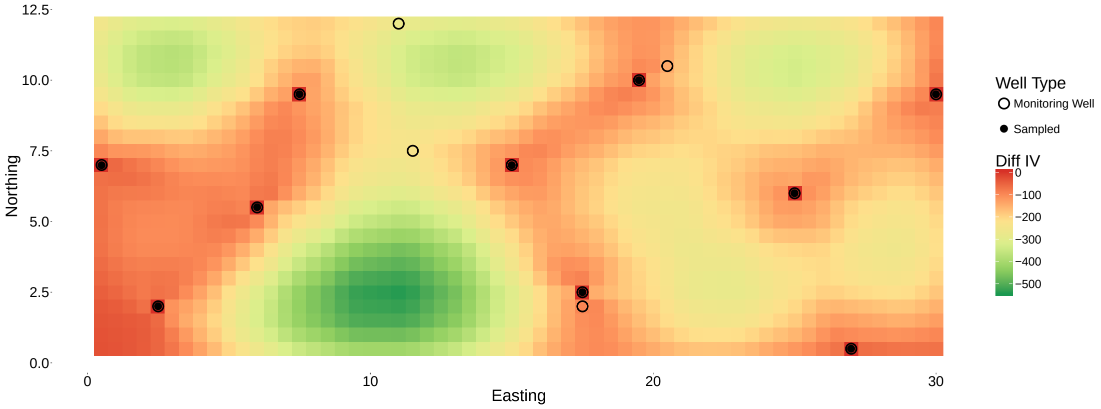

FIGURE 6.10: Change in the IV objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatial p-splines model for Design 1.

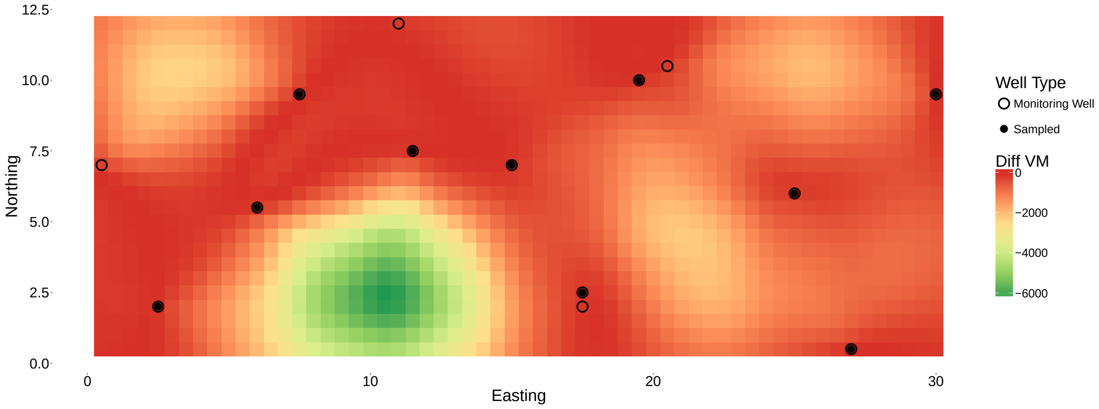

Figure 6.11: Change in the VM objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatio-temporal p-splines model for Design 1.

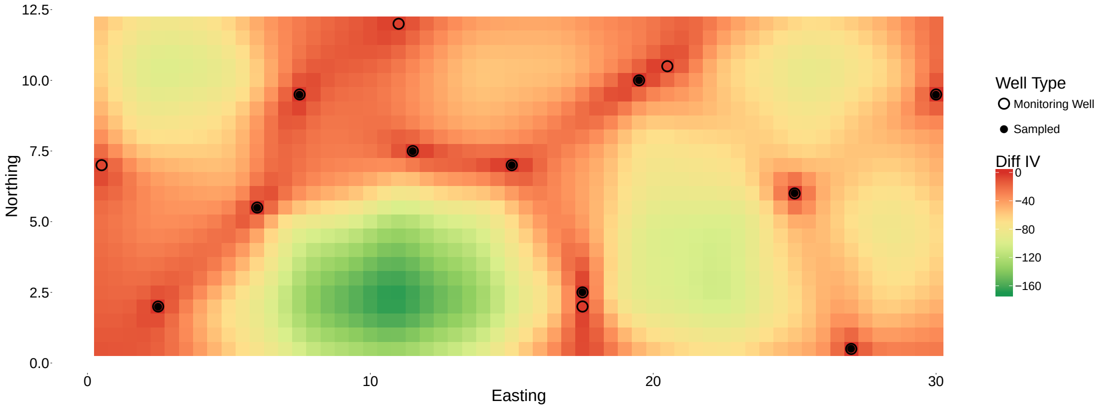

Figure 6.12: Change in the IV objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatio-temporal p-splines model for Design 1.

All of the above figures, i.e. all model and objective function combinations, suggest that the largest improvement in each objective function i.e. reduction, would be achieved by placing a well to the left of the centre at the bottom of the study region. Unsurprisingly, this location is in a region where the distance between neighbouring wells is largest.

6.6.2 Removing a Well from the Network

Groundwater monitoring well maintenance and sampling can be expensive particularly when monitoring wells are located in remote regions, thus it is often of interest to scale down the network to save money. Deciding on the most appropriate well/wells to remove is difficult and involves a trade off between the cost of maintaining and sampling the well and the potential amount of information that will be lost if the well is removed.

To determine which well would be the 'best' to remove, in the sense that it gives the smallest increase in prediction variance, the change in prediction variance that removing each well would cause was assessed. Using the current network, initially all (n) wells were proposed to be sampled in the next sampling event and the prediction variance was calculated, this gives a reference point for the change in variance as each well is removed. Next, iteratively $n-1$ wells were proposed to be sampled at the same future sampling time and the prediction variance was again calculated. The change in the total prediction variance was then computed (shown in the strips at the top of each plot in Figures 6.13 and 6.14) and plotted to give a heat map, shown in Figures 6.13 and 6.14 for spatial and spatio-temporal p-splines respectively. The well which gives the smallest increase in prediction variance when removed from the network is the well that would be proposed to be removed.

The spatial p-splines model begins by indicating that removing one of the two wells in the cluster at the bottom would result in the smallest increase in prediction variance. The next smallest increase in prediction variance is achieved by removing one of the two wells in the cluster at the top right. This is unsurprising given the close proximity of the clustered wells and suggests that one well in these locations is enough. The wells that, when removed, result in the largest increase in prediction variance are located at the boundary, where they do not have any neighbouring wells on one side. Removing these wells removes all information about what is happening at the boundary and hence the large increase in prediction variance.

The order of well insignificance is very different for the spatio-temporal model compared with the spatial model due to the model being able to utilise data from earlier sampling events. Figure 6.14 indicates the first three wells with the smallest change in prediction variance, when removed from the network, are located around the perimeter of the study

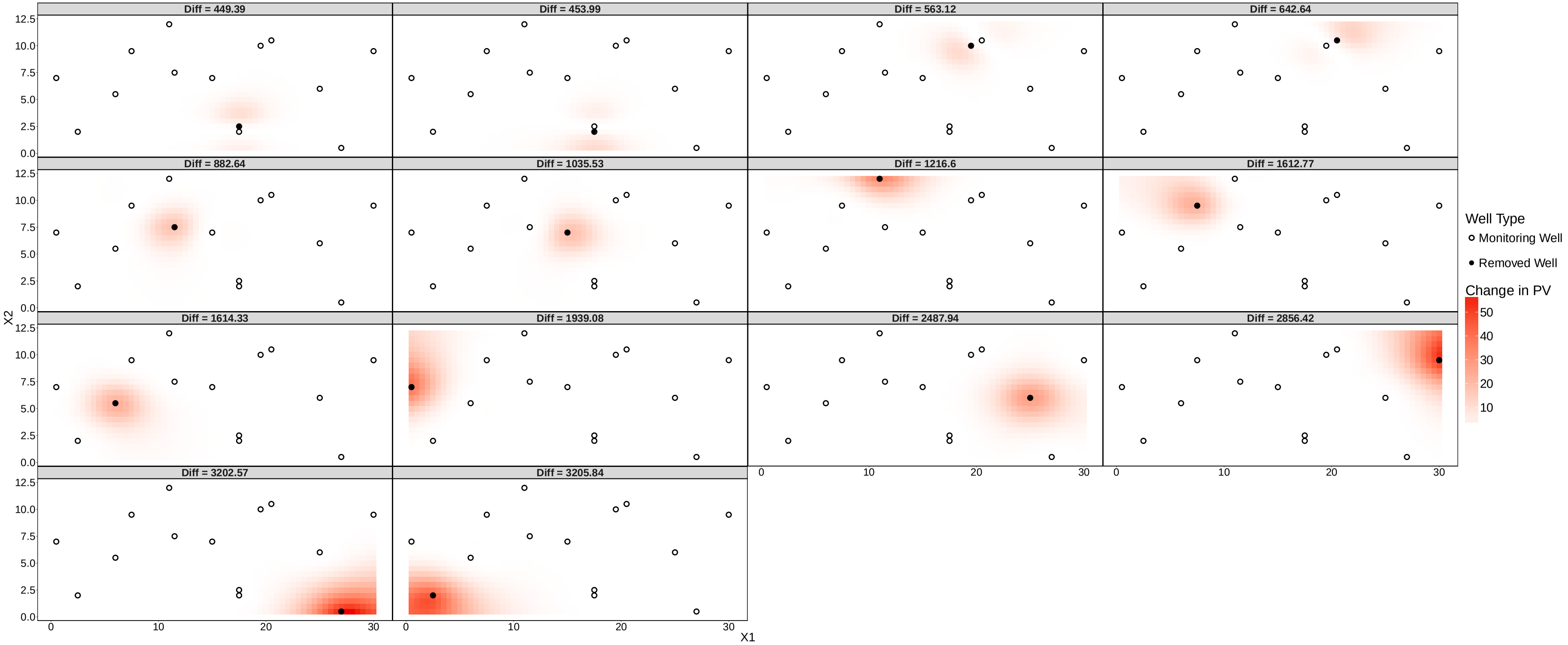

Figure 6.13: Difference in predicted variance if each filled monitoring well is removed at the next time point compared with the prediction variance if all wells were sampled for Design 1 and a spatial p-splines model. The facets are ordered in increasing difference in prediction variance.

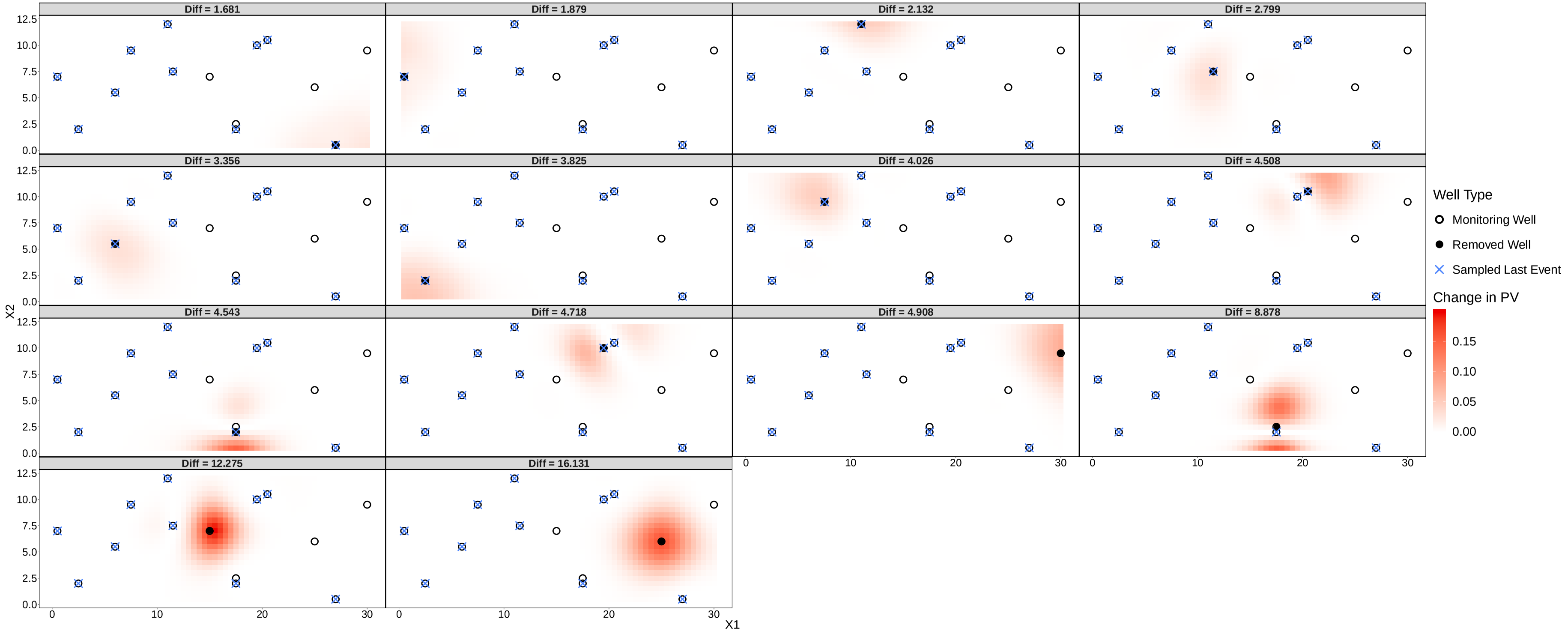

FIGURE 6.14: Difference in predicted variance if each filled monitoring well is removed at the next time point compared with the prediction variance if all wells were sampled for Design 1 and a spatio-temporal p-splines model. The facets are ordered in increasing difference in prediction variance.

region. On the contrary, the four wells which cause the greatest increase in prediction variance are those which were not sampled in the previous sampling event. These wells are also positioned away from the edges of the study region and have distant neighbours causing the increase in prediction variance, when they are removed, to be exaggerated. Interestingly, removing one of the two clustered wells does not give the smallest reduction in prediction variance which is what may have been expected. However one of these wells was not sampled in the previous event, causing an increase in uncertainty. The change in prediction variance as each well is removed is very small compared with the changes seen when the spatial model was used. This again demonstrates the ability of spatio-temporal models to carry forward information from previous sampling events. Obviously, as the time to the next event increases, the less useful previously seen information will be.

6.7 Summary

Two design objective functions have been presented and assessed on two designs (Design 1 here and Design 2 in Appendix D), namely the Variance of the plume Mass (VM) and the Integrated Prediction Variance (IV), for spatial and spatio-temporal p-spline models. Objective functions similar to the VM function have already been discussed in the literature and are most commonly used with Kriging type models here this objective function was also presented for spatial and spatio-temporal p-spline models. The IV objective function is also an already well established design optimisation function. However, again, this is mainly used with Kriging models and here this objective function is also presented for spatial and spatio-temporal p-spline models.

Subject to the imposed restrictions of a fixed well network and a fixed number of samples that can be taken, the resulting designs from the objective functions differ both between functions and models. For the p-spline models, the VM objective function appears to favour wells located in the centre of the study region whereas the IV objective function tends to choose wells that give a good spatial coverage of the study region. The designs chosen by the spatio-temporal model also appear to try to ensure that as many wells as possible were sampled between the current and previous sampling events.

Tables 6.1, 6.2 and 6.3 indicated that the objective function should be chosen carefully and reflect what is desired to be learned about the study region e.g. if the interest is in accurately estimating the plume mass the VM objective should be chosen for design optimisation.

The effect of only sampling some of the wells a few times at the beginning of the sampling period was also investigated for each of the objective functions. It was found that wells which had not been sampled recently were favoured to be sampled in the next sampling event over wells which had been sampled more recently and frequently.

Finally the objective functions were used to determine how the network could be altered i.e. either by adding or removing wells. Each of the p-spline models and both objective functions suggested the same location for a new monitoring well for the dataset and well network used in the simulation study. Both Design 1 used here and Design 2, presented in Appendix D, indicated that the 'best' location was situated in a region where there are no wells located nearby. This is unsurprising since if there are no observed data in these locations the predictions made here will be most uncertain. To determine which well could be removed from the network, the prediction variance was used. The well that, when removed, would result in the smallest increase in prediction variance was seen as the well that could be removed. When using a spatial model, if there were clusters of wells, removing one of these resulted in the smallest increase in prediction variance. On the other hand, the spatio-temporal model suggested wells which had been sampled most recently would cause the smallest increase in prediction variance when removed. Spatio-temporal models are particularly advantageous when wells have been removed from the network due to their ability to use previous sampling information to predict what is going on in the areas where the wells use to be.

When optimising for 25% of the wells, for the network used in this study, this results in only optimising for 4 of the 14 wells. This is a very small sample size to try and make robust inference from when using a spatial model which does not take any previously recorded information into account. Thus, it would be recommended that either a larger sample be taken to use with the spatial models or a spatio-temporal model be used which incorporates previous sampling information into its estimations and predictions. This highlights an advantage of using a spatio-temporal model over a spatial model, with the spatio-temporal model not requiring as many samples at individual sampling events to make robust predictions across the study region.

Chapter 7

Sampling Design for Data with Multiplicative Errors

In Chapter 6, the Integrated Prediction Variance (IV) and Variance of the Mass (VM) objective functions were presented and formulae were derived to compute them for spatial and spatio-temporal models. While these derivations were being computed, it was assumed the data being modelled had additive error. However, commonly the error associated with groundwater quality data is assumed to be multiplicative. In order to work with the standard modelling methodology set out in this thesis, the response variable is log transformed prior to modelling to give an additive interpretation of the error. Once a final model has been decided and fitted, the fitted/predicted values are then transformed back onto the original scale for ease of interpretation.

Log transforming for modelling, then transforming back for interpretation is however, not appropriate for the proposed design objective functions, because integrals computed on the log scale cannot simply be transformed back to the original scale. Computing the objective functions with the log transformed data is sufficient as long as the objective function is also calculated on the log scale, so that the VM objective function to be computed is:

$$\phi_{VM} = \text{var}\left[\int_{x_1}\int_{x_2} \widehat{\log(y)} \; dx_2 dx_1\right], \tag{7.1}$$

and similarly the IV objective function is:

$$\phi_{IV} = \int_{x_1}\int_{x_2} \text{var}\left[\widehat{\log(y)}\right] \; dx_2 dx_1. \tag{7.2}$$

However, often it is of interest to compute and interpret the objective function on the original scale whilst still modelling on the log scale. In this case the VM objective function to be computed is now:

$$\phi_{VM} = \text{var}\left[\int_{x_1}\int_{x_2} \exp\left(\widehat{\log(y)}\right) \; dx_2 dx_1\right], \tag{7.3}$$

and similarly the IV objective function is:

$$\phi_{IV} = \int_{x_1}\int_{x_2} \text{var}\left[\exp\left(\widehat{\log(y)}\right)\right] \; dx_2 dx_1. \tag{7.4}$$

Due to the exponential term inside the integrals, computation of these new objective functions is difficult. The lognormal distribution can be utilised to aid the computation of the IV objective function on the original scale given modelling was performed on the log transformed data. Unfortunately, however, computing the VM objective function for this scenario is complex, and currently unachievable, thus from here in only the IV objective function will be discussed. It is worth noting, however, that the optimal designs chosen by both objective functions are relatively similar.

7.1 The Lognormal Distribution

The lognormal distribution ($\mathcal{LN}$) is a continuous probability distribution of a random variable, X whose logarithm, $Y = \log(X)$, is normally distributed i.e. $Y \sim \mathcal{N}(\mu, \sigma^2)$. Let Z be a standard normal random variable, then:

$$X = \exp(\mu + \sigma Z) \tag{7.5}$$

where μ and σ are the mean and standard deviation, respectively, of the logarithm of X. Thus, by taking logarithms of both sides, the expression above can be redefined as:

$$\log(X) = \mu + \sigma Z. \tag{7.6}$$

Since Z is normally distributed, then $\log(X)$ is also normally distributed.

The lognormal distribution has probability density function (p.d.f):

$$f_{X|\mu,\sigma^2} = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right\}, \tag{7.7}$$

shown in Figure 7.1 with parameters $\mu = 0$ and $\sigma^2 = 1$.

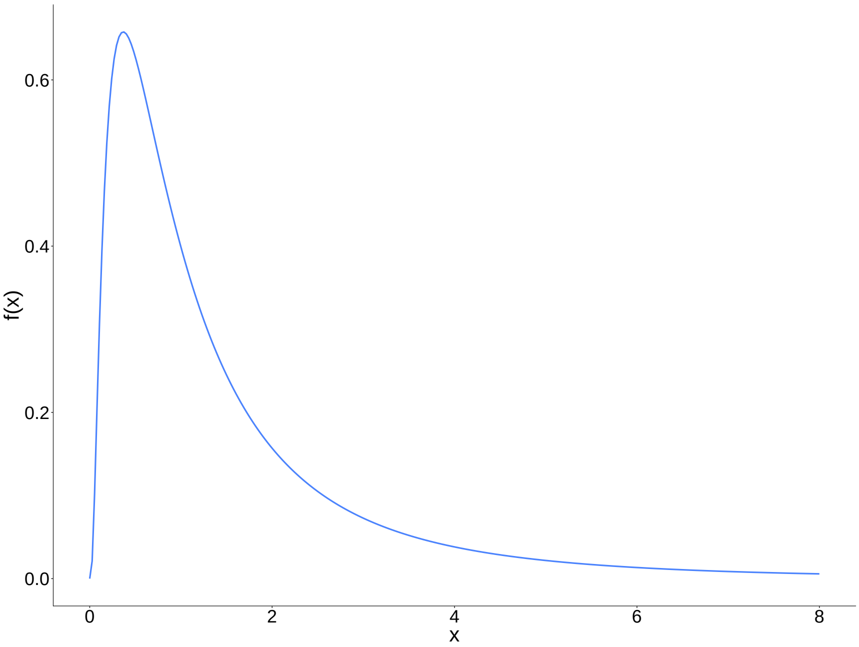

FIGURE 7.1: Probability density function of the lognormal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$

The mean and variance of the lognormal distribution are expressed respectively as:

$$\mathbb{E}(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{7.8}$$

$$\text{var}(X) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1). \tag{7.9}$$

7.2 Predictive Distribution of y

Up until now, for the spatial and spatio-temporal p-spline models, having a formula for the covariance of the basis coefficients, $\mathbf{C}_{\hat{\alpha}|y}$, up to a multiplicative constant has been sufficient for computing the integrated prediction variance when assuming additive error on the data. However, in order to obtain an expression for the lognormal variance that can be subsequently integrated when working with data with multiplicative error, the full predictive distribution of $\mathbf{y}$ is required. This is necessary as the variance and mean of this distribution are embedded within an exponential function.

For a given matrix of basis functions, $\widetilde{\mathbf{B}} \in \mathbb{R}^{r\times m}$ computed over a set of r prediction locations, the predicted response can be described as $\widetilde{\mathbf{Y}}|\boldsymbol{\alpha}, \sigma^2 \sim \mathcal{N}(\widetilde{\mathbf{B}}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_r)$. Therefore, the predictive distribution, using the distributional assumptions given in Section 2.3.2, can be shown to be, for $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{y}}$,

$$\begin{aligned} f_{\widetilde{\mathbf{Y}}|\mathbf{Y}} &= \int f_{\widetilde{\mathbf{Y}}|\boldsymbol{\alpha},\sigma^2} f_{\boldsymbol{\alpha},\sigma^2|\mathbf{y},M_\lambda} \; d\boldsymbol{\alpha} d\sigma^2 \\ &= \int \mathcal{N}(\widetilde{\mathbf{B}}\boldsymbol{\alpha}, \sigma^2\mathbf{I}_r) \times \mathcal{NIG}(\boldsymbol{\mu}^*, \mathbf{V}^*, a^*, b^*) \; d\boldsymbol{\alpha} d\sigma^2 \\ &= \mathcal{MVS}t_{2a^*}\left(\widetilde{\mathbf{B}}\boldsymbol{\mu}^*, \frac{b^*}{a^*}\left(\mathbf{I} + \widetilde{\mathbf{B}}\mathbf{V}^*\widetilde{\mathbf{B}}^\top\right)\right). \end{aligned} \tag{7.10}$$

This is multivariate Student-t distribution, $\mathcal{MVS}t$, with p.d.f:

$$f_{\widetilde{\mathbf{Y}}|\mathbf{Y}} = \frac{\Gamma\left((\nu+p)/2\right)}{(\pi\nu)^{p/2}\Gamma(\nu/2)|\boldsymbol{\Sigma}|^{1/2}}\left[1 + \frac{1}{\nu}(\widetilde{\mathbf{y}} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\widetilde{\mathbf{y}} - \boldsymbol{\mu})\right]. \tag{7.11}$$

The parameters of this distribution,

- Location, $\boldsymbol{\mu}$

$$\boldsymbol{\mu} = \widetilde{\mathbf{B}}\boldsymbol{\mu}^*$$

- Shape, $\boldsymbol{\Sigma}$

$$\Sigma = \frac{b^*}{a^*}\left(\mathbf{I} + \widetilde{\mathbf{B}}\mathbf{V}^*\widetilde{\mathbf{B}}^\top\right)$$

- Degrees of Freedom, ν

$$\nu = 2a^*.$$

where, for a p-splines model:

$$\begin{aligned}
\mathbf{V}^* &= (\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{D}^\top\mathbf{D})^{-1} \\
\boldsymbol{\mu}^* &= \mathbf{V}^*(\mathbf{B}^\top\mathbf{y}) \\
a^* &= a + \frac{n}{2} \\
b^* &= b + \frac{1}{2}\mathbf{y}^\top\left[\mathbf{I}_n - \mathbf{B}(\mathbf{B}^\top\mathbf{B} + \lambda\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{B}^\top\right]\mathbf{y}.
\end{aligned} \tag{7.12}$$

For a more detailed derivation, see Equations 2.46.

It is well known that as $\nu \to \infty$, the multivariate Student-t Distribution becomes the multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, see Bishop [2006] and Kotz and Nadarajah [2004]. Figure 7.2 shows this concept in action with the density of a bivariate normal distribution overlayed with the bivariate Student-t distribution, with increasing degrees of freedom (ν).

As the degrees of freedom increase, the density function of the multivariate Student-t tends towards the multivariate normal. By $\nu = 30$ the densities are almost identical.

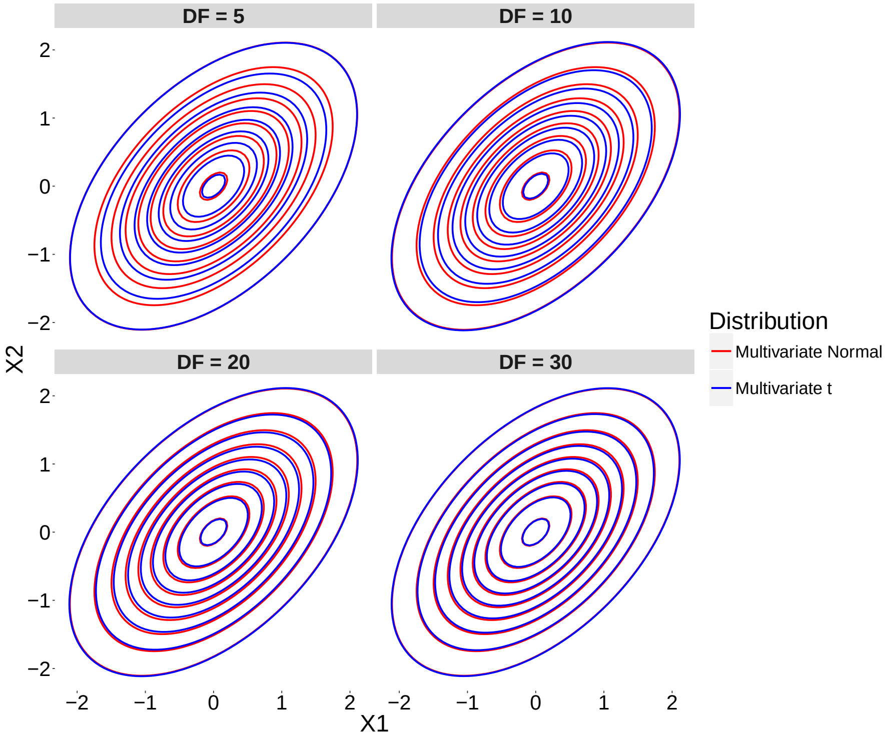

Figure 7.2: Bivariate normal densities overlayed with bivariate student-t densities with increasing degrees of freedom $\nu = 5, 10, 20, 30$.

In the case of the predictive distribution presented here for a p-splines model, $\nu = 2a^*$ with $a^* = a + \frac{n}{2}$. For the spatial models, the number of observations, n, is almost always ≥ 10, therefore, regardless of the value of a, ν will be ≥ 10, indicating the multivariate normal distribution can potentially be used to approximate the multivariate Student-t distribution. This approximation is more robust for spatio-temporal models where the number of observations used to build the model is generally always ≥ 100, therefore again, regardless of a, ν will be ≥ 100. Hence, the multivariate normal distribution with mean, $\boldsymbol{\mu}$ and variance, $\boldsymbol{\Sigma}$, can be used to approximate the multivariate Student-t distribution.

Making use of this idea, the approximate full predictive distribution of $\mathbf{Y}$ is then,

$$\widetilde{\mathbf{Y}}|\mathbf{Y}, \boldsymbol{\alpha}, \sigma^2 \sim \mathcal{N}\left(\widetilde{\mathbf{B}}\hat{\boldsymbol{\alpha}}, \frac{b^*}{a^*}\left(\mathbf{I} + \widetilde{\mathbf{B}}\mathbf{V}^*\widetilde{\mathbf{B}}^\top\right)\right). \tag{7.13}$$

7.3 IV Objective Function using the Lognormal Variance

From the definition of the lognormal, given in Section 7.1, if it is assumed

$$\log(\widetilde{\mathbf{Y}}|\mathbf{Y}, \boldsymbol{\alpha}, \sigma^2) \sim \mathcal{N}\left(\widetilde{\mathbf{B}}\hat{\boldsymbol{\alpha}}, \frac{b^*}{a^*}\left(\mathbf{I} + \widetilde{\mathbf{B}}\mathbf{V}^*\widetilde{\mathbf{B}}^\top\right)\right)$$

then,

$$\widetilde{\mathbf{Y}}|\mathbf{Y}, \boldsymbol{\alpha}, \sigma^2 \sim \mathcal{LN}\left(\widetilde{\mathbf{B}}\hat{\boldsymbol{\alpha}}, \frac{b^*}{a^*}\left(\mathbf{I} + \widetilde{\mathbf{B}}\mathbf{V}^*\widetilde{\mathbf{B}}^\top\right)\right). \tag{7.14}$$

Using the formulae for the mean and variance of a lognormal distribution, given in Equations 7.8 and 7.9 respectively, the mean and the variance of $\widetilde{\mathbf{Y}}|\mathbf{Y}, \boldsymbol{\alpha}, \sigma^2$ are:

- Mean:

$$\exp\left(\widetilde{\mathbf{B}}\hat{\boldsymbol{\alpha}} + \frac{\frac{b^*}{a^*}\left(\mathbf{I} + \widetilde{\mathbf{B}}\mathbf{V}^*\widetilde{\mathbf{B}}^\top\right)}{2}\right) \tag{7.15}$$

- Variance:

$$\exp\left(2\widetilde{\mathbf{B}}\hat{\boldsymbol{\alpha}} + \frac{b^*}{a^*}\left(\mathbf{I} + \widetilde{\mathbf{B}}\mathbf{V}^*\widetilde{\mathbf{B}}^\top\right)\right)\left(\exp\left(\frac{b^*}{a^*}\left(\mathbf{I} + \widetilde{\mathbf{B}}\mathbf{V}^*\widetilde{\mathbf{B}}^\top\right)\right) - 1\right). \tag{7.16}$$

To compute the integrated variance i.e. the integral of Equation 7.16, ideally the formula would be decomposed and simplified into smaller more manageable sections in a similar manner to the derivations in Section 6.4. However, due to the spline functions being embedded within exponential functions, a closed form expression is not available and thus numerical integration is required.

Comparing Equation 7.16, which is to be integrated, with the integrated variance functions for spatial and spatio-temporal p-spline models given in Equations 6.34 and 6.40 respectively, they are clearly very different. The most important difference is the fact that the lognormal variance depends on what has previously been seen through the fitted values i.e. $\widetilde{\mathbf{B}}\hat{\boldsymbol{\alpha}}$, whereas the objective functions for data with additive error, shown in Equations 6.34 and 6.40, only depend on where samples have previously been taken, through the matrices of basis functions, and not their recorded values.

Based on the formulation of the variance of the lognormal it is expected that the optimal design will be made up mainly of wells located where the predicted concentrations are higher.

7.4 Results

Two datasets were considered for this study with differing contaminant plume complexities.

The first, referred to as ‘Dataset 1’, used the same PDE and Design 1, presented in Section 6.5.1. This plume has a very simple shape and only moves marginally in a easterly direction. 10% measurement error was added to the observed data on the log scale to represent multiplicative error. This error value represents noise as a result of sampling and analytical variations and is based on a comparison of blind duplicate samples in a large unpublished groundwater quality dataset for a Shell site (personal communication J. Smith, Shell Global Solutions).

The second dataset, referred to as 'Dataset 2', used PDE2, simulated in Section 4.5.1, with a well network made up of 14 monitoring wells, shown in Figure 7.3. This plume travels further than the plume used in Dataset 1 and the direction of travel is south-westerly. Observations were obtained by interpolating the PDE at these well locations and 40 time points and, as was done with Dataset 1, a random 25% of the data were removed. 10% measurement error was added to the data on the log scale to represent multiplicative error.

The Integrated lognormal Variance objective function (ILNV objective function), and these two datasets, were used to optimise sampling designs with spatial and spatio-temporal p-spline models. Modelling was performed on the log transformed response and design optimisations were carried out for three numbers of wells. For comparison, the methodology set out in Chapter 6 was also used to optimise designs for the same number of wells, again using the log transformed response. The primary aim of this study was to determine whether the optimal designs differ if the objective function is calculated on the original scale, given the response has been log transformed for modelling (ILNV objective function), compared with the objective function being calculated on the log scale (IV objective function), again given modelling was also performed on the log scale.

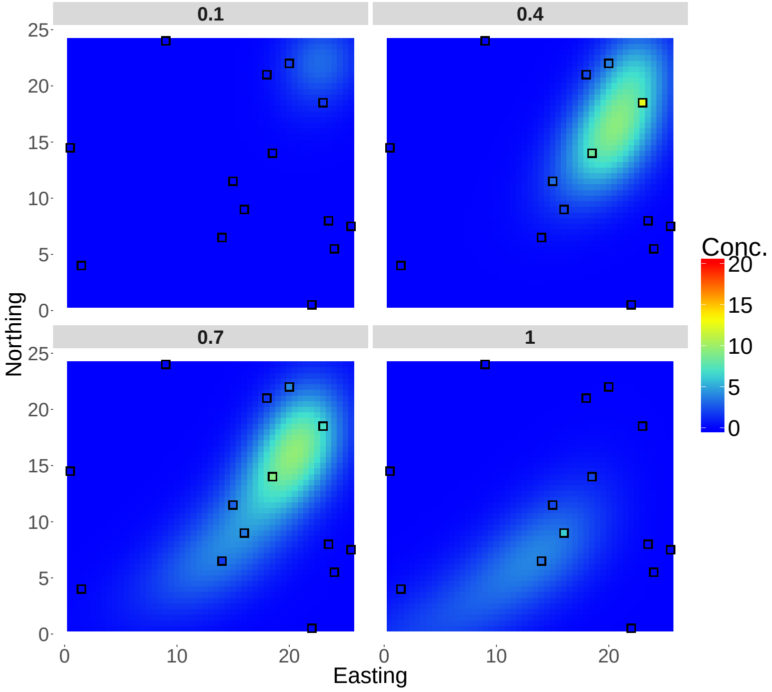

FIGURE 7.3: True surface simulated from the PDE, detailed in Section 4.5.1, at times $t \in \{0.1, 0.4, 0.7, 1\}$ and the well network used for Dataset 2.

7.4.1 Design for a Spatial P-splines Model

For the sampling design optimisation using the spatial p-splines model, degree three basis functions were used with a first order difference penalty. The smoothing parameter was estimated using the MAP estimate, detailed in Section 2.3.2.

Dataset 1

Figure 7.4 shows the optimal designs when computing the IV objective function on the log and original scales i.e. using the IV and ILNV objective functions respectively. The most recent prediction using the spatial p-splines model is displayed in the background to give an idea of where high concentrations were predicted at the last sampling event.

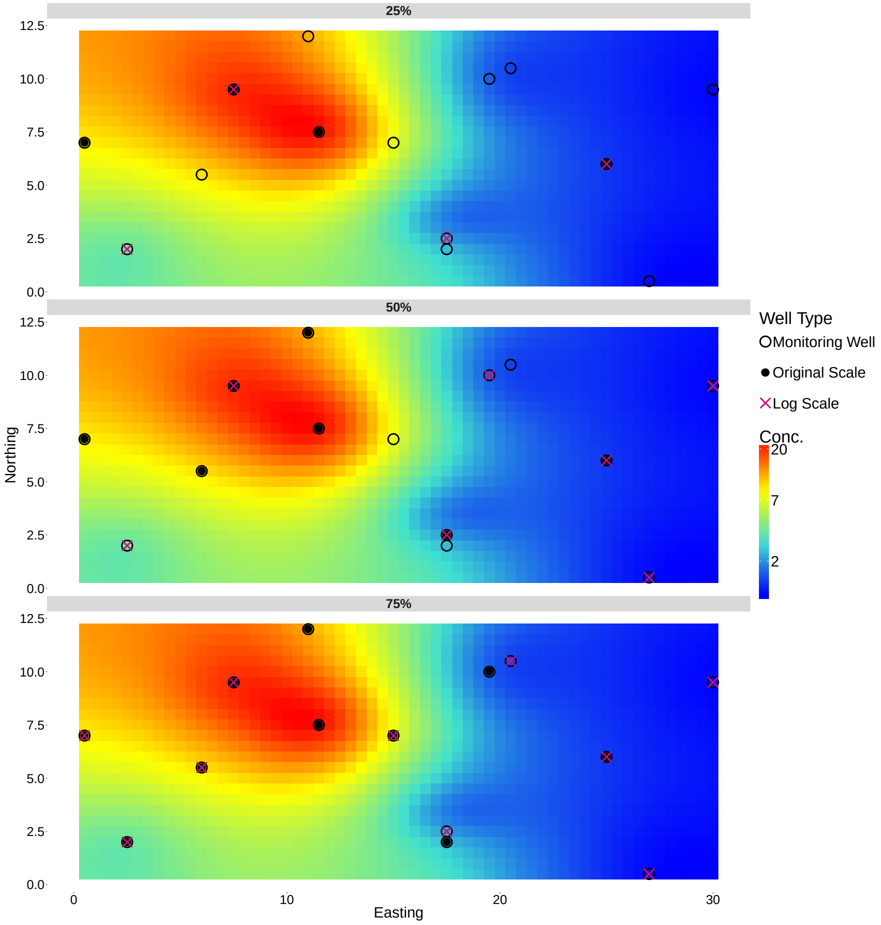

FIGURE 7.4: Optimal designs for the integrated variance (IV) objective function, computed using the lognormal distribution (ILNV objective function) and the methodology set out in Chapter 6 (IV objective function), when optimising for 25%, 50% and 75% of the wells using a spatial p-splines model on Dataset 1. The surface beneath the wells shows the prediction from the previous sampling time.

The optimal designs, when optimising for all three numbers of wells, differ significantly depending on whether the objective function is to be computed on the original or log scales. When using the ILNV objective function, i.e. computing the objective function on the original scale (filled black circles) the wells located where the contaminant was predicted to be high are favoured, as anticipated. On the other hand, the optimal sampling designs for the IV objective function are spread out across the study region for all three optimisations. For example, when optimising for seven wells (50%), five of the wells in the ILNV functions optimal design are positioned on the left of the study region where the contaminant plume was predicted to be. For the IV objective function, the seven optimal wells are positioned more evenly across the study region with only two wells being located at the contaminant plume. It is worth remembering that computation and optimisation of the IV objective function does not depend on the location or value of any previously seen data.

Table 7.1: Cross-tabulation of the Integrated lognormal Variance (ILNV) and Integrated Variance (IV) objective functions and their corresponding optimal designs for Dataset 1. Optimisation was carried out for 25%, 50% and 75% of the total number of wells using a spatial p-splines model, with modelling being performed on the log scale. Bold entries indicate the design with the minimum of each objective function.

		25%		50%		75%	
		Computation Scale					
		Original	Log	Original	Log	Original	Log
Value	ILNV	**20573**	31053	**12411**	28172	**10125**	15027
	IV	228.6	**203.5**	173.6	**147.1**	133.0	**114.3**

Table 7.1 cross tabulates each objective function (ILNV and IV) and their corresponding optimal designs. The analytical results in this table indicate that the scale for which the objective function is to be calculated on should be taken into account when deciding on which version of the objective function to use for optimisation. For this design, assuming computation of the objective function is on the original scale, the integrated prediction variance for a sampling design optimised taking this into account, i.e. using the ILNV objective function, is up to $\sim$ 45% lower compared with the prediction variance for the sampling design optimised assuming computation is on the log scale i.e. using the IV objective function.

Dataset 2

The features of the optimal designs highlighted for Dataset 1 were also present for Dataset 2, shown in Figure 7.5. As was the case with Dataset 1, the ILNV objective function favours wells located over the area where the contaminant concentrations were estimated to be highest in the previous sampling event, in this case the top right corner. Again, the IV objective function opts for wells that give good spatial coverage of the region.

Table 7.2 cross-tabulates the ILNV and IV objective functions with their corresponding optimal designs for Dataset 2. The results for this dataset again highlight the importance of ensuring that the scale of computation of the objective function is taken into account when deciding on which version of the IV objective function to use. If interpretation is to be on the original scale but modelling has been performed on the log scale, the integrated prediction variances of the sampling designs optimised which take this into account i.e. using the ILNV function, are up to 40% lower compared with the designs optimised using the IV function.

Table 7.2: Cross-tabulation of the Integrated lognormal Variance (ILNV) and Integrated Variance (IV) objective functions and their corresponding optimal designs for Dataset 2. Optimisation is carried out for 25%, 50% and 75% of the total number of wells using a spatial p-splines model, with modelling being performed on the log scale. Bold entries indicate the design with the minimum of each objective function.

		25%		50%		75%	
		Computation Scale					
		Original	Log	Original	Log	Original	Log
Value	ILNV	**10506**	21158	**6373**	10946	**4827**	5624
	IV	383.3	**348.1**	287.2	**245.6**	221.0	**207.9**

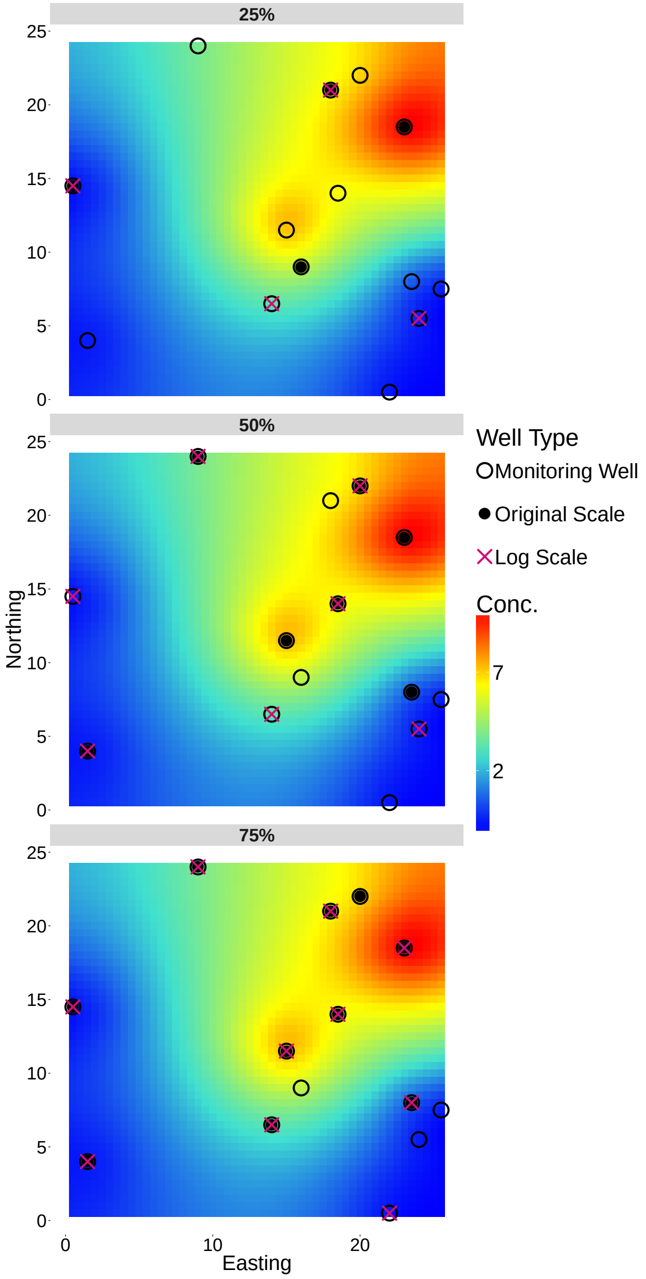

FIGURE 7.5: Optimal designs for the integrated variance (IV) objective function, computed using the lognormal distribution (ILNV objective function) and the methodology set out in Chapter 6 (IV objective function), when optimising for 25%, 50% and 75% of the wells using a spatial p-splines model on Dataset 2. The surface beneath the wells shows the prediction from the previous sampling time.

7.4.2 Design for a Spatio-temporal P-splines Model

For the sampling design optimisations performed using a spatio-temporal p-splines model, the time of the next sampling event was fixed at 5% of the current temporal range into the future. The model was fitted with 13 basis functions for the easting and temporal components and the number of basis functions for the northing component was scaled by the dimensions of the study region. The two smoothing parameter spatio-temporal p-splines model, presented in Chapter 4, was used along with a first order penalty.

Dataset 1

Figure 7.6 shows the optimal designs when computing the IV objective function on the original and log scales using Dataset 1 and a spatio-temporal p-splines model.

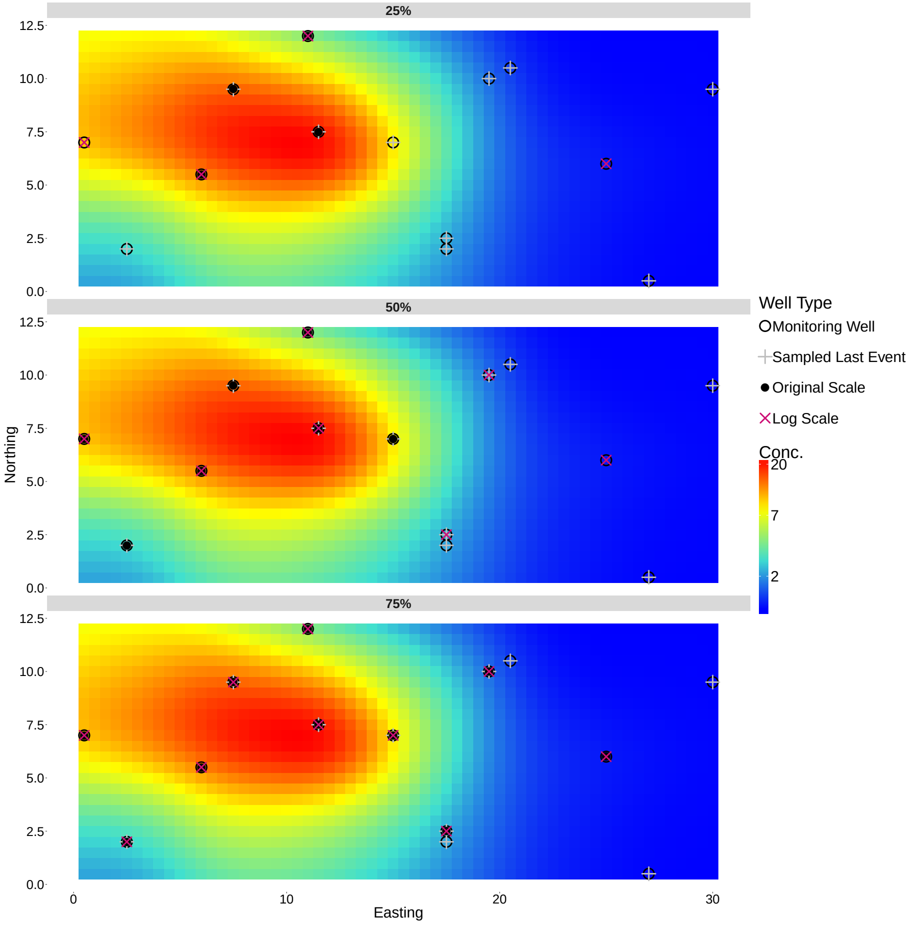

FIGURE 7.6: Optimal designs for the integrated variance (IV) objective function, computed using the lognormal distribution (ILNV objective function) and the methodology set out in Chapter 6 (IV objective function), when optimising for 25%, 50% and 75% of the wells using a spatio-temporal p-splines model on Dataset 1. The surface beneath the wells shows the prediction from the previous sampling time.

The differences that were observed between the designs from the spatial model can also be seen in the optimal sampling designs from the spatio-temporal model. Although the sampling designs for each computation scale differ by a few wells, the positioning of these differences are important. The sampling designs for the objective function calculated on the original scale, optimised using the ILNV objective function, clearly favour wells located on the left of the study region where the contaminant concentrations are predicted to be higher. In contrast the designs for computation of the objective function on the log scale, using the IV objective function, are more spread out across the study region and attempt to ensure all wells are sampled during the current and previous sampling events.

Table 7.3 cross-tabulates each version of the IV objective function with their corresponding optimal designs. The % reduction in integrated variance between designs is not nearly as significant for a spatio-temporal model compared with the results from the spatial model, with the largest difference being $\sim 5\%$. This is probably due to the spatio-temporal model carrying forward information from previous sampling events into the regions where wells, that were not chosen in the selected designs, are located. This in turn reduces the uncertainty associated with the predictions in these areas. However Table 7.3 does show that the sampling designs optimised using the ILNV objective function always have a lower integrated prediction variance when it is computed on the original scale compared with when it is computed on the log scale i.e. the IV design, when the designs differ.

Table 7.3: Cross-tabulation of the Integrated lognormal Variance (ILNV) and Integrated Variance (IV) objective functions and their corresponding optimal designs for Dataset 1. Optimisation is carried out for 25%, 50% and 75% of the total number of wells using a spatio-temporal p-splines model, with modelling being performed on the log scale. Bold entries indicate the design with the minimum of each objective function.

		25%		50%		75%	
		Computation Scale					
		Original	Log	Original	Log	Original	Log
Value	ILNV	**3061.7**	3231.1	**2956.4**	3079.8	2947.4	2947.4
	IV	53.3	**52.5**	51.9	**51.0**	49.7	49.7

Dataset 2

The sampling designs optimised for Dataset 2 using the spatio-temporal model showed similarities to those of the spatial model. Again, wells located near where the contaminant plume was predicted to be in the previous event were favoured by the ILNV objective function, with the 25% and 50% optimisations only choosing wells located in this region. As was previously seen in Dataset 1, the IV objective function attempts to obtain samples from all of the wells during the current and previous sampling events.

The cross-tabulation results for Dataset 2 and the spatio-temporal model, shown in Table 7.4, show similar trends to those previously seen for this model and Dataset 1. If the objective function is to be interpreted on the original scale whilst modelling is performed on the log scale, then use of an objective function that reflects this, namely the ILNV function, reduces the integrated prediction variance by up to 5% compared with the IV function, which calculates the objective function on the log scale.

Table 7.4: Cross-tabulation of the Integrated lognormal Variance (ILNV) and Integrated Variance (IV) objective functions and their corresponding optimal designs for Dataset 2. Optimisation is carried out for 25%, 50% and 75% of the total number of wells using a spatio-temporal p-splines model, with modelling being performed on the log scale. Bold entries indicate the design with the minimum of each objective function.

		25%		50%		75%	
		Computation Scale					
		Original	Log	Original	Log	Original	Log
Value	ILNV	**1370.7**	1433.4	**1323.9**	1391.3	**1311.7**	1333.1
	IV	91.5	**88.9**	89.6	**87.2**	86.0	**86.0**

7.4.2.1 The Effect of Previous Sampling Frequency

In Section 6.5.3 of Chapter 6 the effect previous well-specific sampling frequency had on the next sampling design was investigated. It was found that wells which had not been sampled recently were favoured over those which had more recent samples. The previous study, in Section 7.4.2, found that when the Integrated lognormal Variance is used as the objective function, wells where high concentrations have been predicted are favoured in the selected design over those where lower concentrations were predicted. Thus, it

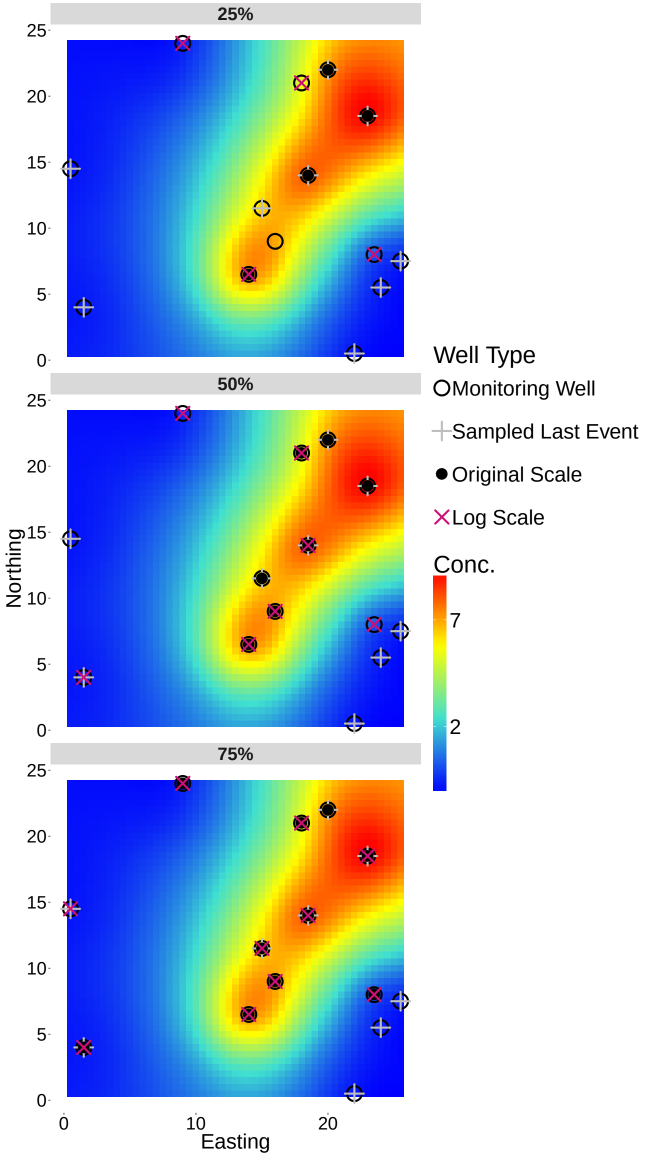

FIGURE 7.7: Optimal designs for the integrated variance (IV) objective function, computed using the lognormal distribution (ILNV objective function) and the methodology set out in Chapter 6 (IV objective function), when optimising for 25%, 50% and 75% of the wells using a spatio-temporal p-splines model on Dataset 2. The surface beneath the wells shows the prediction from the previous sampling time.

was of interest to determine whether these 'high prediction' wells would still be favoured if the data being used to build the model contained 'low frequency wells', namely wells where only a few samples were taken at the beginning of the time window. The first dataset used for the study in Section 6.5.3, where seven wells were randomly allocated to the 'low frequency well' set, was again used here, but this time multiplicative error was added and the response was log transformed prior to modelling.

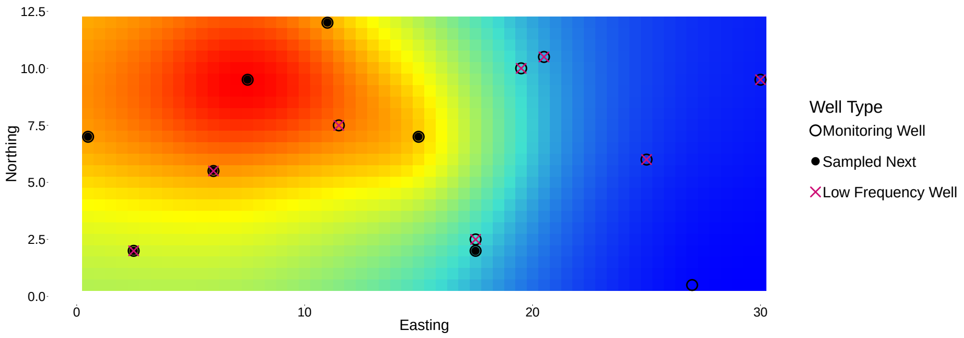

Figure 7.8: Optimal design for the integrated variance (IV) objective function computed using the log-normal distribution and a spatial p-spline model when optimising for 7 wells using the dataset with a 'low frequency well' set.

Figure 7.8 shows the resulting design optimised using a spatio-temporal p-splines model and the Integrated lognormal Variance objective function. From this plot it is apparent that the 'low frequency wells' are not favoured for the next design, with the 'high concentration' wells still being preferred for the next sampling event.

7.5 Summary

Optimising sampling designs for data which are assumed to have multiplicative error results in designs being chosen that are different from those when the data are assumed to have additive error, when the objective function is interpreted on the same scale. Commonly, to allow for modelling data with multiplicative error using standard methods, the response variable is log transformed to give an additive interpretation of the error. These models, with the log transformed response, can subsequently be used to optimise sampling designs. However, the scale on which the design objective function is to be computed must be determined prior to the design optimisation, as the resulting designs differ depending on whether the computation of the objective function is on the original or log scale.

Assuming multiplicative error on the data and that modelling is performed on the log scale, when the objective function is calculated on the original scale the optimal design tends to favour wells located in the region where, in the previous sampling event, concentrations of the contaminant were predicted to be highest. This is due to the variance of the lognormal distribution, which is the predictive distribution, depending on both the mean and variance of the log transformed response. When the objective function is to be calculated on the log scale, in the spatio-temporal setting the objective function only depends on where observations were previously recorded, not their recorded values and the resulting designs tend to be space filling and favour observations that were not sampled in the previous event. The results seen in Chapter 6 illustrate this. In the spatial setting, similar results are seen, with wells where predictions were highest being favoured when calculating the objective function on the original scale. For the log scale calculation, previously recorded data are only used to determine the smoothing parameter and the resulting designs try to give good spatial coverage, as seen in Chapter 6.

Chapter 8

Discussion & Future Work

The main aim of this thesis was to develop spatio-temporal models for groundwater contamination data that could subsequently be used to optimise sampling designs on monitoring networks. Methods have been compared, applied and developed to produce spatio-temporal models that can be efficiently built and two design objective functions have been presented for optimising sampling designs over a fixed monitoring network for data with additive and multiplicative errors.

8.1 Spatio-temporal P-spline Models

There are many advantages to using a spatio-temporal model over multiple temporally independent spatial models. These were demonstrated in the study in Chapter 3, with the main benefit being the improvement in prediction accuracy and potential reduction in the amount of data that need to be sampled whilst retaining the predictive accuracy.

In earlier work by Molinari [2014], a single smoothing parameter spatio-temporal p-splines model was used with the number of basis functions for the temporal dimension being chosen to reflect the belief that the contaminant concentrations vary more across space than they do over time. By formulating the model in such a way, linear algebra methods could be used to efficiently estimate the smoothing parameter. However, deciding on a scaling rule for the number of temporal basis functions is subjective and makes justification of the one smoothing parameter problematic. Given that space and time are measured on different scales it may not even be considered appropriate to scale one based on the other and thus a model which controls the smoothness of the temporal component separately from the spatial components is more appealing. Chapter 4 developed an algorithm for tuning two smoothing parameters that avoided the computationally expensive grid search that is required in the naive approach for estimating multiple parameters. Initially it was hoped that the contours of the optimisation surface would lie parallel with one of the axes, allowing one parameter to be tuned whilst fixing the other. This would allow the efficient methodology set out by Molinari [2014] still to be used for parameter estimation through an augmented data formulation of the second smoothing parameter. However, this was not the case upon examination of several hypothetical and real groundwater datasets. Interestingly though, there did appear to be a trend in the optimal combination of smoothing parameters as the number of basis functions in the model increased. This trend was exploited, with a low basis resolution p-splines model being used to find starting points for the smoothing parameter optimisation on a coarse grid. Theses starting points then allowed each smoothing parameter to be tuned separately using a model with the desired basis resolution and the methodology set out by Molinari [2014]. The efficiency of tuning two parameters is unlikely ever to be equivalent to that of tuning one parameter, but the improvement in modelling flexibility out weighs this added computational expense.

Inevitably there are also some drawbacks to using spatio-temporal p-spline models. Several steps have been taken to improve the computational speed of estimating the parameters in the spatio-temporal model but ultimately the computational speed of the spatial model is unlikely to be matched. The tensor product structure of the spatio-temporal basis results in significant increases in the computational effort when only a few extra basis functions are added to enhance the models flexibility. This subsequently causes the computation time to increase exponentially and given it is recommended that a two parameter model is used, the time taken to estimate the parameters can very quickly become unmanageable.

Chapters 3 and 4 highlighted, and Chapter 5 investigated, an effect known as ‘ballooning’ which appears in some of the spatio-temporal p-spline models predictions. The simulation study, and the measure developed to flag up when ballooning might be present, found that the likelihood of ballooning occurring reduced when the number of basis functions was increased. However, increasing the number of basis functions begins to require a trade-off between obtaining robust predictions and the computational time required to achieve these predictions. Alternatively, increasing the smoothing parameter seems to contain ballooning.

Rather than trying to highlight when ballooning is present, a new penalty was proposed aimed at trying to stop ballooning from happening in the first place. The penalty was based on the idea that although the contaminant plume would move around in space, its mass should not change significantly over time, i.e. the plume should just diffuse out over the study region. Thus, if a sudden change in the mass is detected through ballooning, the penalty should suppress this unusual fluctuation in the mass. Two datasets were simulated to assess the effectiveness of the penalty. The first was made up of one spatial dimension and one temporal dimension. Studies of this dataset suggest the penalty was working in the intended manner. Problems began to arise, however, with the second dataset, made up of two spatial dimensions and one temporal dimension. This dataset had a point in time where the wells suddenly picked up the contamination and thus the predicted contaminant mass would suddenly increase. It was hoped that the penalty would suppress this sudden increase in mass, but instead the model forced mass into the earlier time points, in regions with no well coverage. In hindsight this is unsurprising since the simulated dataset is not mimicing exactly what happens when ballooning occurs since there are data to support this increase in mass.

The penalty itself is doing what was hoped in conserving the contaminant mass. This can be seen by looking at the total mass over time plotted in Figure 5.12. However, the model does not know where to store the excess mass and so places it where there are no data. The results suggest that the penalty may not be a solution for ballooning however, it could be used to help inform models of where contamination is located if it is not detected by the monitoring wells until later sampling time. Therefore, future work could involve extending the proposed penalty by incorporating information on where the contamination is likely to be located if there is a sudden change in mass with data support. Methodology developed by Frasso [2013], Frasso et al. [2016*a*] and Frasso et al. [2016*b*] seeks to build spline penalties based on differential equations (DEs) for DE parameter estimation and the analysis of dynamic systems. These types of penalties can potentially be incorporated into the mass penalty to inform the model of the groundwater flow and thus on where to place excess mass that may suddenly appear with data support.

8.2 Monitoring Network Design

The second half of the thesis focussed on optimising sampling designs for fixed monitoring networks using spatial and spatio-temporal models. Two design objective functions were considered for determining the optimal sampling design for the next sampling event, namely, the variance of the plume mass (VM) and the integrated prediction variance (IV).

Variations of these objective functions are widely developed in the literature for Kriging-based models, but to the best of our knowledge these have not yet been developed for spline-based models. Here a primary focus is on spatial and spatio-temporal p-spline models. The sampling design optimisations were carried out subject to three constraints:

1. It is assumed that the monitoring well network is in place and fixed,
2. The number of wells to be sampled in the next event is also fixed,
3. The time of the next sampling event is predetermined.

The resulting optimal sampling designs differed for each model and objective function combination. When using the spatial and spatio-temporal p-splines model, the VM objective function appeared to favour wells located in the centre of the study region whilst the IV function tended to choose a sampling design that gave good spatial coverage of the region. Cross-tabulation of the objective functions and their optimal designs highlighted the importance of choosing an appropriate objective function that answers the aim of the study. If the interest is in reducing the prediction variance optimisation of the sampling design should be through the IV objective function. The sampling designs chosen by the spatio-temporal model appeared to select wells such that the majority of the wells are sampled during the current and previous sampling events. Simulation studies also indicated that wells which had not been sampled recently were favoured for the next sampling design over those that had been sampled more recently and frequently.

There are several lines for future work on the sampling design aspect. Currently the well network is exhaustively searched to determine the optimal combination of wells. This is adequate for the size of the network used here and ensures the optimal combination of wells is not missed. However, as the size of the network increases, computing the objective functions for $\begin{pmatrix} N \\ n \end{pmatrix}$ combinations of wells, where N is the size of the network and n is the number of wells to be sampled, quickly becomes too computationally intensive and thus adopting an optimisation method would be more appropriate. Commonly, genetic algorithms are used, as discussed in the literature review presented in Section 6.1. Alternatively, another widely used method for optimisation is, spatial simulated annealing (SSA), which is discussed by Helle and Pebesma [2012] and used successfully by Abida et al. [2008] and Heuvelink et al. [2010] for optimising air quality monitoring networks. Although these methods are frequently used for optimising the positions of the wells from a dense grid, in essence the idea is the same for optimising over a set of fixed locations, for example see Zhang et al. [2005].

To aid the calculations of the integrals contained within the objective functions, for now the domain over which they were calculated was taken as the rectangle around the outermost wells. However, by setting the domain to this shape, areas of the study region, in particular the corners, have little or no data support, resulting in higher prediction variances in these regions. Ideally the region of the domain contained within the convex hull of the wells would be where the objective functions would be computed. Figure 8.1 shows the current rectangle domain and the suggested convex hull domain for Design 1 from Chapter 6.

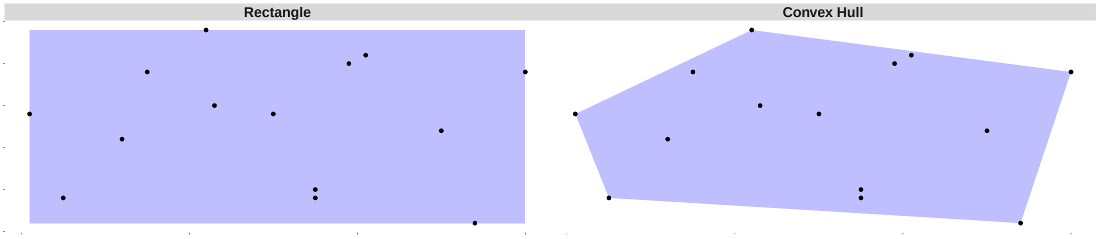

FIGURE 8.1: Rectangle domain currently used for design optimisation and the proposed convex hull domain

Computing the objective functions over the convex hull of the wells is complex and requires partial integrals of the b-spline functions to achieve the more complicated domain shape. This adaptation of the domain shape is scope for future work.

8.2.1 Accounting for Multiplicative Errors

Finally, the case where the observed data are assumed to have multiplicative error, a common assumption for groundwater contamination data, was considered. Throughout this thesis, with the exception of Chapter 6, it has been assumed that the data have multiplicative error and thus the response was log transformed prior to modelling to give an additive interpretation of the error. Once a final model had been achieved, the fitted and predicted values were exponentiated back onto the original scale for interpretation. This method of transforming for modelling then transforming back for interpretation does not, however, hold for the design objective functions as a result of the integral terms. Use of the log transformed response and the objective functions designed for data with additive error, developed in Chapter 6, is appropriate if the objective functions are also to be calculated on the log scale. However, commonly it is often of interest to calculate the objective functions on the original scale whilst modelling on the log scale.

Chapter 7 employed the Normal approximation to the Student-t distribution, for a suitably high degrees of freedom parameter, and the lognormal distribution to allow the IV objective function to be calculated on the original scale given modelling was performed on the log scale. The lognormal variance function, shown in Equation 7.16, that was to be integrated, is significantly more complex than the IV function derived in Chapter 6 for additive error and no closed form solution was available. Thus numerical integration was required. The results indicated the designs differed depending on whether the IV objective function was computed on the original scale or the log scales, with the former favouring wells located where concentrations were predicted to be highest and the latter choosing wells that gave good spatial coverage. Therefore, future work could seek a more efficient way of calculating the integrated lognormal variance.

Figure 8.2 displays a flow chart that can be used to aid in deciding which set of objective functions to use when the error type is known, and if it is multiplicative, the desired scale of computation is also known.

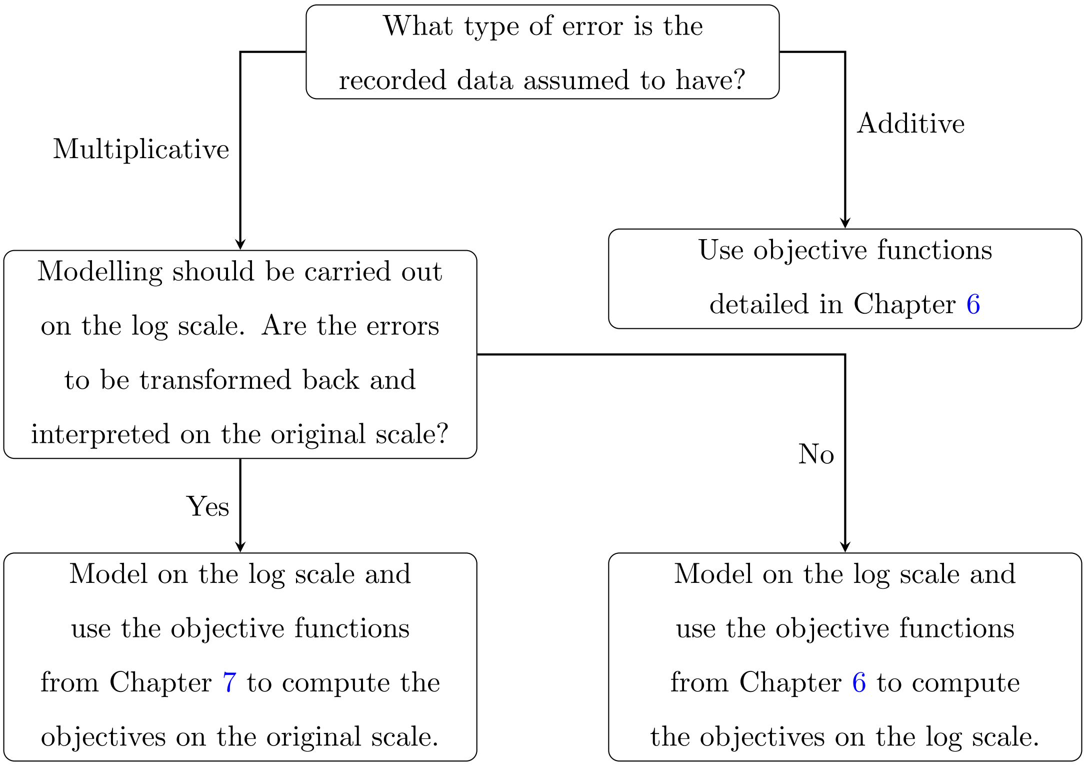
What type of error is the recorded data assumed to have?
Multiplicative
Additive
Modelling should be carried out on the log scale. Are the errors to be transformed back and interpreted on the original scale?
Use objective functions detailed in Chapter 6
No
Yes
Model on the log scale and use the objective functions from Chapter 7 to compute the objectives on the original scale.
Model on the log scale and use the objective functions from Chapter 6 to compute the objectives on the log scale.

FIGURE 8.2: Flow chart for deciding which objective function to use based on the assumed error type.

Due to the integral of the mean function being embedded within an exponential function, calculating the VM objective function for this scenario was unachievable at the current time. Establishing a way of calculating the VM objective function for a lognormal response would be another interesting future project.

Appendix A

Statistical Theorems and Properties

A.1 Conditional Distribution Property of a Multivariate Gaussian Distribution

Given the joint distribution of $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ is,

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right). \tag{A.1}$$

Then, the conditional distribution of $\mathbf{X}_1|\mathbf{X}_2$ is:

$$\mathbf{X}_1|\mathbf{X}_2 \sim \mathcal{N}\left(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right). \tag{A.2}$$

A.2 Variance of a linear combination of variables

Given random variables $X_1, \cdots, X_n$ and constants $a_1, \cdots, a_n$ the variance of the sum of the random variables X_i scaled by their corresponding constant a_i is:

$$\mathrm{var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{cov}(X_i, X_j) \tag{A.3}$$

Appendix B

Matrix Properties

B.1 The Spectral decomposition

Given matrix A of dimension $n \times n$. There exists n vectors $\boldsymbol{\gamma}_i$ with corresponding scalar values $\delta_i \geq 0$ such that:

$$A\boldsymbol{\gamma}_i = \delta_i \boldsymbol{\gamma}_i \tag{B.1}$$

The δ_i scalars are known as *eigenvalues* of A and the corresponding vectors, $\boldsymbol{\gamma}_i$ are known as *eigenvectors*.

$$A = \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}^\top \tag{B.2}$$

where the columns of the orthogonal matrix $\boldsymbol{\Gamma}$ are the vectors $\boldsymbol{\gamma}_i$ and $\boldsymbol{\Delta}$ is a diagonal matrix with the δ_is as its diagonal entries.

B.2 Kronecker product properties

- Let matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ be given matrices of such size that the matrix products $\mathbf{AC}$ and $\mathbf{BD}$ exist, then

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD}) \tag{B.3}$$

- **Transpose** Let $\mathbf{A}$ and $\mathbf{B}$ be given matrices, then

$$(\mathbf{A} \otimes \mathbf{B})^\top = \mathbf{A}^\top \otimes \mathbf{B}^\top \tag{B.4}$$

- **Inverse** Let $\mathbf{A}$ and $\mathbf{B}$ be given invertible matrices, then

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1} \tag{B.5}$$

- **Matrix Equations** Let $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ be given matrices and $\mathbf{X}$ be an unknown matrix. Consider the equation $\mathbf{AXB} = \mathbf{C}$, this can be re-written as

$$(\mathbf{B}^\top \otimes \mathbf{A})\mathrm{vec}(\mathbf{X}) = \mathrm{vec}(\mathbf{AXB}) = \mathrm{vec}(\mathbf{C}) \tag{B.6}$$

Appendix C

Predicted Surfaces for the Basis Function Simulation Study in Chapter 5

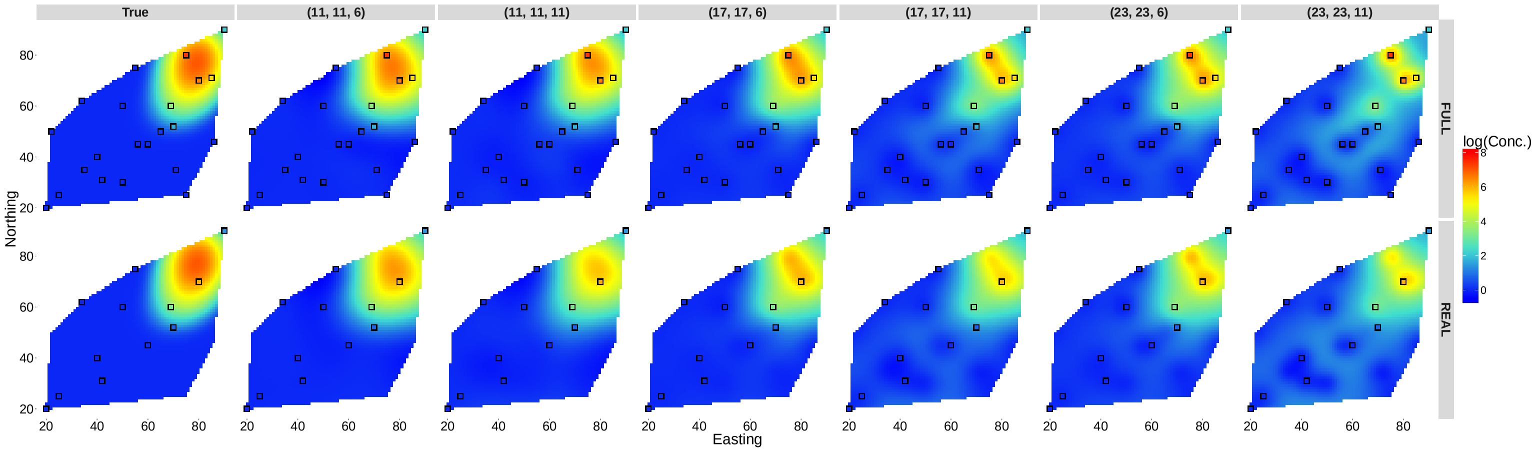

Figure C.1: Predicted surfaces for spatio-temporal spline models with one smoothing parameter and an increasing number of basis functions for the 10th time point in PDE2

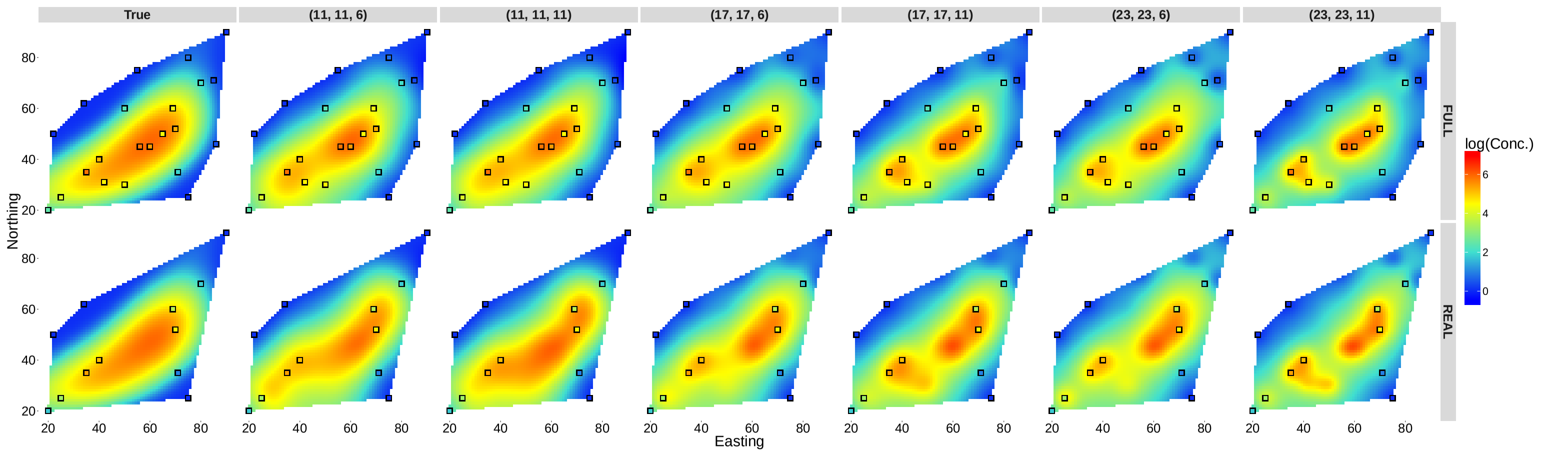

FIGURE C.2: Predicted surfaces for spatio-temporal spline models with one smoothing parameter and an increasing number of basis functions for the 32nd time point in PDE2

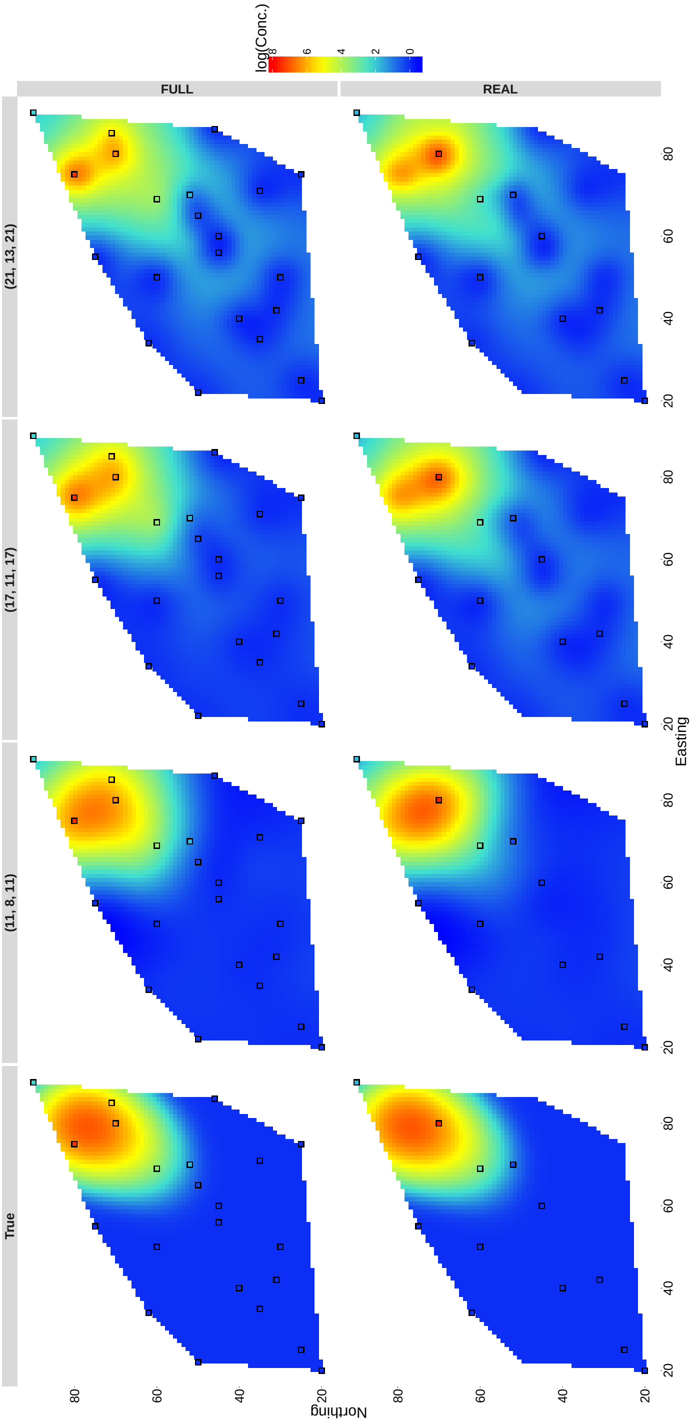

FIGURE C.3: Predicted surfaces for spatio-temporal spline models with two smoothing parameters and an increasing number of basis functions for the 10th time point in PDE2

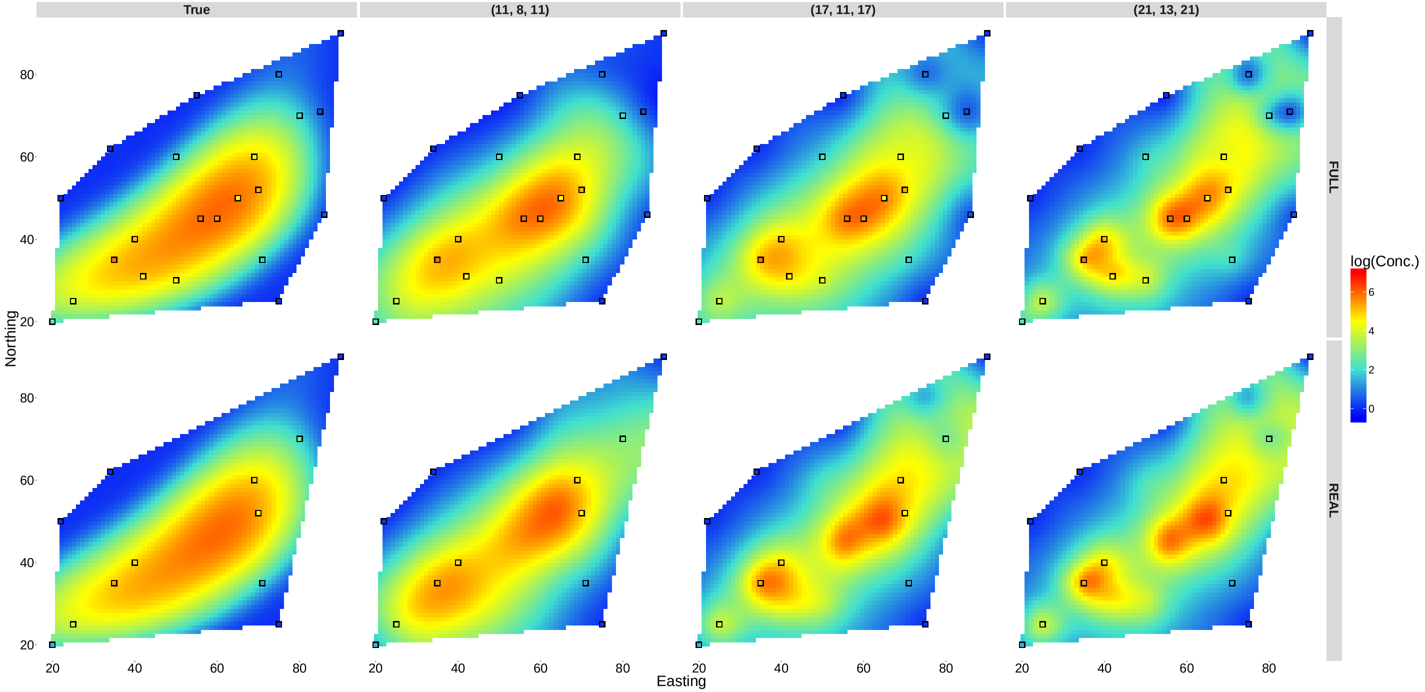

Figure C.4: Predicted surfaces for spatio-temporal spline models with two smoothing parameters and an increasing number of basis functions for the 32nd time point in PDE2

Appendix D

Analysis of a Second Design using the Objective Functions from Chapter 6

D.1 Objective Functions in Action

The analysis carried out in section 6.5, on Design 1, was repeated for Design 2, shown in Figure 6.1. Again, to compare the designs chosen by each model and (spatial p-splines, spatial Kriging and spatio-temporal p-splines) objective function combination, three proportions (25%, 50% and 75%) of the total number of wells were optimised for the next sampling event.

D.1.1 Designs for Spatial Models

For the spatial p-splines model degree three basis functions were used with a first order penalty. For Kriging, a Matérn covariance function was used with $\kappa = 2$.

The characteristics of the optimal designs seen for Design 1 are similar for Design 2. Figure 6.2 shows the optimal designs for each objective function for the spatial p-splines models. Here, as the proportion of wells being optimised increases the designs chosen by each objective function become more similar, with each design appearing to try and fill the study region.

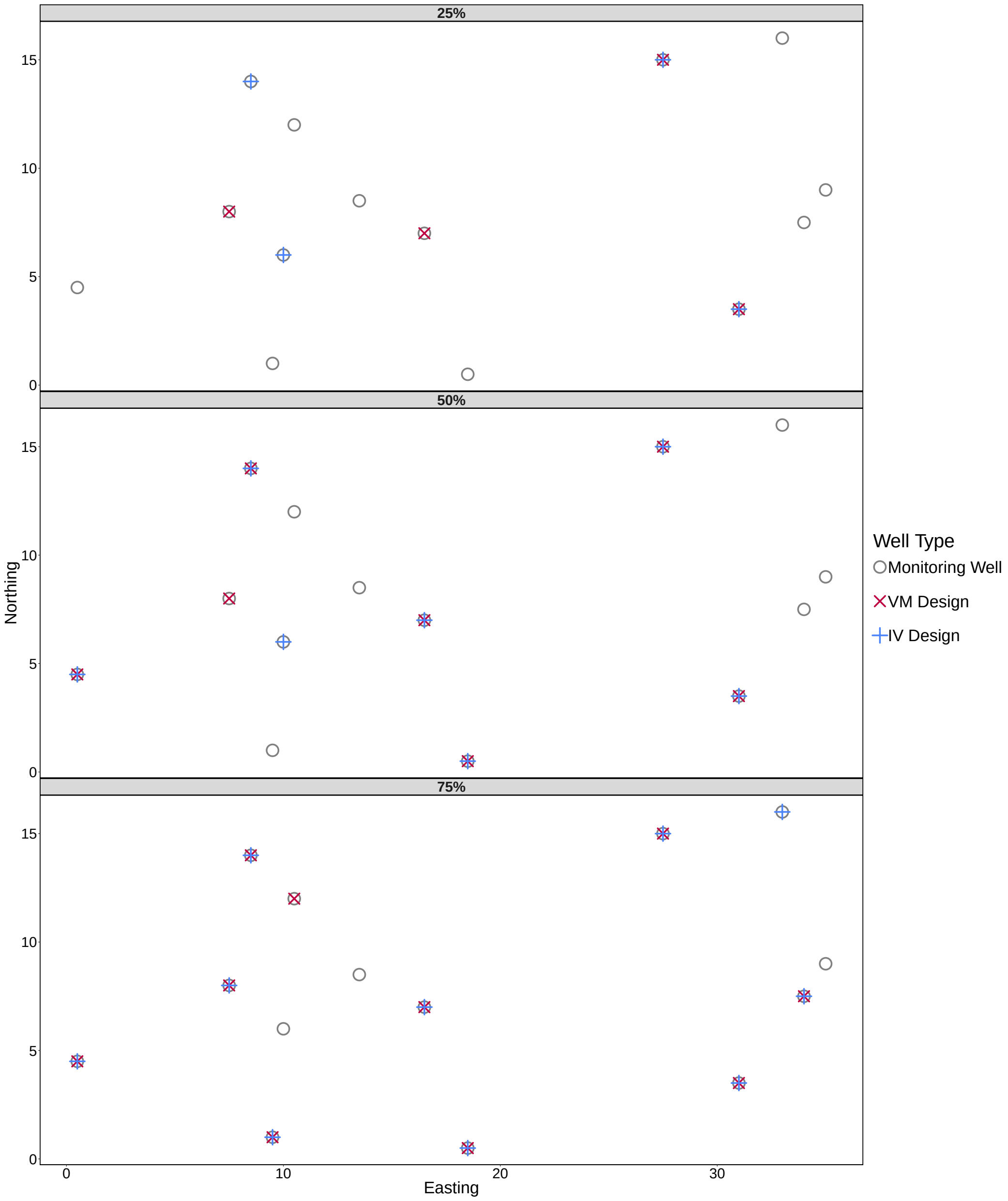

FIGURE D.1: Optimal designs for the integrated variance (IV) and variance of the mass (VM) objective functions using a spatial p-spline model when optimising for three proportions of the total number of wells (25%, 50% and 75%) using Design 2.

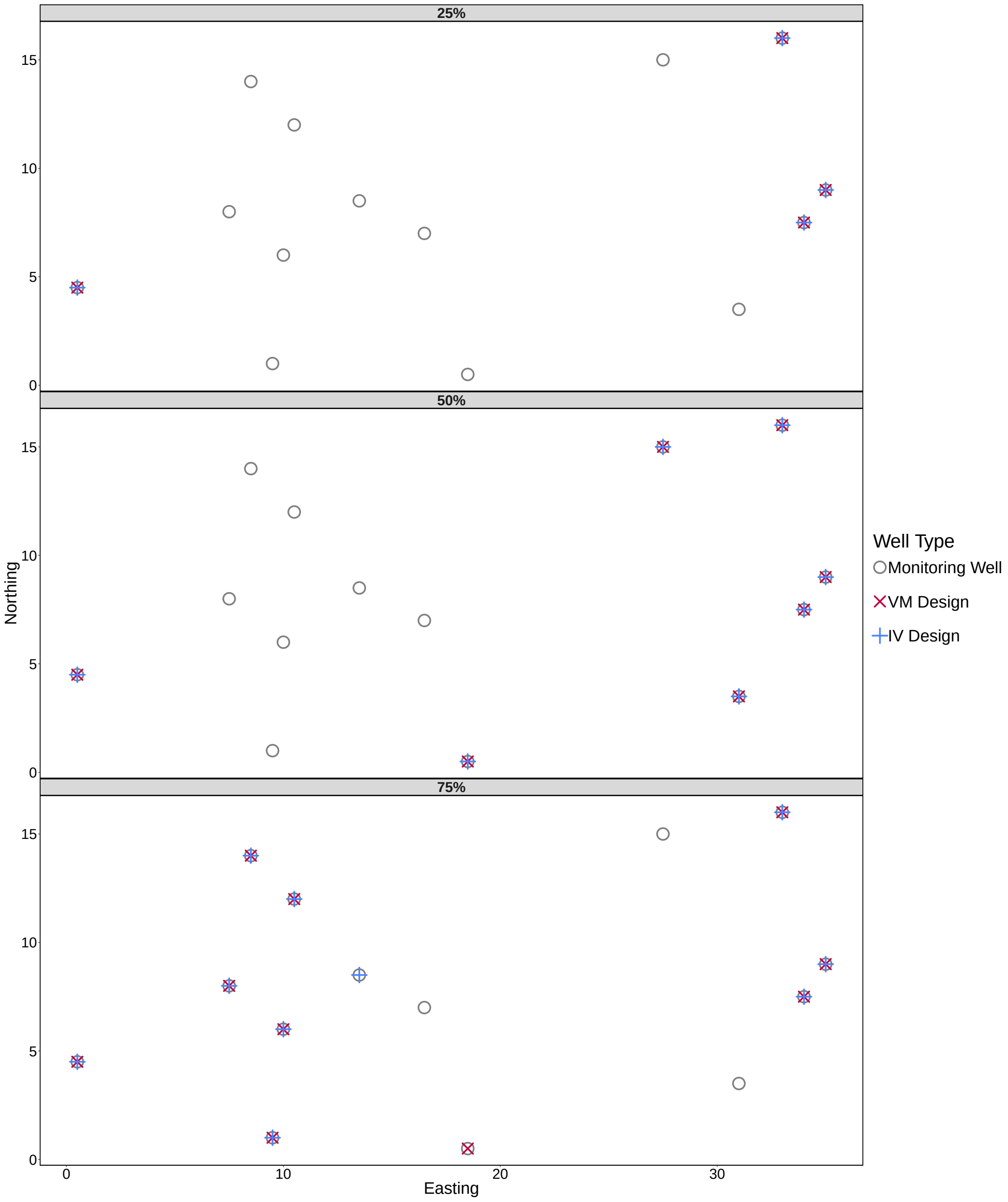

Figure D.2: Optimal designs for the integrated variance (IV) and variance of the mass (VM) objective functions using a spatial Kriging model when optimising for three proportions of the total number of wells (25%, 50% and 75%) using Design 2.

The designs chosen by the Kriging model are vary different to those chosen by the spatial p-splines model. For all three optimisations the Kriging model selects wells around the perimeter of the study as the optimal design. The designs for 25% and 50% of the wells are the same for both objective functions, with the designs for 75% of the wells only differing by one well. Wells located on the left side of the network are again preferred by both designs.

Tables D.1 and D.2 cross-tabulate the objective function values for each optimal design for the spatial p-splines and Kriging models respectively. Bold values indicate the design with the minimum of each objective function i.e. across each row in each subsection.

Table D.1: Cross-tabulation of the objective function values for each objective functions optimal design when optimising for 25%, 50% and 75% of the total number of wells using a spatial p-splines model using Design 2. Bold entries indicate the design with the minimum of each objective function.

		Optimal Design					
		25%		50%		75%	
		VM	IV	VM	IV	VM	IV
Value	VM	**898520**	991296	**446519**	450999	**329799**	340726
	IV	12784.1	**12661.2**	9334.4	**9297.1**	7808	**7758.1**

Table D.2: Cross-tabulation of the value of the objective functions for each objective functions optimal design when optimising for 25%, 50% and 75% of the total number of wells using a spatial Kriging model using Design 2. Bold entries indicate the design with the minimum of each objective function.

		Optimal Design					
		25%		50%		75%	
		VM	IV	VM	IV	VM	IV
Value	VM	160646	160646	353796	353796	**548173**	555038
	IV	1751.7	1751.7	3464.8	3464.8	5178.5	**5141.0**

Design 2 exhibits similar results to those of Design 1. Again, the design optimised using an objective function has the minimum value of the corresponding objective function across both designs when the designs differed between objective functions. However, for this design the percentage reductions in objective function values between designs is not as large as it was for Design 1. For spatial p-splines, the largest reduction is seen when optimising for 25% of the wells with the VM function being $\sim$ 9% lower for the design optimised using this function. As was the case for Design 1, the difference in the IV objective function between designs is much smaller, with the largest difference being 1%.

D.1.2 Designs for a Spatio-temporal Model

In a similar manner to the spatial models, a spatio-temporal p-splines model was used to optimise sampling designs using the two objective functions on Design 2. The two smoothing parameter spatio-temporal p-splines model was used with 15 basis functions for the easting and temporal components and the northing component had its number of basis functions scaled by the dimensions of the study region. As before, a first order difference penalty was used for both the spatial and temporal components.

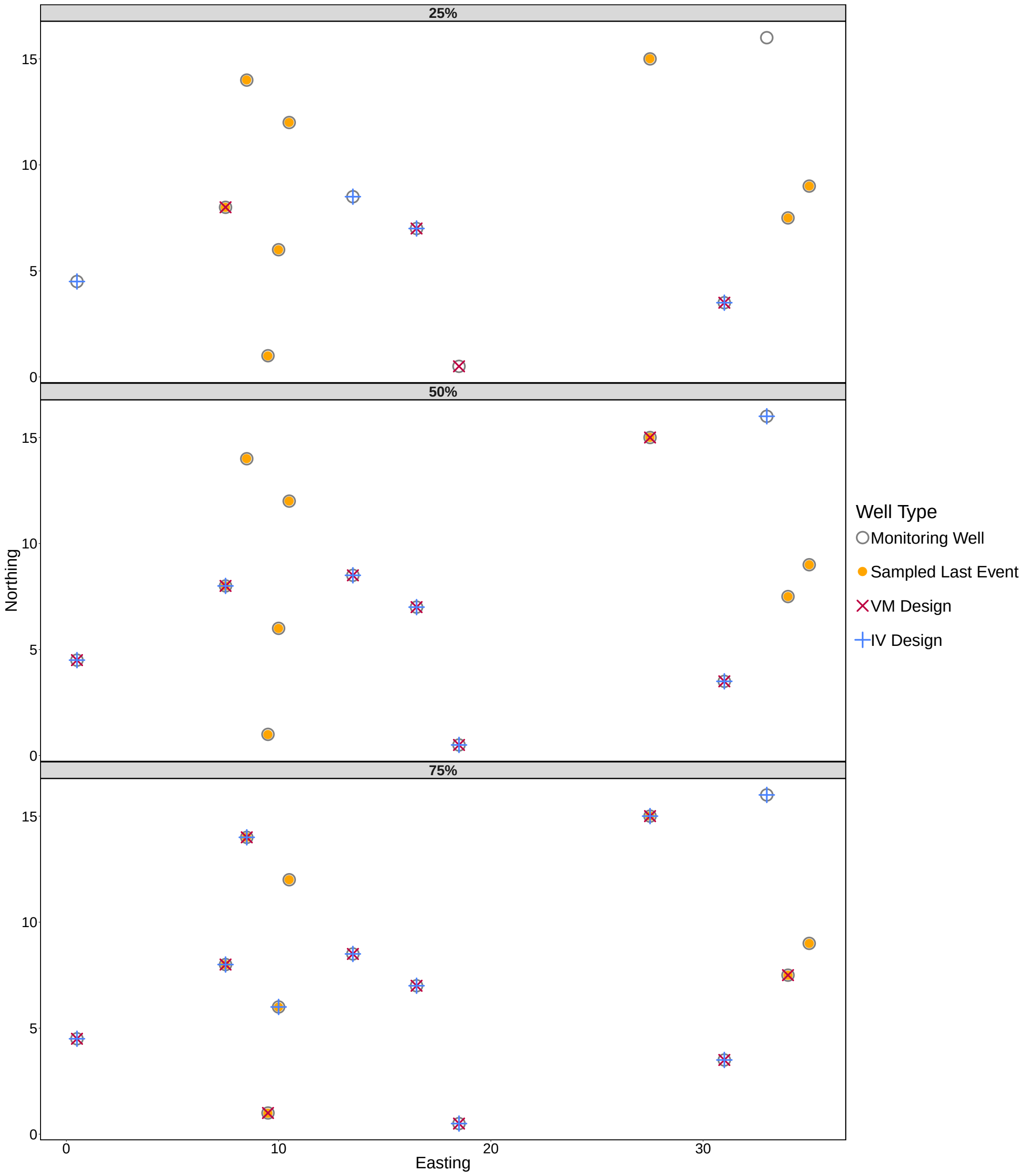

Figure D.3: Optimal designs for the integrated variance (IV) and variance of the mass (VM) objective functions using a spatio-temporal p-spline model when optimising for three proportions of the total number of wells (25%, 50% and 75%) using Design 2.

The designs for each objective function are different when optimised using a spatio-temporal model. The IV objective function chooses sampling designs such that as many of the 14 wells are sampled across the current and previous sampling event as possible. The designs chosen by both objective functions are similar in the sense that they both provide good spatial coverage of the region.

Table D.3: Cross-tabulation of the value of the objective functions for each objective functions optimal design when optimising for 25%, 50% and 75% of the wells using a spatio-temporal splines model for Design 2. Bold entries indicate the design with the minimum of each objective function.

		Optimal Design					
		25%		50%		75%	
		VM	IV	VM	IV	VM	IV
Value	VM	**154026**	154214	**152602**	152866	**151827**	151936
	IV	3533.2	**3526.7**	3501.9	**3499.6**	3486.1	**34816**

In a similar manner to the spatial models, the objective functions were cross-tabulated for each design for the spatio-temporal p-splines model. Again, the value of the objective function was lowest for the design optimised using the corresponding objective function. As was the case with Design 1, the increase in the objective function by using the opposite design is not as significant as it was for the spatial models.

D.1.3 The Effect of Previous Sampling Frequency

The effects of previous sampling frequency on the next sampling design were investigated for Design 2 in the same way they were for Design 1. For Design 2 only one data set was simulated with 7 randomly chosen wells in the "low frequency well" set. Figure D.4 shows the resulting designs using the two objective functions.

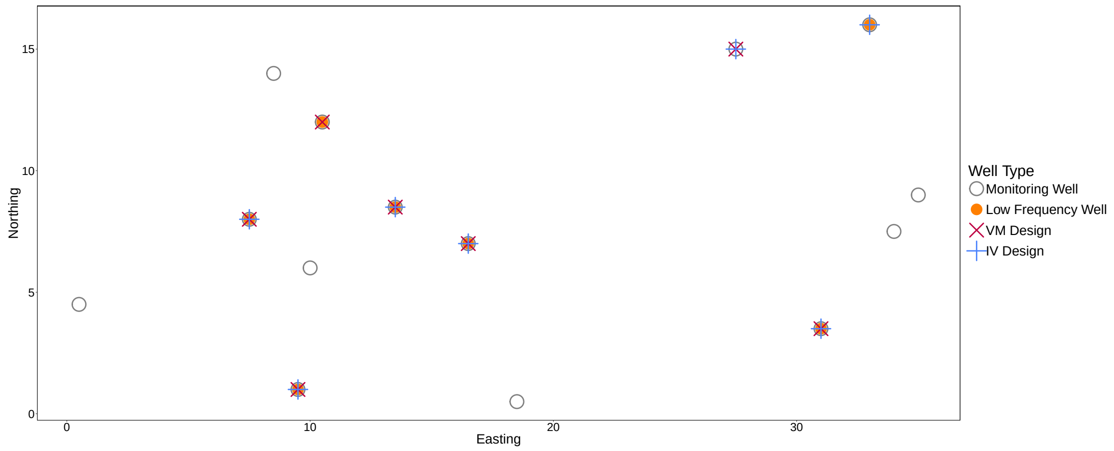

Figure D.4: Optimal design of seven wells using the low frequency dataset simulated on Design 2.

For Design 2, 6 of the 7 wells in the "low frequency well" set were chosen for the next design, with the 5 of the wells being the same for each design. Instead of choosing the other well from the "low frequency well" set, both designs select the same well from the "frequently sampled well" set in their next sampling design.

D.2 Using the Objective Functions to Change the Well Network

D.2.1 Adding a New Well to the Network

The pixel simulations performed on Design 1 in section 6.6.1 were then repeated for Design 2. Figures D.5, D.6, D.7 and D.6 show the resulting changes in objective function values when each pixel is proposed as a new location along with the optimal 10 wells.

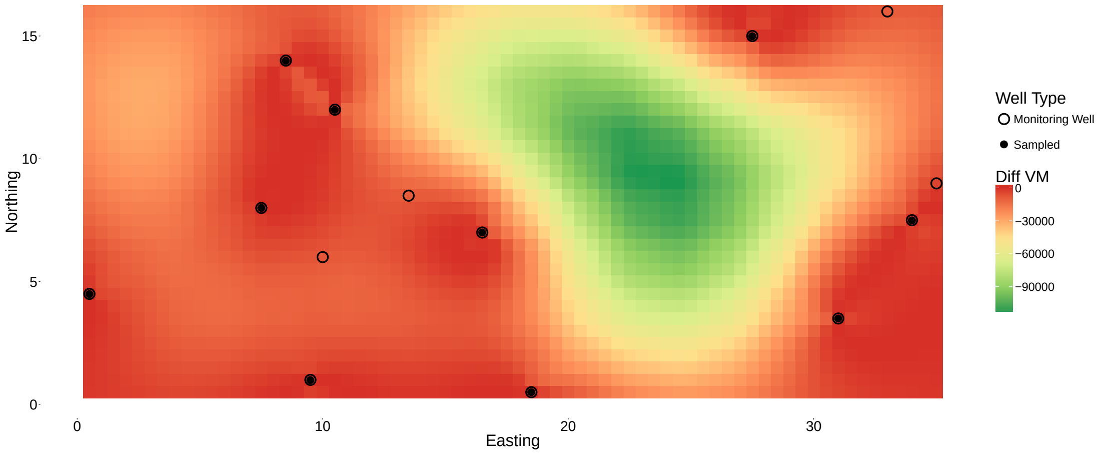

FIGURE D.5: Change in the VM objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatial p-splines model for Design 2.

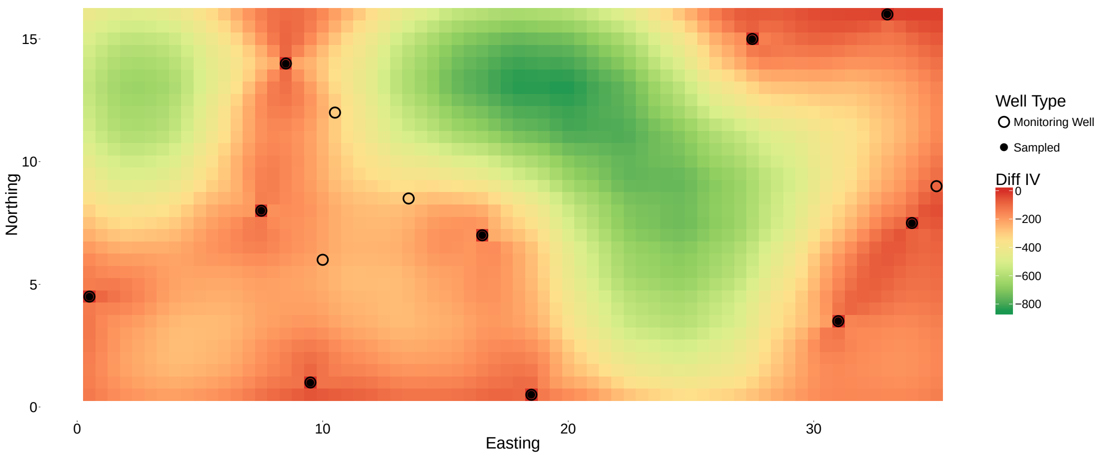

FIGURE D.6: Change in the IV objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatial p-splines model for Design 2.

Using spatial p-splines (Figures D.5, and D.6), the objective functions select slightly different locations as the optimal point for a new well, with the VM objective function suggesting the position would be to the right of the centre of the region, whereas the IV objective function suggests the best position is in the centre at the top. Both objective functions do however indicate that the largest reduction in objective function value would occur by positioning a well in a region roughly where the distance between neighbouring wells is greatest. The IV function also highlights the top left corner as a potential position for a new well, however the improvement in IV objective function value here is not quite as large.

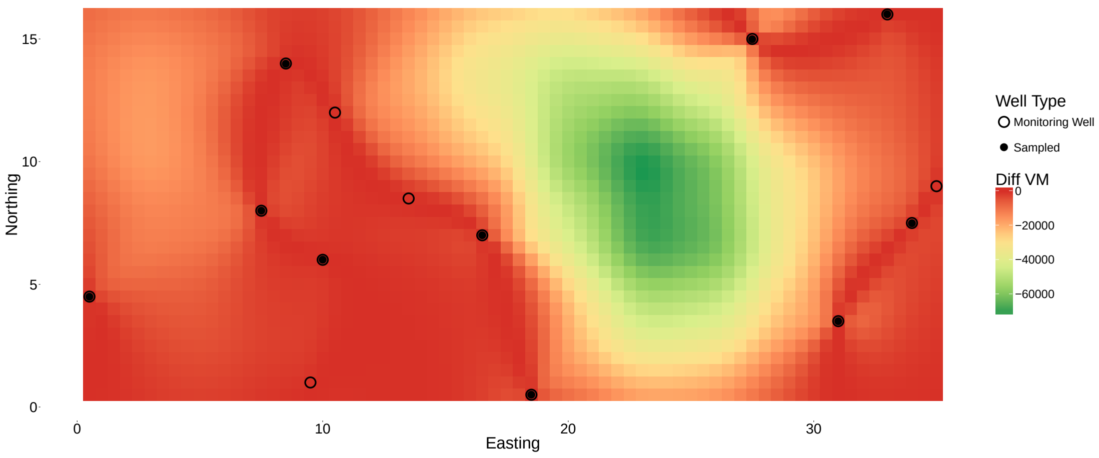

FIGURE D.7: Change in the VM objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatio-temporal p-splines model for Design 2.

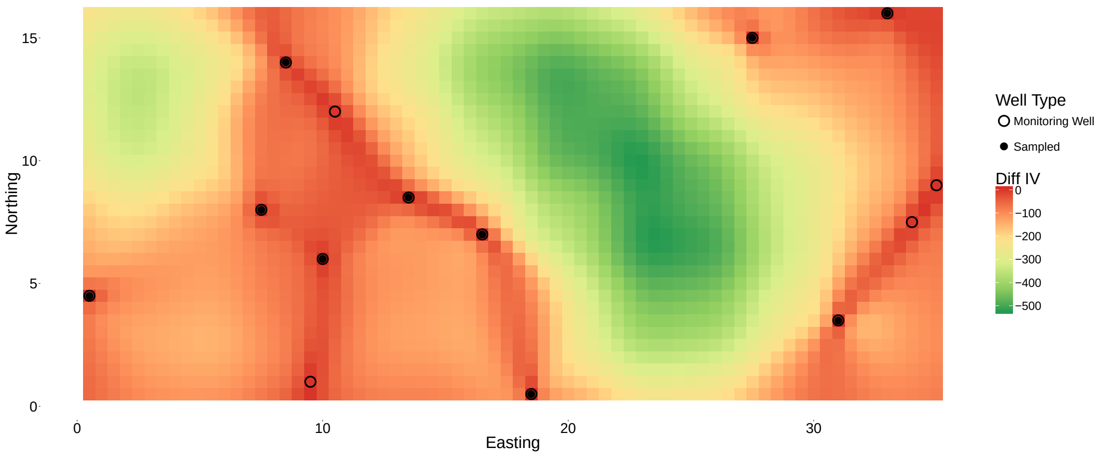

FIGURE D.8: Change in the IV objective function if each pixel were to be a new well location along with the current optimal 10 locations using a spatio-temporal p-splines model for Design 2.

The resulting objective function surfaces for the spatio-temporal p-splines model show similar trends to those of the spatial model. Both objective functions again highlight the region to the right of the centre as the most suitable location for a new wells. The IV objective function again suggests a potential location in the top left corner where the reduction in IV objective function is just marginally smaller than the most suitable locations.

D.2.2 Removing wells from the network

The method of identifying which well, when removed, would cause the smallest increase in prediction variance, discussed in Section 6.6.2, was used on Design 2. Figures D.9 and D.10 show heat maps of the change in prediction variance as each well is removed for spatial and spatio-temporal p-spline models respectively.

For the spatial p-splines model the change in prediction variance, as each well is removed, increases as the distance to the removed wells nearest neighbour increases. Wells located near the corners of the study region give the greatest increase in prediction variance when removed.

As was the case with Design 1, the wells which cause the smallest increase in prediction variance are very different for the spatio-temporal model compared with the spatial model. The wells which cause the largest increase are those which have not been sampled in the most recent sampling event, with those which are located around the perimeter and have been sampled in the last event giving the smallest increase.

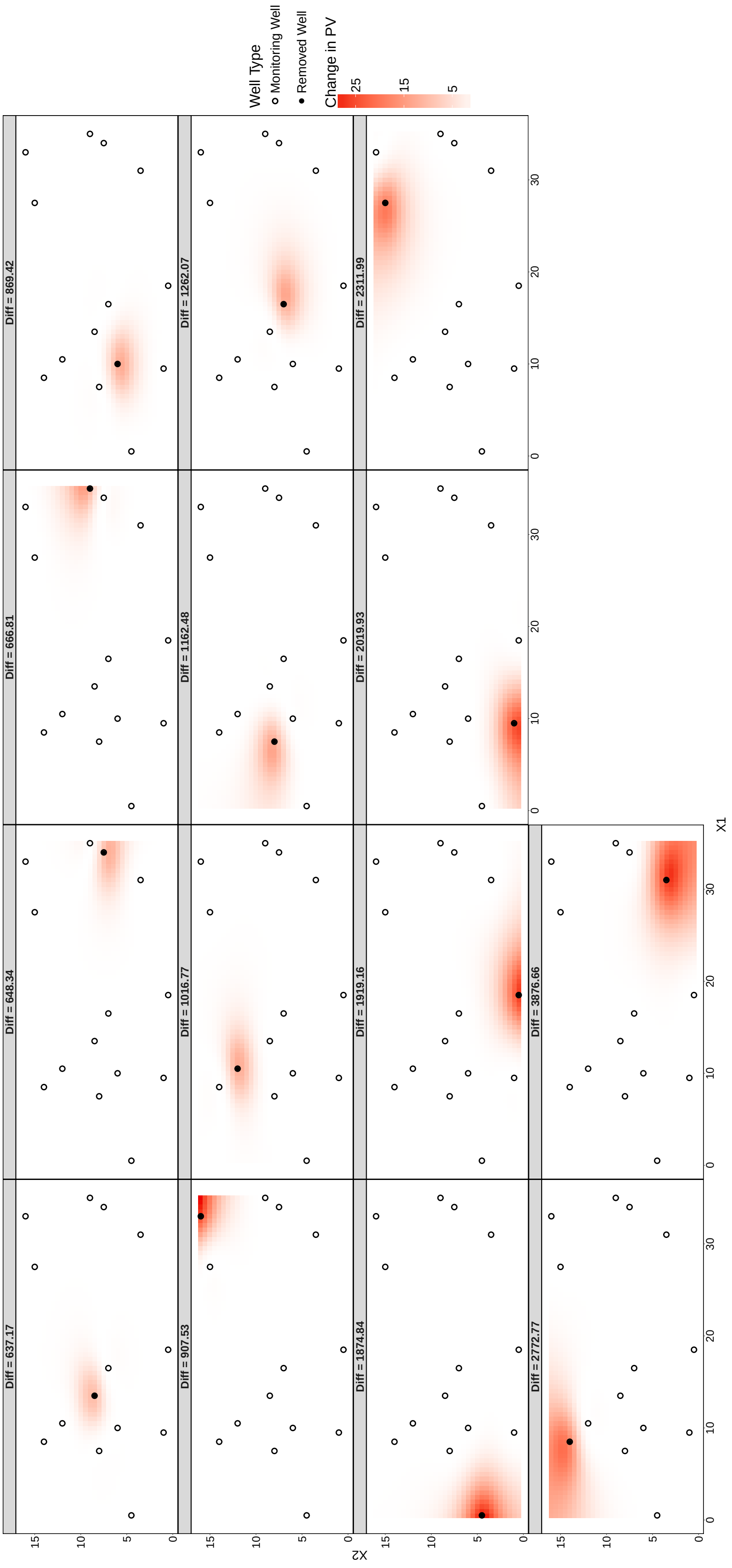

FIGURE D.9: Difference in predicted variance if each filled monitoring well is removed at the next time point compared with the prediction variance if all wells were sampled for Design 2 and a spatial p-splines model. The facets are ordered in increasing difference in prediction variance.

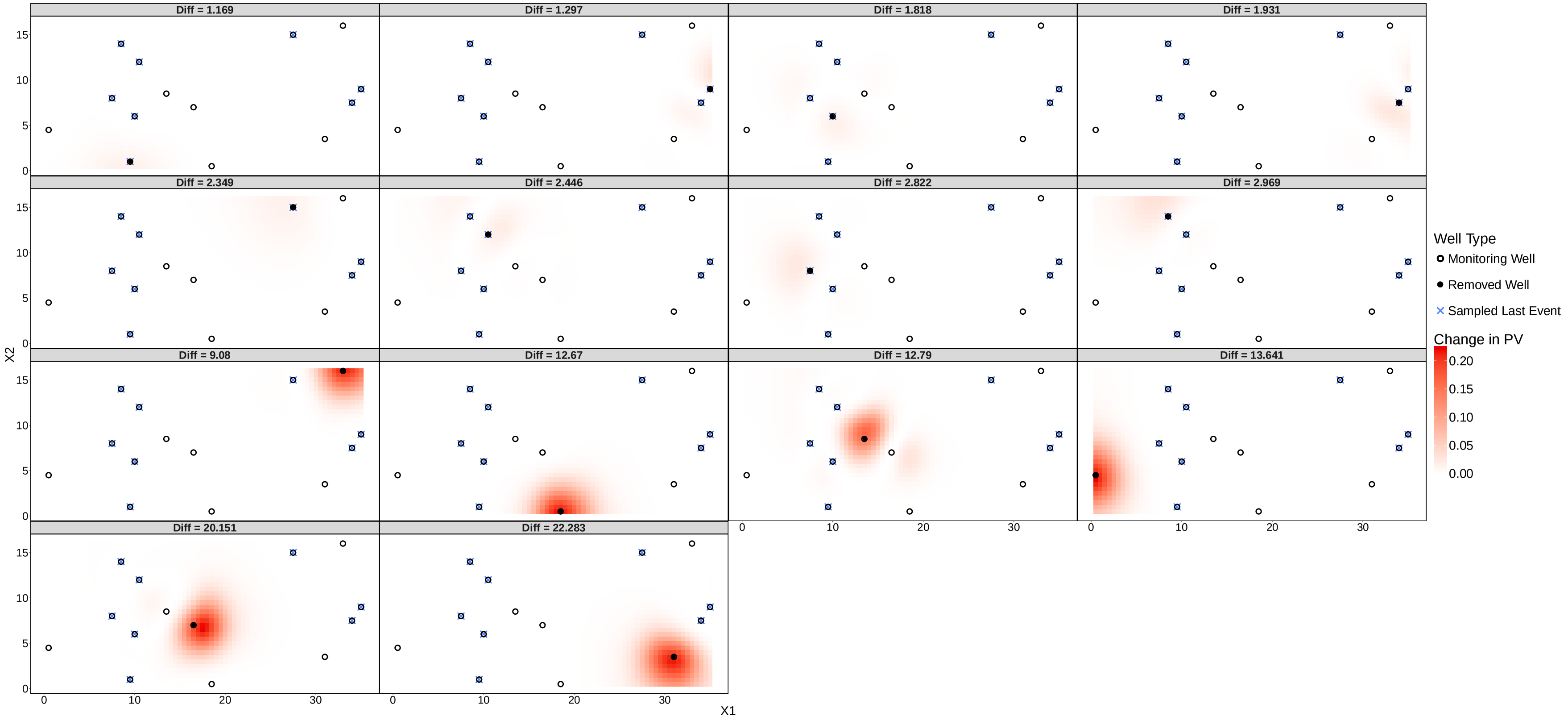

FIGURE D.10: Difference in predicted variance if each filled monitoring well is removed at the next time point compared with the prediction variance if all wells were sampled for Design 2 and a spatio-temporal p-splines model. The facets are ordered in increasing difference in prediction variance.

Bibliography

Abida, R., Bocquet, M., Vercauteren, N. and Isnard, O. [2008], 'Design of a monitoring network over france in case of a radiological accidental release', *Atmospheric Environment* **42**(21), 5205 – 5219.

Air Quality in Scotland [2018], 'Data'.
URL: *http://www.scottishairquality.co.uk/data/*

Akaike, H. [1973], *Information theory and an extension of the maximum likelihood principle*, Akadmiai Kiad, pp. 267 – 281.

Andricevic, R. [1990], 'Cost-effective network design for groundwater flow monitoring', *Stochastic Hydrology and Hydraulics* **4**(1), 27–41.

Angulo, M. and Tang, W. H. [1999], 'Optimal ground-water detection monitoring system design under uncertainty', *Journal of geotechnical and geoenvironmental engineering* **125**(6), 510–517.

Appelo, C. A. J. and Postma, D. [2004], *Geochemistry, groundwater and pollution*, CRC press.

Arnell, N. W. [1999], 'Climate change and global water resources', *Global environmental change* **9**, S31–S49.

Bear, J. [1979], *Hydraulics of groundwater*, Courier Corporation.

Bierkens, M. F. P. [2006], 'Designing a monitoring network for detecting groundwater pollution with stochastic simulation and a cost model', *Stochastic Environmental Research and Risk Assessment* **20**(5), 335–351.

Bishop, C. M. [2006], *Pattern Recognition and Machine Learning*, Springer.

Bohorquez, M., Giraldo, R. and Mateu, J. [2016], 'Optimal dynamic spatial sampling', *Environmetrics* **27**(5), 293–305.

Bowman, A. and Azzalini, A. [1997], *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*, Oxford Statistical Science Series, OUP Oxford.

Bowman, A. W., Giannitrapani, M. and Scott, E. M. [2009], 'Spatio-temporal smoothing and sulphur dioxide trends over europe', *Journal of the Royal Statistical Society Series C-Applied Statistics* **58**, 737–752.

Brus, D. J. and Heuvelink, G. B. M. [2007], 'Optimization of sample patterns for universal kriging of environmental variables', *Geoderma* **138**(1), 86 – 95.

Cameron, K. and Hunter, P. [2000], 'Optimization of ltm networks using gts: Statistical approaches to spatial and temporal redundancy', *Air Force Center for Environmental Excellence, Brooks AFB, TX* .

Cameron, K. and Hunter, P. [2002], 'Using spatial models and kriging techniques to optimize long-term ground-water monitoring networks: a case study', *Environmetrics* **13**(5-6), 629–656.

Chadalavada, S.and Datta, B. [2008], 'Dynamic optimal monitoring network design for transient transport of pollutants in groundwater aquifers', *Water Resources Management* **22**(6), 651–670.

Cressie, N. and Wikle, C. [2011], *Statistics for Spatio-Temporal Data*, CourseSmart Series, Wiley.

Dhar, A. [2013], 'Geostatistics-based design of regional groundwater monitoring framework', *ISH Journal of Hydraulic Engineering* **19**(2), 80–87.

Diggle, P. and Lophaven, S. [2006], 'Bayesian geostatistical design', *Scandinavian Journal of Statistics* **33**(1), 53–64.

Diggle, P. and Ribeiro, P. [2007], *Model-based Geostatistics.*, Springer Series in Statistics, Springer.

Eilers, P. H. C. and Marx, B. D. [1996], 'Flexible smoothing with b-splines and penalties', *Statistical Science* **11**(2), 89 – 102.

Eldén, L. [1977], 'Algorithms for the regularization of ill-conditioned least squares problems', *BIT Numerical Mathematics* **17**(2), 134–145.

Elumalai, V., Brindha, K., Sithole, B. and Lakshmanan, E. [2017], 'Spatial interpolation methods and geostatistics for mapping groundwater contamination in a coastal area', *Environmental Science and Pollution Research* **24**(12), 11601–11617.

Environmental Protection Agency [2017], 'Ground water monitoring requirements for hazardous waste treatment, storage and disposal facilities'.
URL: *https://www.epa.gov/hwpermitting/ground-water-monitoring-requirements-hazardous-waste-treatment-storage-and-disposal*

Environmental Protection Agency [2018], 'Meteorological data'.
URL: *https://www.epa.gov/ceam/meteorological-data*

European Commission [2018], 'Groundwater as a resource'.
URL: *http://ec.europa.eu/environment/water/water-framework/groundwater/resource.htm*

European Union [1999], 'Council directive 1999/31/ec of 26 april 1999 on the landfill of waste.', *Journal of the European Union* .

European Union [2016], 'Directive 2006/118/ec of the european parliament and of the council of 12 december 2006 on the protection of groundwater against pollution and deterioration', *Journal of the European Union* **372**, 19 –31.

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', *Environmetrics* **26**(6), 431–441.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. [2013], *Regression: Models, Methods and Applications*, Springer Berlin Heidelberg.

Frasso, G. [2013], *Splines, differential equations, and optimal smoothing*, Università Degli Studi Di Napoli Federico II.

Frasso, G., Jaeger, J. and Lambert, P. [2016*a*], 'Inference in dynamic systems using b-splines and quasilinearized ode penalties', *Biometrical Journal* **58**(3), 691–714.

Frasso, G., Jaeger, J. and Lambert, P. [2016*b*], 'Parameter estimation and inference in dynamic systems described by linear partial differential equations', *AStA Advances in Statistical Analysis* **100**(3), 259–287.

Fretwell, B. A., Short, R. I. and Sutton, J. S. [2006], 'Guidance on the design and installation of groundwater quality monitoring points', *Environment Agency* .

Groundwater Foundation [2018], 'Groundwater contamination'.
URL: *http://www.groundwater.org/get-informed/groundwater/contamination.html*

Helle, K. B. and Pebesma, E. [2012], 'Stationary sampling designs based on plume simulations', *Spatio-Temporal Design: Advances in Efficient Data Acquisition* pp. 319–344.

Herrera, G. S. and Pinder, G. F. [2005], 'Space-time optimization of groundwater quality sampling networks', *Water Resources Research* **41**(12).

Heuvelink, G. B. M., Griffith, D. A., Hengl, T. and Melles, S. J. [2012], 'Sampling design optimization for space-time kriging', *John Wiley, Oxford, doi* **10**, 207–230.

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

Holly, S., Pesaran, M. H. and Yamagata, T. [2010], 'A spatio-temporal model of house prices in the usa', *Journal of Econometrics* **158**(1), 160 – 173.

Hornberger, G. M., Wiberg, P. L., Raffensperger, J. P. and D'Odorico, P. [2014], *Elements of physical hydrology*, JHU Press.

Kiefer, J. and Wolfowitz, J. [1959], 'Optimum designs in regression problems', *The Annals of Mathematical Statistics* pp. 271–294.

Kotz, S. and Nadarajah, S. [2004], *Multivariate t-distributions and their applications*, Cambridge University Press.

Krige, D. G. [1951], 'A statistical approach to some basic mine valuation problems on the witwatersrand', *Journal of the Chemical, Metallurgical and Mining Society of South Africa* **52**, 119139.

Le, N. D. and Zidek, J. V. [2006], *Statistical Analysis of Environmental Space-Time Processes*, Springer Series in Statistics, Springer.

Lee, D. J. and Durban, M. [2011], 'P-spline anova-type interaction models for spatio-temporal smoothing', *Statistical Modelling* **11**(1), 49–69.

Lee, D., Mukhopadhyay, S., Rushworth, A. and Sahu, S. K. [2017], 'A rigorous statistical framework for spatio-temporal pollution prediction and estimation of its long-term impact on health', *Biostatistics* **18**(2), 370–385.

Li, J. and Heap, A. D. [2014], 'Spatial interpolation methods applied in the environmental sciences: A review', *Environmental Modelling and Software* **53**, 173 – 189.

Loaiciga, H. A. [1989], 'An optimization approach for groundwater quality monitoring network design', *Water Resources Research* **25**(8), 1771–1782.

Loaiciga, H. A., Charbeneau, R. J., Everett, L. G., Fogg, G. E., Hobbs, B. F. and Rouhani, S. [1992], 'Review of groundwater quality monitoring network design', *Journal of Hydraulic engineering* .

Mackay, D. M. and Cherry, J. A. [1989], 'Groundwater contamination: Pump-and-treat remediation', *Environmental Science & Technology* **23**(6), 630–636.

Maher, W. A., Cullen, P. W. and Norris, R. H. [1994], 'Framework for designing sampling programs', *Environmental Monitoring and Assessment* **30**(2), 139–162.

Malchow, H., Petrovskii, S. V. and Venturino, E. [2007], *Spatiotemporal patterns in ecology and epidemiology: theory, models, and simulation*, Chapman and Hall/CRC.

Mateu, J. and Muller, W. G. [2013], *Spatiotemporal Design: Advances in efficient data aquisition*, Wiley.

McPhee, J. and Yeh, W. W.-G. [2005], *Optimal Experimental Design for Parameter Estimation and Contaminant Plume Characterization in Groundwater Modelling*, John Wiley & Sons, chapter 9, pp. 219–245.

Meyer, P. D., Valocchi, A. J. and Eheart, J. W. [1994], 'Monitoring network design to provide initial detection of groundwater contamination', *Water Resources Research* **30**(9), 2647–2659.

Miller, C., Magdalina, A., Willows, R., Bowman, A., Scott, E., Lee, D., Burgess, C., Pope, L., Pannullo, F. and Haggarty, R. [2014], 'Spatiotemporal statistical modelling of long-term change in river nutrient concentrations in england & wales', *Science of The Total Environment* **466-467**, "914 – 923.

Molinari, D. [2014], *Spatiotemporal Modelling of Groundwater Contaminants*, University of Glasgow.

Montas, H. J., Mohtar, R. H., Hassan, A. E. and AlKhal, F. A. [2000], 'Heuristic space-time design of monitoring wells for contaminant plume characterization in stochastic flow fields', *Journal of Contaminant Hydrology* **43**(34), 271 – 301.

National Health Service [2018], 'Organisations'.
URL: *https://www.scot.nhs.uk/organisations/*

Ng, J. C., Wang, J. and Shraim, A. [2003], 'A global health problem caused by arsenic from natural sources', *Chemosphere* **52**(9), 1353 – 1359. Environmental and Public Health Management.

Nowak, W. [2010], 'Measures of parameter uncertainty in geostatistical estimation and geostatistical optimal design', *Mathematical Geosciences* **42**(2), 199–221.

Nowak, W., De Barros, F. and Rubin, Y. [2010], 'Bayesian geostatistical design: Task-driven optimal site investigation when the geostatistical model is uncertain', *Water Resources Research* **46**(3).

Nunes, L. M., Cunha, M. C. and Ribeiro, L. [2004], 'Groundwater monitoring network optimization with redundancy reduction', *Journal of Water Resources Planning and Management* **130**(1), 33–43.

Nunes, L. M., Cunha, M. and Ribeiro, L. [2013], 'Coverage methods for early groundwater contamination detection', *Bulletin of environmental contamination and toxicology* **90**(5), 531–536.

Nunes, L., Paralta, E., Cunha, M. and Ribeiro, L. [2004], 'Groundwater nitrate monitoring network optimization with missing data', *Water Resources Research* **40**(2).

O'Donnell, D., Rushworth, A., Bowman, A. W., Scott, M. E. and Hallard, M. [2013], 'Flexible regression models over river networks', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **63**(1), 47–63.

Reed, P. M., Ellsworth, T. R. and Minsker, B. S. [2004], 'Spatial interpolation methods for nonstationary plume data', *Groundwater* **42**(2), 190–202.

Reed, P. M. and Minsker, B. S. [2004], 'Striking the balance: Long-term groundwater monitoring design for conflicting objectives', *Journal of Water Resources Planning and Management* **130**(2), 140–149.

Reed, P., Minsker, B. S. and Goldberg, D. E. [2001], 'A multiobjective approach to cost effective long-term groundwater monitoring using an elitist nondominated sorted genetic algorithm with historical data', *Journal of Hydroinformatics* **3**(2), 71–89.

Reed, P., Minsker, B. and Valocchi, A. J. [2000], 'Cost-effective long-term groundwater monitoring design using a genetic algorithm and global mass interpolation', *Water resources research* **36**(12), 3731–3741.

Romary, T., Malherbe, L. and Fouquet, C. [2014], 'Optimal spatial design for air quality measurement surveys', *Environmetrics* **25**(1), 16–28.

Rouhani, S. [1985], 'Variance reduction analysis', *Water Resources Research* **21**(6), 837–846.

Schwarz, G. [1978], 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464.

Scottish Environment Protection Agency [2018], 'Groundwater'.
URL: *https://www.sepa.org.uk/regulations/water/groundwater/*

Shaddick, G. and Zidek, J. V. [2015], *Spatio-temporal methods in environmental epidemiology*, CRC Press.

Singh, B. and Sekhon, G. S. [1979], 'Nitrate pollution of groundwater from farm use of nitrogen fertilizers a review', *Agriculture and Environment* **4**(3), 207 – 225.

Sugiura, N. [1978], 'Further analysts of the data by akaike' s information criterion and the finite corrections', *Communications in Statistics - Theory and Methods* **7**(1), 13–26.

Surfer® 16 from Golden Software, LLC [2018].
URL: *www.goldensoftware.com*

van Geer, F. C., Bierkens, M. F. and Broers, H. P. [2008], *Groundwater Monitoring Strategies*, John Wiley & Sons, Ltd.

Ventrucci, M., Bowman, A. W., Miller, C. and Gross, J. [2014], 'Quasi-periodic spatiotemporal models of brain activation in single-trial meg experiments', *Statistical Modelling* **14**(5), 417–437.

Wagner, B. J. [1995], 'Sampling design methods for groundwater modeling under uncertainty', *Water Resources Research* **31**(10), 2581–2591.

Waller, L. A., Carlin, B. P., Xia, H. and Gelfand, A. E. [1997], 'Hierarchical spatio-temporal mapping of disease rates', *Journal of the American Statistical Association* **92**(438), 607–617.

Whitaker, S. [1986], 'Flow in porous media i: A theoretical derivation of darcy's law', *Transport in Porous Media* **1**(1), 3–25.

Wood, S. N. [2000], 'Modelling and smoothing parameter estimation with multiple quadratic penalties', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(2), 413–428.

Wood, S. N. [2004], 'Stable and efficient multiple smoothing parameter estimation for generalized additive models', *Journal of the American Statistical Association* **99**(467), 673–686.

Wood, S. N. [2006], *Generalized Additive Models: An Introduction with R*, Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.

Wood, S. N. [2011], 'Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models', *Journal of the Royal Statistical Society Series B-Statistical Methodology* **73**, 3–36.

Wu, J., Zheng, C. and Chien, C. C. [2005], 'Cost-effective sampling network design for contaminant plume monitoring under general hydrogeological conditions', *Journal of Contaminant Hydrology* **77**(1), 41–65.

Yeh, M. S., Lin, Y. P. and Chang, L. C. [2006], 'Designing an optimal multivariate geostatistical groundwater quality monitoring network using factorial kriging and genetic algorithms', *Environmental Geology* **50**(1), 101–121.

Yeh, W. W. G. [2015], ‘Review: Optimization methods for groundwater modeling and management’, *Hydrogeology Journal* **23**(6), 1051–1065.

Zhang, Y. Q., Pinder, G. F. and Herrera, G. S. [2005], ‘Least cost design of groundwater quality monitoring networks’, *Water Resources Research* **41**(8).

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

Zhu, Z. and Stein, M. L. [2006], ‘Spatial sampling design for prediction with estimated parameters’, *Journal of agricultural, biological, and environmental statistics* **11**(1), 24–44.

Zimmerman, D. L. and Li, J. [2013], ‘Model-based frequentist design for univariate and multivariate geostatistics’, *Spatio-Temporal Design: Advances in Efficient Data Acquisition* pp. 37–53.