Flexible Regression for River Systems

by

Alastair Rushworth

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

June 2014

Declaration of Authorship

I, Alastair Rushworth declare that this thesis titled, 'Flexible Regression for River Systems' and the work presented in it are my own. I confirm that 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.

The work presented in Chapter 3 has been published in Biometrics with the title *"Distributed lag models for hydrological data"* (Rushworth *et al.* (2013)) and was presented at the 26th International Workshop on Statistical Modelling (2011) in Valencia with the same title.

The work presented in Chapter 4 has been published in the Journal of the Royal Statistical Society Series C: Applied Statistics with the title *"Flexible regression models over river networks"* (O'Donnell *et al.* (2013)).

A manuscript describing the work presented in Chapter 5 has been prepared for publication with the title *"Validation and comparison of geostatistical and spline models for spatial stream networks"*. The work was also presented at the The International Environmetrics Conference (2013) in Alaska with the title *"Assessing and comparing the performance of models for stream network data"*. The software described in Chapter 5 is available to download from `http://alastairrushworth.wordpress.com/r-packages/`.

“Everywhere is walking distance if you have the time.”

Steven Wright

Abstract

Maintaining river health is of vital importance to the human populations that depend on them for drinking water, and for the income generated from industry and leisure activities. The key to a clear understanding of the current state of the river environment lies in assimilating the various data that are available for a particular river catchment. As a result of the large expense involved in extensive data collection programmes, measurements are often only taken at a handful of monitoring locations, resulting in large portions of a river network remaining unmonitored and rendering it difficult to assess the health of the river as a whole. Interpreting observations associated with a particular response variable pivots on understanding many other variables whose underlying relationships are often highly complex and which may not be routinely measured. Cutting-edge statistical methods can play a crucial role in the interpretation of such data, particularly when faced with small sample sizes and the presence of latent processes. In particular, developing models for environmental data that relax the assumption of simple linear dependencies between response and covariate is a core theme of this thesis, which can enable powerful descriptions of such complex systems. This approach adopts and promotes modern flexible regression techniques based on penalised splines, which are motivated and summarised in Chapter 2; these permit regression relationships to assume a wide variety of non-linear shapes, without requiring the modeller to impose *a priori* structure.

This thesis aims to address two related, but distinct regression problems for data collected within a river catchment. Firstly, the relationship between rainfall data collected at a rain gauge and subsequent river flow rates collected at a point downstream is tackled in Chapter 3. In this application, it is of particular interest to understand the degree, duration and time-lag of the influence of a rainfall event on a measurable increase in river flow rates at a downstream location. This relationship is complex because it is governed by attributes of

the surrounding river environment that may not be readily available, such as soil composition, land use and ground strata. However, rainfall and flow data are frequently collected at a high temporal resolution, and Chapter 3 develops models that exploits this feature that are able to express complex lagged dependence structures between a sequence of flow rates and a rainfall time series. The chapter illustrates how the resulting model enables insight into the sensitivity of the river to additional rainfall, and provides a mechanism for obtaining predictions of future flow rates, without recourse to traditional computationally intensive deterministic modelling.

This thesis also tackles the problem of constructing appropriate models for the spatial structure of variables that are carried by water along the channels of the river network. This problem cannot be approached using traditional spatial modelling tools due to the presence of the different volumes of water that mix at confluence points, often causing sudden changes in the levels of the measured variable. Very little literature is available for this type of spatial problem, and none has been developed that is appropriate for the large data sets that are becoming increasingly common in many environmental settings. Chapters 4 and 5 develop new regression models that can incorporate spatial variation on a stream network that respects the presence of confluences, flow rates and direction, while including non-linear functional representations for the influence of covariates. These different model components are constructed using the same modern flexible regression framework as used in Chapter 3, and the computational benefits of adopting this approach are highlighted. Chapter 4 illustrates the utility of the new models by applying them to a large set of dissolved nitrate concentrations collected over a Scottish river network. The application reveals strong trends in both space and time, and evidence of a subtle interaction between temporal trend and the location in space; both conclusions would have been difficult to reach using other techniques.

Acknowledgements

I would like to thank my supervisors Adrian Bowman and Mark Brewer for their support and patience throughout all of the stages of my PhD research. Without their gentle guidance and advice, I would not have been able to develop and learn so much. I'd also like to thank the staff and students in both the School of Mathematics and Statistics at Glasgow University and at Biomathematics and Statistics Scotland, for providing such nurturing, diverse and absorbing environments in which to develop the skills of a researcher.

Thanks to the staff at the James Hutton Institute for providing river flow and rainfall data and for their support during the early stages of my work. I would also like to acknowledge the input made by the Scottish Environment Protection Agency and the British Atmospheric Data Centre for providing the River Tweed and rainfall data, respectively.

Big thanks are due to Dr. Jay ver Hoef and Dr. Erin Peterson for giving me the opportunity to spend two fantastic months working with them at the Commonwealth Scientific and Industrial Research Organisation in Brisbane, which broadened my understanding of river network models and software writing. I am particularly grateful to the Jim Gatheral Travel Scholarship for providing the financial support that made this collaboration possible. Thank you also to Dr. Mark Hallard and the members of the data team at the Scottish Environment Protection Agency in Stirling for the chance to work with them during a secondment from my PhD research, and to Prof. Marian Scott for providing the opportunity for carrying out this work.

Finally, to my friends and family, for support both explicit and implicit, for beers, coffees, chocolate, hearty meals, climbs, distracting conversations, walks and bike rides.

Contents

List of Figures

List of Tables

To Alexis

Chapter 1

Introduction

1.1 Monitoring of river networks

Rivers and the reservoirs they fill, are primary sources of drinking water for human populations worldwide, are widely used for leisure and recreational activities and are important resources for transport, industrial and economic output. Equally as important as the human benefits, rivers and river banks are habitats and food sources for many endemic species that are sensitive to their immediate physical environment and climatic change. For example, Vörösmarty *et al.* (2010) find that "*nearly 80% (4.8 billion) of the world's population (for 2000) lives in areas where either incident human water security or biodiversity threat exceed the 75th percentile*". Preserving and improving the health of rivers worldwide is therefore a vitally important issue, and has received much political attention in recent years. In 1991, the Nitrates Directive (European Parliament (1991)) was introduced, obligating European member states to monitor and report on aspects of water quality. In particular, the Nitrates Directive requires areas of land that drain into nitrate-polluted water bodies to be designated 'Nitrate vulnerable zones' (NVZ), and upon classification, measures must be implemented to improve water quality. In 2000, the European Water Framework Directive (WFD) (European Parliament (2000)) was introduced and has been part of UK law since 2003. The WFD consolidated previous policies into a single broad legislative framework intended to protect all water bodies including rivers, lakes and groundwaters by setting targets for minimum water health, and encouraging a coordinated approach to river basin management and monitoring.

Adherence to the WFD obligates government bodies, such as the Scottish Environment Protection Agency (SEPA) to undertake systematic monitoring and reporting on the state of their water bodies. As a result of subsequent national-scale water monitoring, a large amount of data is collected. However, due to the costs associated with obtaining samples, the data are often collected with a limited spatial coverage. Furthermore, the available data are often noisy, due to factors such as equipment measurement error and observations made at the limits of detection of the equipment. It is also common to find large gaps in the measurement record due to the establishment or termination of sampling regimes at a particular location. As a result of these and other problems, visual assessment of the data alone is not sufficient to obtain an understanding of the current state of an aquatic environment. Statistical modelling therefore plays a crucial role in the interpretation of these data, and provides a framework within which scientific hypotheses can be tested even when faced with multiple sources of uncertainty and incomplete data sets.

1.2 Data on river networks

Data collected on quantities occurring in the natural environment often exhibit non-linear relationships, caused by the complex latent physical systems from which they are generated. In the case of dissolved pollution carried by river flow, observed processes such as differing land usages, meteorological conditions and the presence of sewage outlets are known to induce strong spatial and temporal dependencies into the observed data. Much of the literature that deals with hydrological and environmental systems takes a deterministic approach to modelling, by describing the system through physical process models, for summaries in hydrology, see for example Beven (1985) and Shaw (2010). This approach is very powerful and seeks to build a model that describes a detailed mechanistic representation of the system of interest. However, these models are often highly computationally intensive to run, and it is not straightforward to accommodate uncertainty about model parameters or measurement error resulting from the data collection process.

This thesis explores the use of alternative statistical approaches to describing an environmental system that relies only on the observed data to determine the most appropriate model structure. An attractive and well established method of statistical modelling is linear regression analysis, where a set of explanatory variables are modelled as each having a linear contribution to a single dependent outcome variable. However, when the explanatory variables exhibit strong spatial and temporal dependence, standard linear regression is not an appropriate tool. It is therefore the aim of this thesis to extend current regression modelling techniques so that the incorporation of complex variation in space and time is permitted, making use of a flexible framework that avoids imposing the restrictions such as linearity on relationships between variables in the model that would poorly describe the true dependence in the data. Capturing spatial and temporal patterns is a computationally challenging problem in the field of spatial analysis, and so this thesis places a particular emphasis on choosing algorithms that result in efficient computation.

1.3 Thesis Outline

Although the development of appropriate flexible regression models is the central theme of this thesis, two related but distinct problems are tackled in particular. The first concerns the statistical representation of the relationship between river flow rates and rainfall, utilising data from an individual rain gauge and from a single high frequency stream flow monitoring site. The second problem that this thesis seeks to address is the challenge of modelling observations on variables such as water temperature, nitrate loads and dissolved oxygen concentrations that are attributes of the water that flows through the channels of a spatial stream network.

1.3.1 Chapter 2: Introducing flexible regression

Non-linearity between variables can result because highly complex deterministic processes may govern the systems underlying in the variable or because the behaviour of a variable is dynamic, for example with time, and that the form of the change has no *a priori* deterministic representation nor does it follow a simple parametric form. The latter form of non-linearity forms a key motivation of this thesis, and Chapter 2 summarises the problem that arises when standard linear regression is used and the variables involved exhibit these types of non-linear relationships. An alternative flexible modelling strategy that avoids these issues and is based on a particular penalised spline regression (P-splines) is presented together with an overview of how such models are fitted, with some discussion of techniques for obtaining optimal levels of flexibility. The chapter describes extensions to basic flexible regression models that are often useful in practice, such as modelling bivariate smooth interactions between two covariates. The chapter summarises key features from the wider semiparametric modelling literature including the different frameworks available for achieving parameter estimation. The relevance and importance of sparse matrix algorithms in model fitting is outlined, with emphasis on how these can be exploited during smoothness parameter selection.

1.3.2 Chapter 3: Flexible rainfall and flow modelling

Chapter 3 investigates the time-lagged relationship between river flow and rainfall, in the particular setting where only high resolution rainfall and flow time series are available. Drivers of flow generation are summarised alongside a review of existing modelling approaches to rainfall-flow modelling. Flexible time-varying coefficient models are proposed, that are based on P-splines that are capable of capturing complex time-lagged dependence between the time series. The proposed models develop existing distributed lag models that are more commonly utilised in short-term epidemiological studies, by allowing the flexible lag structure to vary in time. A more general modelling framework is described that allows the construction of varying-coefficient models in which the coefficients vary in multiple dimensions. Both of the models developed are applied to data collected on the River Dee in the North East of Scotland. Chapter 3 closes with some criticism of the adequacy of the proposed models in the context of the data considered and some potential improvements that could be made.

1.3.3 Chapters 4 and 5: Flexible models for river networks

Chapter 4 investigates space-time modelling of covariates that are observed at a set of monitoring sites on a river network. Standard spatial models are not appropriate because of the unique features that underpin variation on a flow-driven network, such as the flow direction and presence of confluences that connect different flow channels. Some recently developed models for such structures from the spatial modelling literature are reviewed and discussed. A new approach utilising P-splines to represent both the spatial propagation of network-varying covariate values and non-linear covariate effects is introduced. The models are extended to allow the inclusion of network-varying smooth effects. Chapter 4 finishes with a discussion of the current shortcomings of the network models developed, and some potential avenues for future research.

Chapter 5 builds on Chapter 4 by rigorously testing the performance of the network models and comparing them to others available in the current literature. The relative performance is measured and comparisons are made by implementing a large simulation study where data are generated from a wide variety of realistic river network structures. Predictive performance is particularly emphasised as this is a particularly common goal in environmental space-time analysis. Issues such as computation and user-friendly implementations of the models are also addressed.

1.3.4 Chapter 6: Main findings and future extensions

In Chapter 6, a summary of the main findings and contributions of the thesis is presented. In addition to this summary, a substantial extension to the work of Chapter 3 is proposed, that could make use of highly detailed rainfall RADAR data in order to improve model accuracy and reduce the bias that could arise as a result of the spatio-temporal structure in rainfall. This extension describes a distributed lag model in which flow levels are dependent on rainfall at both spatial and temporal lags. Chapter 6 also sketches a route for further developing the river network models of Chapter 4, in which variables measured on the network could be modelled continuously, rather than with the discrete segment-wise representation developed in Chapter 4. This model would make use of a more complex P-splines specification for capturing smooth structure on the bodies of a stream network.

Chapter 2

Flexible modelling with P-splines

In this chapter, the problem of representing non-linear dependence in a regression context is discussed, and how an attractive solution may be obtained using P-splines, a particular type of smoothing technique. Dealing with non-linearity in a regression context is central to appropriately handling environmental data, a point which was particularly emphasised in Section 1.2. A summary of the ideas underlying P-splines is provided in this chapter and is framed within the wider context of modelling with generalised additive models, semi-parametric models and nonparametric regression that are supported by a well developed literature (Hastie & Tibshirani (1990), Green *et al.* (1994), Ruppert *et al.* (2003), Wood (2006)). Simple motivating examples are provided to support the discussion, and a strong emphasis is placed on efficient computation, and the different techniques by which the smoothness control parameters can be estimated.

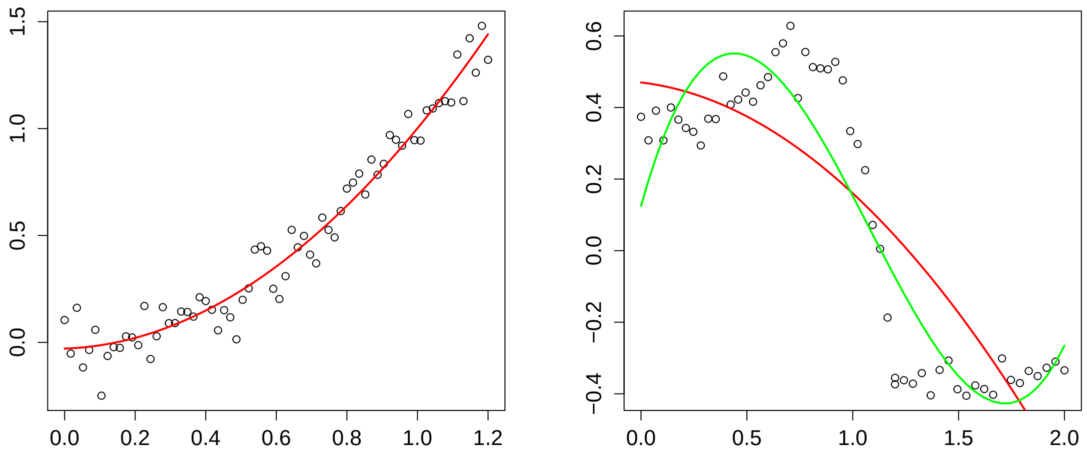


FIGURE 2.1: Two simulated examples of noisy, non-linear data relationships. The data shown in the left panel have been fitted with a quadratic polynomial (red line), while the data shown in the right panel have been fitted with quadratic (red line) and cubic (green line) polynomials.

2.1 Regression, smoothing and basis functions

2.1.1 Regression with polynomials

Standard regression models that fit straight-line relationships between response and explanatory variable, are not appropriate when the true relationship exhibits curvature, and will typically result in strongly correlated residuals. In many instances, the curvature can be accommodated by fitting polynomials, or other parametric functions of the explanatory variable: an example of this is illustrated by Figure 2.1. The left hand panel of Figure 2.1 shows the result of fitting a linear model with polynomial terms of the form

$$y_i = f(x_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \qquad \text{where} \quad \epsilon_i \sim N(0, \sigma^2)$$

where $1 \leq i \leq n$ for n data points. The fit appears to be reasonable, as the curvature present is relatively simple and easily accommodated by a quadratic function. In contrast, the data shown in the right hand panel of Figure 2.1 exhibit more complex curvature, and it is less clear that either the quadratic or the cubic functions fitted can offer a good description of the observed data structure. In practice, the fits shown in Figure 2.1 might be deemed reasonable if there were good reason to believe the true regression relationship to be polynomial *a priori*. In many instances, polynomial functions require the practitioner to make strong assumptions about the *global* behaviour of the fitted function that cannot be easily justified; for example, in the left panel of Figure 2.1 it is unclear that at lower values on the x-axis, the response should encounter a turning point and begin to increase as the fitted quadratic function would. In addition, the range of shapes permitted by polynomial functions are typically too restrictive to capture non-trivial curvature, which is a feature that is particularly visible in the data in the right panel of Figure 2.1.

In pursuit of a better alternative, it is helpful to consider the influence of assuming global parametric structures such as those imposed by polynomial regression. For the fitted quadratic functions shown in Figure 2.1, the terms $\{1, x_i, x_i^2\}$ are non-zero for all values of x (except at $x = 0$) which means that each of the parameter estimates $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ dictate the trajectory of the fitted function f at *all* observed and unobserved values of x_i. This could be a particularly undesirable property if the goal is to predict outside of the range of the observed data. Slightly more formally, $\{1, x_i, x_i^2\}$ is a *basis* for f that spans the space of all quadratic polynomials, and if the true function does not lie in this space, then the resulting fit to the data may be poor.

2.1.2 Approaches to flexible regression

Flexible regression methods attempt to avoid these undesirable properties by constructing estimates for f that capture the behaviour of the data *locally*, where f is typically only assumed to be continuous and differentiable to some degree. There are a number of techniques available to achieve this, and one of the most prominent is known as *kernel regression* (see Bowman & Azzalini (1997) for an overview) which estimates the function f by weighting the data points with a kernel function K. A common choice is Nadaraya-Watson kernel regression (Nadaraya (1964), Watson (1964)), which fits f using the estimator

$$E(y_i|x_i, h) = \hat{f}(x_i) = \frac{\sum_{j=1}^{n} K_h(x_i - x_j)y_j}{\sum_{j=1}^{n} K_h(x_i - x_j)}. \tag{2.1}$$

In Equation 2.1, h is a bandwidth parameter that controls the spread of the chosen kernel function K_h, and the subsequent smoothness of the fitted function $\hat{f}$. As seen in Equation 2.1, kernel regression relies on weighting all of the data points, and its implementation requires manipulation of $n \times n$ matrices. This feature may limit the use of kernel methods with large data sets, although it should be noted that in many situations the matrices involved are banded and sparse because the kernel function K_h is usually close to 0 outside of a narrow interval around each datum. An alternative approach that aims to achieve a computational cost that scales more favourably with n, starts instead by projecting the data on to a much smaller set of *locally defined basis functions*, and then seeks to weight these functions in such a way that their linear combination provides a good representation of the data. The basis functions can take a very diverse range of forms, and their local nature may be as a result of each function having a prominent maximum or minimum, or in the more direct sense that they have compact support. Some examples of basis functions in regular use are truncated power functions Ruppert *et al.* (2003), B-splines (De Boor (1978), Eilers & Marx (1996)), monotonic I-splines (Ramsay (1988)) and thin-plate splines (Duchon (1977), Wood (2006)). Regardless of the basis chosen, this is generally referred to as low-rank smoothing, in the sense that the number of parameters required to construct the estimated smooth function is typically far fewer than the number of data points.

The main differences between competing approaches for low-rank smoothing lies in the choice of the number and of the locations of the basis functions, and the subsequent mechanism for obtaining parameter estimates. For example, Ruppert *et al.* (2003) prefer truncated power functions as basis functions, with functions spaced uniformly on the quantile scale of the data, using a ridge penalty to control smoothness and a mixed model approach to parameter estimation. Although easy to construct, truncated power functions are in some instances prone to poor numerical condition, highlighted for example by Eilers & Marx (2010). The P-splines approach of Eilers & Marx (1996) in contrast, chooses a rich set of uniformly spaced B-spline basis functions whose linear combination represents $\hat{f}$, together with a roughness penalty on pairs or higher-order neighbourhoods of basis functions. In order to maintain a clear and simple presentation of the models developed in this thesis, the smoothing framework proposed by Eilers & Marx (1996) will be adopted and used throughout this thesis due to its conceptually intuitive approach to fitting smooth functions. It is acknowledged that different smoothing frameworks could have been chosen, which would have likely resulted in slight differences in the models subsequently developed. However, each of the different frameworks are intended to fit smooth functions with minimal input from the modeller, and so it is also likely that any differences between them would be small, and that the results would not be affected.

2.2 P-splines

Section 2.2.1 proceeds by defining B-spline basis functions, which are the functional building blocks of the P-splines approach. Following this, in Section 2.2.2 the basic principles of using a set of B-spline basis functions to construct a smooth function is outlined, and the use of a roughness penalty to control smoothness in the smooth function is also described.

2.2.1 B-splines

A spline is simply a function constructed from polynomial pieces joined together in a specific way, and a B-spline of degree q in particular has the following general properties that were summarised by Eilers & Marx (1996):

1) It consists of $q+1$ polynomial pieces, each of degree q

2) The polynomial pieces are joined at q inner knots

3) At the joining points, derivatives up to order $q-1$ are continuous

4) The B-spline is positive on a domain spanned by $q+2$ knots; elsewhere it is zero

5) At a given x, $q+1$ B-splines are nonzero

A comprehensive reference that describes the mathematical properties is given by De Boor (1978). B-splines are popular choices for smoothing and regression as they are easy to construct and have attractive numerical properties. Each B-spline is composed of $q+1$ polynomial pieces joined at q inner knots, therefore constructing a B-spline *basis* consisting of p individual basis functions requires the choice of the location of these knots. Any choice of knot placement is possible, for example placing them at q evenly-spaced percentiles of the data may result in an appropriate smooth when the response is observed rarely over large intervals of a covariate. However, it is more commonly assumed that the knots are placed at a set of equally spaced at locations $(v_1, \ldots, v_{p+q+1})$. Regardless of the knot choice, De Boor (1978) describes a recursive procedure for evaluating the i^{th} B-spline basis function of degree q at a point x:

$$B_i^q(x) = \frac{x - v_i}{v_{i+q+1} - v_m} B_i^{q-1}(x) + \frac{v_{i+q+2} - x}{v_{i+q+2} - v_{i+1}} B_{i+1}^{q-1}(x)$$

and where

$$B_i^{-1}(x) = \begin{cases} 1 & \text{if } v_q \leq x < v_{q+1}, \\ 0 & \text{otherwise.} \end{cases}$$

However, for practical purposes, it is computationally convenient to compute differences of truncated power functions to construct a B-spline basis of degree q, see Eilers & Marx (2010) for a summary and an implementation in the R language (R Development Core Team (2011)). For an illustration of B-spline bases of 10 functions with degrees 0, 1, 2 and 3, see Figure 2.2.

2.2.2 Simple flexible regression

The general idea of smoothing with basis functions is simply to recast the mean function, f, as a linear combination of the set of basis functions evaluated at each covariate value, rather than global polynomial functions as used in ordinary regression. For fitting the effect of a smooth function f of a covariate $\boldsymbol{x} = (x_1, \ldots, x_n)$ on a response $\boldsymbol{y} = (y_1, \ldots, y_n)$, using a set of p basis functions $\{B_j() : 1 \leq j \leq p\}$, then a flexible model can be expressed as

$$y_i = f(x_i) + \epsilon_i = \sum_{j=1}^{p} B_j(x_i)\alpha_j + \epsilon_i \tag{2.2}$$

where $\epsilon_i \sim N(0, \sigma^2)$. For the data shown in the top-left panel Figure 2.3 an example set of 30 uniformly-spaced basis functions can be constructed, evaluated at each of the data points x_i, so that the resulting matrix of basis function evaluations is given by

$$\boldsymbol{B} = \begin{pmatrix} B_1(x_1) & B_2(x_1) & \ldots & B_{30}(x_1) \\ B_1(x_2) & B_2(x_2) & \ldots & B_{30}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_1(x_n) & B_2(x_n) & \ldots & B_{30}(x_n) \end{pmatrix}. \tag{2.3}$$

The relatively large number of 30 functions were chosen to illustrate the overfitting that results from fitting curves using this basis and no smoothness control. For illustration, the set of 30 basis functions evaluated at each point on a fine grid over the x-axis are shown in the top right-hand panel of Figure 2.3. In matrix notation, the model described

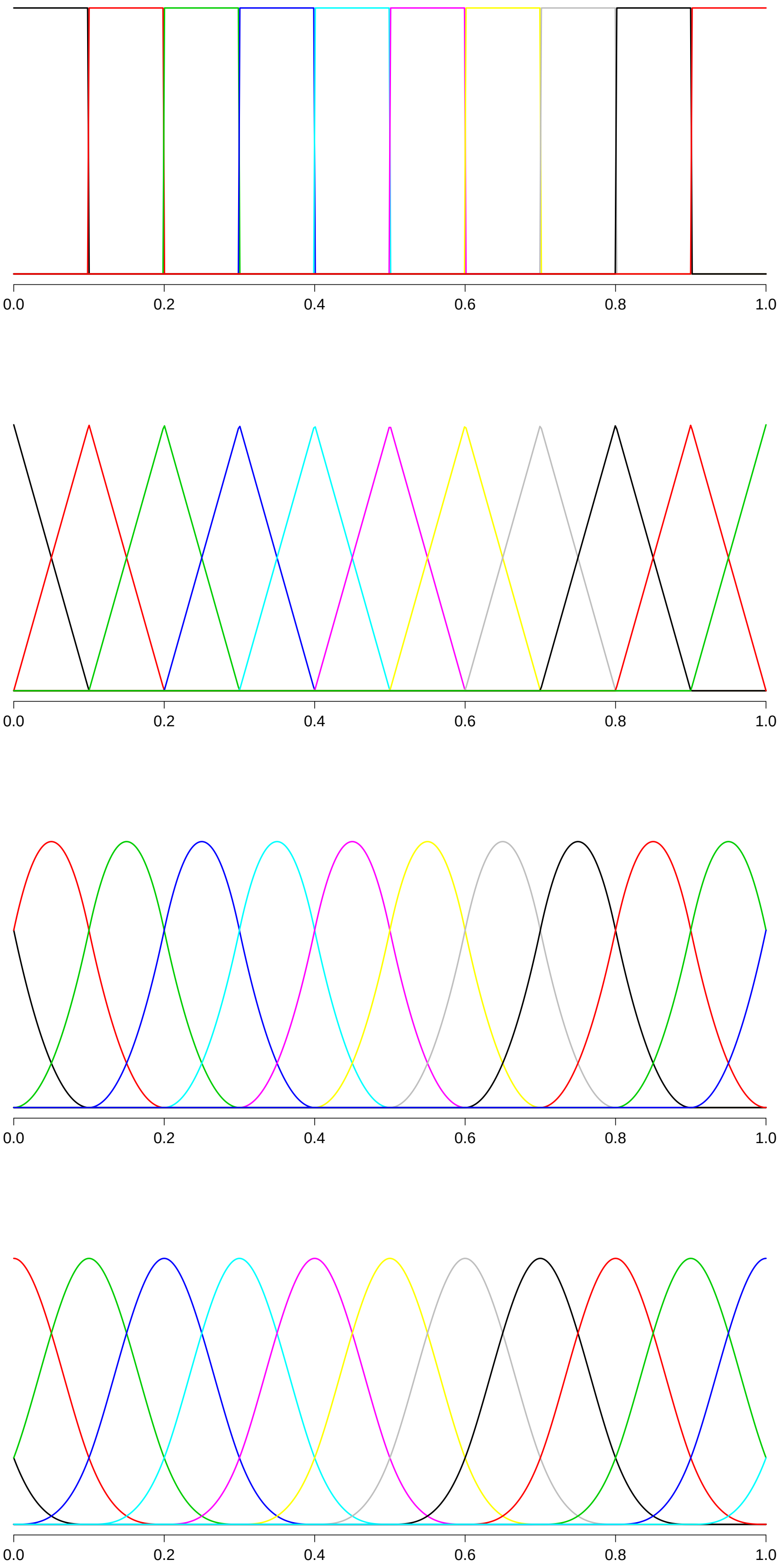


FIGURE 2.2: From top panel: 10 overlapping B-spline basis functions of degree 0, 1, 2 and 3 respectively.

in Equation 2.2 can be rewritten

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

parameter estimates $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \dots, \hat{\alpha}_p)$ associated with each of the basis functions in Equation 2.3 are available by least squares by minimising the quadratic form

$$(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha})^\top(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha}) \tag{2.4}$$

with respect to $\boldsymbol{\alpha}$. The closed form expression for the minimising value for $\boldsymbol{\alpha}$ of Equation 2.4 is $\hat{\boldsymbol{\alpha}} = (\boldsymbol{B}^\top\boldsymbol{B})^{-1}\boldsymbol{B}^\top\boldsymbol{y}$; fitted values for $\boldsymbol{y}$ are then given by

$$\hat{\boldsymbol{y}} = \boldsymbol{B}\hat{\boldsymbol{\alpha}} = \boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B})^{-1}\boldsymbol{B}^\top\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \tag{2.5}$$

where $\boldsymbol{H}$ is known as the *smoother* or *hat* matrix. Here, the choice of the basis size p was arbitrary and the fitted function f is sensitive to this choice, with wiggliness of f increasing with p. This is particularly visible for the example data shown in Figure 2.3: the bottom left panel shows the result of fitting with the basis without any restriction on the curvature, where the estimated curve is clearly overfitting the data. Fitting with P-splines under the framework set out by Eilers & Marx (1996) avoids sensitivity to the chosen basis dimension by incorporating a penalty on neighbourhoods of the parameters $\boldsymbol{\alpha}$. This is achieved by including in the least squares objective function the sum-of-squared first order differences

$$\sum_{i=1}^{p-1}(\alpha_{i+1} - \alpha_i)^2, \tag{2.6}$$

which has an interpretation as a measure of roughness of $\boldsymbol{\alpha}$ and therefore the smoothness of the fitted function $\hat{f}$. Going further, the roughness expression in Equation 2.6 can be multiplied by a scalar quantity λ, which in turn increases the influence of the roughness

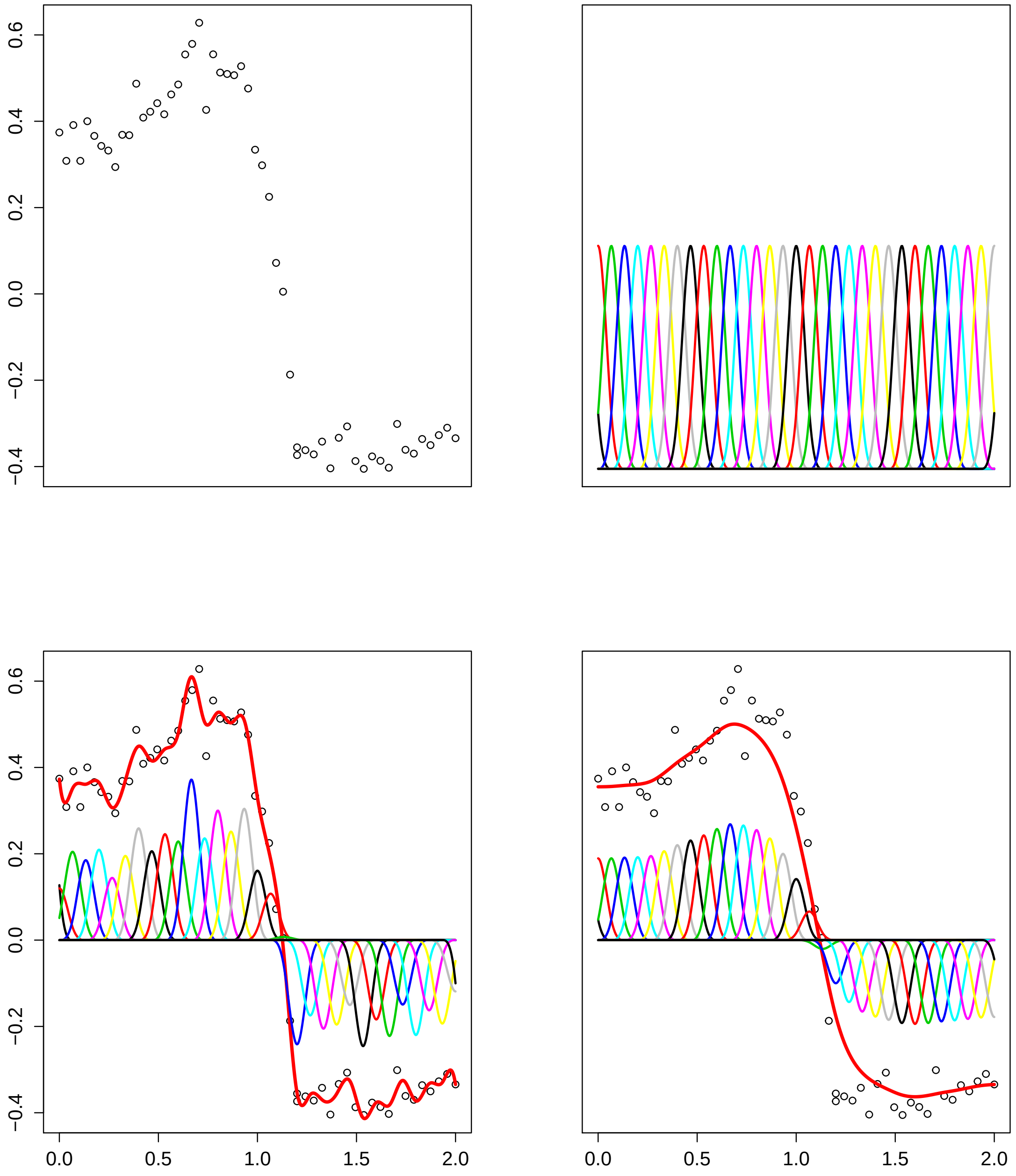


FIGURE 2.3: Top panels from left: plot of non-linear data series; set of 30 overlapping and uniformly spaced B-spline basis functions which will be used to obtain a smooth fit to the data series. Bottom panels from left: red line shows an unpenalised fit resulting from the chosen basis, basis function heights are proportional to the estimated coefficients $\hat{\boldsymbol{\alpha}}$; red line shows a penalised fit to the data, basis function heights are visibly smoother than under the unpenalised fit.

measure on the estimator of $\boldsymbol{\alpha}$ and curves with different smoothness properties will result. The new objective function then becomes

$$(\boldsymbol{y} - \boldsymbol{B\alpha})^\top(\boldsymbol{y} - \boldsymbol{B\alpha}) + \lambda\boldsymbol{\alpha}^\top\boldsymbol{D}^\top\boldsymbol{D\alpha} \tag{2.7}$$

where

$$\boldsymbol{D} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \end{pmatrix}. \tag{2.8}$$

Higher order penalties than that implied by Equation 2.8 are possible, for example by incorporating the squared second order differences

$$\sum_{i=1}^{p-2}(\alpha_{i+2} - 2\alpha_{i+1} + \alpha_i)^2, \tag{2.9}$$

whose corresponding difference matrix is defined as

$$\boldsymbol{D} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \end{pmatrix}, \tag{2.10}$$

different shaped smooths with different properties will result. It is helpful to consider why this happens: in the case of Equation 2.6, fitted functions that are just horizontal lines incur the lowest penalty (because in this case $\alpha_1 = \alpha_2 = \ldots = \alpha_p$), and so the first order roughness penalty represents deviations from a constant function. For the second order case in Equation 2.9, the lowest penalty is incurred when $\alpha_{i+2} - \alpha_{i+1} = \alpha_{i+1} - \alpha_i$, for $1 \leq i \leq p-2$, which occurs when the fitted function is linear, and hence the penalty represents deviations from a straight-line relationship for a given set of $\boldsymbol{\alpha}$. As a result of these differencing properties, when $\lambda \rightarrow \infty$ under first order difference penalties the fitted function $\hat{f}$ tends to a constant, whereas under second order differences the result is a linear $\hat{f}$. The latter is typically the preferred behaviour when $\lambda \rightarrow \infty$, because a linear model is often viewed as a default model choice when no smoothness is expected, and makes sense as a limiting case under strong smoothing. Other differencing choices could be used, but without strong prior scientific knowledge they risk introducing bias, for example when the true relationship in the data is linear and strong smoothing results in some other fitted function. As a result of these considerations, second order difference penalties are used throughout this thesis in combination with a set of degree 3 B-spline basis functions, the latter ensures that the resulting curves are relatively smooth and twice differentiable which are attractive properties for many environmental modelling applications. Although these choices are made in a way that seems slightly ad-hoc, it is not likely that the results and subsequent inferences would be substantially altered if different choices (i.e. higher order differences and higher degree basis functions) were made.

In the case of continuous responses and assuming Gaussian errors, for a fixed value of λ, $\boldsymbol{\alpha}$ is estimated by solving the equations

$$\boldsymbol{B}^{\top}\boldsymbol{y} = (\boldsymbol{B}^{\top}\boldsymbol{B} + \lambda\boldsymbol{D}^{\top}\boldsymbol{D})\boldsymbol{\alpha}$$

The scalar $\boldsymbol{\alpha}^{\top}\boldsymbol{D}^{\top}\boldsymbol{D}\boldsymbol{\alpha}$ measures the roughness of f via the parameters $\boldsymbol{\alpha}$. λ is a smoothness control parameter that modulates the extent to which $\boldsymbol{\alpha}^{\top}\boldsymbol{D}^{\top}\boldsymbol{D}\boldsymbol{\alpha}$ influences the estimates of $\boldsymbol{\alpha}$ resulting from 2.7, and restricts the range of shapes that f can take across values of x. The benefit of enforcing this restriction is clear by considering the bottom right panel of Figure 2.3 where $\hat{\boldsymbol{\alpha}}$ has been estimated with $\lambda = 10$ and the result is a smoother, more parsimonious fit to the data.

When $\hat{f}$ is wiggly, the fit to each data point is closer and the sum of squares $(\boldsymbol{y}-\boldsymbol{B}\boldsymbol{\alpha})^{\top}(\boldsymbol{y}-$

$\boldsymbol{B\alpha}$) is small and $\boldsymbol{\alpha}^\top \boldsymbol{D}^\top \boldsymbol{D\alpha}$ is large; correspondingly when $\hat{f}$ is smooth, the fitted function has less scope to achieve a good fit to each datum and $(\boldsymbol{y} - \boldsymbol{B\alpha})^\top(\boldsymbol{y} - \boldsymbol{B\alpha})$ is larger and $\boldsymbol{\alpha}^\top \boldsymbol{D}^\top \boldsymbol{D\alpha}$ is smaller. The introduction of the λ parameter provides a mechanism for choosing an optimal smoothing level: small λ yields less penalised models with low residual variance but a large effective number of parameters, while larger λ yields more structured models with higher residual variance and a lower dimension. Therefore, a trade off between model compexity and fit must be achieved for which some algorithms are described in Section 2.3.

2.2.3 Additive modelling

General and detailed discussions of additive models are available in Hastie & Tibshirani (1990), Green *et al.* (1994), Ruppert *et al.* (2003), Wood (2006) and Fahrmeir *et al.* (2013), that each include extensions such as complex mixed effects models and modelling with non-normal responses. Here attention is focused on additive models using the P-splines framework described in Marx & Eilers (1998). The flexible regression described in Section 2.2.2 generalises to the setting of additive models, where a set of q covariates $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_q\}$ is to be regressed on $\boldsymbol{y}$ and the contribution of each is assumed to take the form of some unknown smooth function f_j:

$$y_i = \beta_0 + \sum_{j=1}^{q} f_j(x_{i,j}) + \epsilon_i.$$

Parameter estimates are available by penalised least squares using

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{P})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

where the model matrix $\boldsymbol{X}$ can be written in the augmented form

$$\boldsymbol{X} = \left[\begin{array}{c|c|c|c|c} \mathbf{1} & \boldsymbol{B}_1(\boldsymbol{x}_1) & \boldsymbol{B}_2(\boldsymbol{x}_2) & \dots & \boldsymbol{B}_q(\boldsymbol{x}_q) \end{array}\right]. \tag{2.11}$$

In Equation 2.11, $\boldsymbol{B}_j()$ is the B-spline basis expansion of q_j basis functions of the j^{th} covariate, $\boldsymbol{x}_j$. The penalty matrix $\boldsymbol{P} = \boldsymbol{P}(\boldsymbol{\lambda}) = \boldsymbol{P}(\lambda_1, \ldots, \lambda_q)$ is a block diagonal matrix which is now a function of m smoothness control parameters such that

$$\boldsymbol{P} = \begin{pmatrix} \lambda_1 \boldsymbol{D}_1^\top \boldsymbol{D}_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 \boldsymbol{D}_2^\top \boldsymbol{D}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_q \boldsymbol{D}_q^\top \boldsymbol{D}_q \end{pmatrix}.$$

Expressions for the variance of the fitted values and model parameters are available via

$$\begin{aligned} \mathsf{Var}\,(\hat{\boldsymbol{\alpha}}) &= \mathsf{Var}\left[(\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{P}(\boldsymbol{\lambda}))^{-1} \boldsymbol{X}^\top \boldsymbol{y}\right] \\ &= \hat{\sigma}^2 (\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{P}(\boldsymbol{\lambda}))^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{P}(\boldsymbol{\lambda}))^{-1} \end{aligned}$$

$$\begin{aligned} \mathsf{Var}\,(\hat{\boldsymbol{y}}) &= \mathsf{Var}\left[\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{P}(\boldsymbol{\lambda}))^{-1} \boldsymbol{X}^\top \boldsymbol{y}\right] \\ &= \hat{\sigma}^2 \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{P}(\boldsymbol{\lambda}))^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{P}(\boldsymbol{\lambda}))^{-1} \boldsymbol{X}^\top \qquad (2.12) \end{aligned}$$

Estimating the residual variance requires some notion of the model and residual degrees of freedom, $\mathsf{df}_{\mathsf{model}}$ and $\mathsf{df}_{\mathsf{error}}$ respectively. For a standard Gaussian linear model these are defined simply as a function of the idempotent projection matrix, $\boldsymbol{H}$ defined as $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ so that

$$\begin{aligned} \mathsf{df}_{\mathsf{model}} &= \mathsf{tr}\,(\boldsymbol{H}) & \mathsf{df}_{\mathsf{error}} &= \mathsf{tr}\,(\boldsymbol{I} - \boldsymbol{H})\,(\boldsymbol{I} - \boldsymbol{H})^\top \\ &= \mathsf{tr}\left[\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top\right] & &= \mathsf{tr}(\boldsymbol{I}) + \mathsf{tr}(\boldsymbol{H}\boldsymbol{H}^\top) - 2\mathsf{tr}(\boldsymbol{H}) \\ &= \mathsf{tr}\left[\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right] & &= \mathsf{tr}(\boldsymbol{I}) - \mathsf{tr}(\boldsymbol{H}) \\ &= \mathsf{tr}\,[\boldsymbol{I}_p] = p & &= n - p. \end{aligned}$$

Here, p is simply the number of columns of $\boldsymbol{X}$ or the number of parameters being fitted, and it follows that the residual variance is estimated as $\hat{\sigma}^2 = \frac{\mathsf{RSS}}{\mathsf{df}_{\mathsf{error}}} = \frac{\mathsf{RSS}}{n-p}$. By analogy with linear modelling, Hastie & Tibshirani (1990) estimate that the model and residual degrees of freedom in a semiparametric model can be obtained from the equivalent smoothing matrix $\boldsymbol{H}$ as defined in Equation 2.5. In this case, $\boldsymbol{H}$ is not a projection matrix and is not idempotent and therefore for additive and other semiparametric models

$$\begin{aligned} \mathsf{df}_{\mathsf{model}} &= \mathsf{tr}(\boldsymbol{H}) = \mathsf{tr}\left[\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{P})^{-1}\boldsymbol{X}^\top\right] & \mathsf{df}_{\mathsf{error}} &= \mathsf{tr}\left(\boldsymbol{I}-\boldsymbol{H}\right)\left(\boldsymbol{I}-\boldsymbol{H}\right)^\top \\ &= \mathsf{tr}\left[\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{P})^{-1}\right] & &= \mathsf{tr}(\boldsymbol{I})+\mathsf{tr}(\boldsymbol{H}\boldsymbol{H}^\top)-2\mathsf{tr}(\boldsymbol{H}) \end{aligned} \tag{2.13}$$

In particular, Ruppert *et al.* (2003) show that for moderate levels of λ, $\mathsf{RSS}/(n-\mathsf{df}_{\mathsf{model}})$, is a biased estimator for $\hat{\sigma}^2$ and that $\hat{\sigma}^2 = \mathsf{RSS}/(n-2\mathsf{tr}(\boldsymbol{H})+\mathsf{tr}(\boldsymbol{H}\boldsymbol{H}^\top))$ should be used instead. $\mathsf{df}_{\mathsf{model}}$ is known as the *effective degrees of freedom* or *effective dimension*, and in order to distinguish it from the ordinary linear model definition it is denoted ED throughout this thesis. Since each smooth component involved in an additive model involves several basis parameters, it is also useful to define the model degrees of freedom associated with each smooth component. Assuming a model matrix of the form described in Equation 2.11, the fitted values can be written

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{H}\boldsymbol{y} \\ &= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{P}(\boldsymbol{\lambda}))^{-1}\boldsymbol{X}^\top\boldsymbol{y} \\ &= \boldsymbol{X}_{\boldsymbol{1}}(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{P}(\boldsymbol{\lambda}))^{-1}\boldsymbol{X}_{\boldsymbol{1}}^\top\boldsymbol{y} \\ &\quad +\boldsymbol{X}_{\boldsymbol{B}_1}(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{P}(\boldsymbol{\lambda}))^{-1}\boldsymbol{X}_{\boldsymbol{B}_1}^\top\boldsymbol{y} \\ &\quad\quad +\boldsymbol{X}_{\boldsymbol{B}_2}(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{P}(\boldsymbol{\lambda}))^{-1}\boldsymbol{X}_{\boldsymbol{B}_2}^\top\boldsymbol{y} \\ &\quad\quad\quad +\vdots \qquad\qquad \vdots \\ &\quad\quad\quad\quad +\boldsymbol{X}_{\boldsymbol{B}_q}(\boldsymbol{X}^\top\boldsymbol{X}+\boldsymbol{P}(\boldsymbol{\lambda}))^{-1}\boldsymbol{X}_{\boldsymbol{B}_q}^\top\boldsymbol{y} \\ &= \boldsymbol{H}_0\boldsymbol{y}+\boldsymbol{H}_1\boldsymbol{y}+\boldsymbol{H}_2\boldsymbol{y}+\ldots+\boldsymbol{H}_q\boldsymbol{y} \end{aligned}$$

where $\boldsymbol{X}_{\boldsymbol{B}_i}$ is the model matrix $\boldsymbol{X}$ in which all elements not associated with the smooth component $\boldsymbol{B}_i$ have been set to zero. Then the effective degrees of freedom associated with smooth component i is given by $\mathsf{ED}_i = \mathsf{tr}(\boldsymbol{H}_i)$.

2.2.4 Interaction terms

Often in practical modelling applications, the individual effects of two covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ may not be adequate to account for variation in a response, because the smooth effect of $\boldsymbol{x}_1$ depends in turn on the level of $\boldsymbol{x}_2$. In such a setting, a *bivariate* smooth interaction term is required, and in the setting of P-splines this corresponds to fitting a surface to represent the joint effect of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. If the marginal basis matrices for the two covariates of interest are called $\boldsymbol{B}_{p_1}(\boldsymbol{x}_1)$ and $\boldsymbol{B}_{p_2}(\boldsymbol{x}_2)$ with dimensions $n \times p_1$ and $n \times p_2$ respectively, then the basis matrix for the interaction surface is the row-wise Kronecker product or *box product* (Eilers *et al.* (2006)) of these, which is defined as

$$\boldsymbol{B}_{p_1}(\boldsymbol{x}_1) \Box \boldsymbol{B}_{p_2}(\boldsymbol{x}_2) = (\boldsymbol{B}_{p_1}(\boldsymbol{x}_1) \otimes \mathbf{1}_{1\times p_2}) \odot (\mathbf{1}_{p_1 \times 1} \otimes \boldsymbol{B}_{p_2}(\boldsymbol{x}_2)) \tag{2.14}$$

where $\odot$ is the element-wise or Hadamard product and $\otimes$ is the Kronecker product. The matrix in Equation 2.14 has $p_1 \times p_2$ columns, each of which is associated with a parameter in the parameter vector $\boldsymbol{\alpha} = (\alpha_{11}, \ldots, \alpha_{1p_2}, \ldots, \alpha_{p_1 1}, \ldots, \alpha_{p_1 p_2})$. It then remains to choose a suitable penalty on the flexibility of the fit described by $[\boldsymbol{B}_{p_1}(\boldsymbol{x}_1) \Box \boldsymbol{B}_{p_2}(\boldsymbol{x}_2)]\,\boldsymbol{\alpha}$. Using the same ideas that underlie the penalties constructed for univariate smooth functions, a sensible roughness measure might sum squared differences along each margin of the grid of parameters expressed by the 'flattened' vector $\boldsymbol{\alpha}$:

$$\sum_{j=1}^{p_2} \sum_{i=1}^{p_1-2} (\alpha_{i,j} - 2\alpha_{i+1,j} + \alpha_{i+2,j})^2 \\ + \sum_{j=1}^{p_2-2} \sum_{i=1}^{p_1} (\alpha_{i,j} - 2\alpha_{i,j+1} + \alpha_{i,j+2})^2$$

In matrix notation, Equation 2.15 can be expressed

$$\boldsymbol{\alpha}^\top \left(\boldsymbol{D}_{p_1}^\top \boldsymbol{D}_{p_1} \otimes \boldsymbol{I}_{p_2} + \boldsymbol{I}_{p_1} \otimes \boldsymbol{D}_{p_2}^\top \boldsymbol{D}_{p_2} \right) \boldsymbol{\alpha}$$

where $\boldsymbol{D}_p$ is a second order difference matrix with p columns, and $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. Finally, at least a single smoothness control parameter is required to control the influence of the roughness penalty and the resulting wiggliness of the surface estimated by the parameters $\boldsymbol{\alpha}$. There are two possible options, using a single smoothing parameter λ:

$$\lambda \boldsymbol{\alpha}^\top \left(\boldsymbol{D}_{p_1}^\top \boldsymbol{D}_{p_1} \otimes \boldsymbol{I}_{p_2} + \boldsymbol{I}_{p_1} \otimes \boldsymbol{D}_{p_2}^\top \boldsymbol{D}_{p_2} \right) \boldsymbol{\alpha}, \tag{2.15}$$

or using a pair of parameters (λ_1, λ_2)

$$\boldsymbol{\alpha}^\top \left(\lambda_1 \boldsymbol{D}_{p_1}^\top \boldsymbol{D}_{p_1} \otimes \boldsymbol{I}_{p_2} + \boldsymbol{I}_{p_1} \otimes \lambda_2 \boldsymbol{D}_{p_2}^\top \boldsymbol{D}_{p_2} \right) \boldsymbol{\alpha}. \tag{2.16}$$

Equation 2.15 describes an *isotropic* smooth, in which the strength of penalisation is the same across both axes ($\boldsymbol{x}_1$ and $\boldsymbol{x}_2$) of the space and results in an *isotropic* smooth. Isotropic smooth surfaces are special cases that are useful when the units of the two covariates are the same, or where the smoothness associated with axis defined by each variable is thought to be the same. Under the assumption of isotropy, only a single smoothing parameter is needed to represent the smoothness of the surface, which grants a level of computational ease. However, it is usually not possible to make the assumption of isotropy, in which case Equation 2.16 describes the more appropriate penalty term.

In general, interaction smooths are computationally expensive due to the requirement of handling a matrix of dimension $p_1 \times p_2$. However, in the special case where the data lie on a regular 2D (or higher dimensional) grid of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, Equation 2.14 is column equivalent to

$$\boldsymbol{B}(\boldsymbol{x}_1)_{n \times p_1} \otimes \boldsymbol{B}(\boldsymbol{x}_2)_{n \times p_2}$$

and Eilers *et al.* (2006) show how in this case, and in more general cases where the model matrix is constructed from a sequence of Kronecker products, that much of the computational overhead can be reduced by exploiting properties of Kronecker products.

2.2.5 Varying coefficient models

So far, the models described are designed to smooth the data directly or to represent the influence of a covariate or a set of covariates on a response variable as smooth functions. However, it is sometimes necessary to construct a model in which the linear (or smooth) effect of one covariate is allowed to change smoothly and non-linearly according to the influence of another, and the result is called a varying coefficient model (Hastie & Tibshirani (1993)). This is particularly useful in settings where patterns change with time, and is the main idea underpinning the models that are developed in Chapter 3. For example, considering a covariate $\boldsymbol{x} = (x_1, \ldots, x_n)$, a response $\boldsymbol{y} = (y_1, \ldots, y_n)$ and a time index $\boldsymbol{t} = (t_1, \ldots, t_n)$, then a time-varying coefficient model for $\boldsymbol{y}$ and $\boldsymbol{x}$ could be expressed using P-splines as follows

$$\begin{aligned} y_i &= f(t_i)x_i + \epsilon_i \\ &= \sum_{j=1}^{p} \alpha_j B_j(t_i)x_i + \epsilon_i \end{aligned} \tag{2.17}$$

where f is a smooth function that is constructed from a set of p evenly spaced B-spline basis functions. The model described by Equation 2.17 is different from the bivariate interaction described in Section 2.2.4 because at a particular time point, the impact of x_i on y_i is linear, but the gradient associated with this linear effect has the flexibility to vary with time. In matrix notation, Equation 2.17 can be re-expressed as

$$\boldsymbol{y} = \left((\boldsymbol{1}_{1\times p} \otimes \boldsymbol{x}) \odot \boldsymbol{B}(\boldsymbol{t}) \right) \boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{2.18}$$

where $\odot$ is the element-wise product and $\otimes$ is the Kronecker product. Parameter estimates can be obtained in the same manner as previously described, by penalised least squares, by minimising

$$(\boldsymbol{y} - ((\mathbf{1}_{1\times p} \otimes \boldsymbol{x}) \odot \boldsymbol{B}(\boldsymbol{t})\boldsymbol{\alpha}))^\top (\boldsymbol{y} - ((\mathbf{1}_{1\times p} \otimes \boldsymbol{x}) \odot \boldsymbol{B}(\boldsymbol{t})\boldsymbol{\alpha})) + \lambda\boldsymbol{\alpha}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\alpha}$$

where $\lambda\boldsymbol{\alpha}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\alpha}$ is a roughness penalty over neighbourhoods of the parameter vector $\boldsymbol{\alpha}$. Applications of this type of model include 3D imaging (Heim *et al.* (2007)), proportional hazards modelling (Lambert & Eilers (2005)), phenology (Roberts (2008)) and spectroscopy (Eilers & Marx (2002)).

2.2.6 Sparse matrices

A matrix is deemed to be *sparse* when a high proportion of its elements are equal to zero. The importance of sparsity lies in its potential to reduce the computational cost associated with linear algebra, by avoiding performing any multiplications that involve zero elements. Well established algorithms exist for storing sparse matrices and for performing linear algebra, described for example by Davis (2006) and Ng & Peyton (1993). Most of these algorithms are implemented in low level `C` code for speed, but have more recently been incorporated in R (R Development Core Team (2011)) via high level functions in the packages `Matrix` (Bates & Maechler (2013)) which is a general-purpose suite of sparse matrix software, `spam` (Furrer & Sain (2010)) which is tailored to some specific computations that arise in Bayesian spatial modelling and `SparseM` (Koenker & Ng (2013)).

Sparsity is relevant to the current flexible regression context because a defining characteristic of a B-spline basis function is that it is non-zero over only a short interval. Model matrices for P-splines are therefore typically sparse, particularly when the basis dimension is very large ($p > 100$) and some authors have taken steps to exploit this structure, see for example Eilers & Marx (2010). This is especially helpful in situations where very large Kronecker product bases are required for capturing smooth interaction terms, as described in Section 2.2.4, and also when matrices composed of dummy variables are involved in the

Kronecker product as is encountered in Section 4.5.

There are a number of steps in the estimation of semiparametric models that are particularly computationally burdensome and can be sped up by sparse matrix algorithms: constructing and storing the model matrix $\boldsymbol{X}$ may be difficult when large Kronecker product matrices are involved, and most importantly, obtaining repeated evaluations of the diagonal elements of the hat matrix $\boldsymbol{H}$ when selecting smoothing parameters.

2.3 Smoothing parameter selection

High values of λ result in stronger smoothing, while small values allow more flexibility in the shape of the fitted curve. An example is given in Figure 2.4, where estimation with small values of λ results in fitted curves that overfit the data, and higher values oversmoothing.

After inspection of the different strengths of smoothness shown in Figure 2.4 it seems uncontroversial that the green line, representing a moderate level of smoothness, with $\lambda = 2$, provides a good representation of the underlying signal in the data. In fact, for simple univariate regression settings such as that presented in Figure 2.4, visual selection of different smoothness scenarios may be adequate for selecting an optimal value that neither overfits or oversmooths. However, in more complex settings where the model is composed of smooth functions of multiple variables, implying several smoothing parameters, selection of the optimal λ parameters is much more difficult or impossible by such visual inspection alone, and a more formal approach is usually adopted. The problem of automatic smoothing parameter selection is still an area of intensive research, and a number of competing approaches currently exist that are suitable for fitting models in which multiple smoothing parameters are to be chosen.

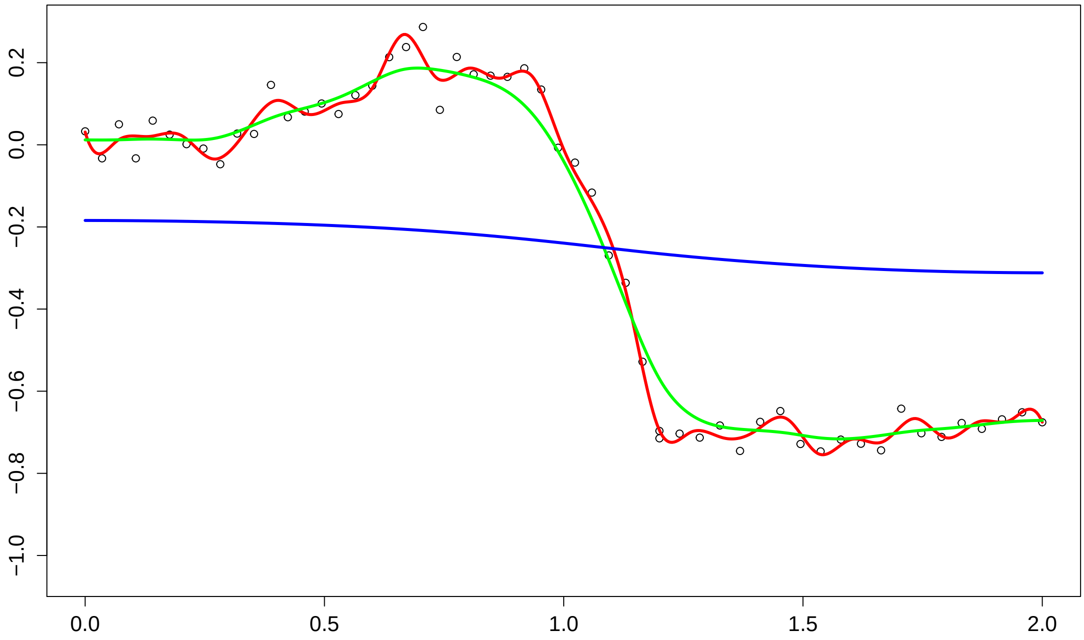


FIGURE 2.4: Three different smoothing strengths applied to the data example shown in Figure 2.1. Red curve corresponds to $\lambda = 10^{-4}$, green curve to $\lambda = 2$ and the blue curve to $\lambda = 10^3$.

2.3.1 Automatic selection with performance criteria

The optimal smoothing parameter vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)$ should strike a balance between a good model fit as measured by the residual sum of squared errors, and model complexity as measured by the effective degrees of freedom described in Equation 2.13. A popular suggestion to achieve this balance has been to select the $\boldsymbol{\lambda}$ that optimises some model fit performance criterion, for example, Eilers & Marx (1996) advocate minimising Akaike's Information Criterion, Akaike (1973), (AIC) which is defined by Hastie & Tibshirani (1990) for semiparametric models as

$$\mathsf{AIC} \;=\; \frac{1}{n}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y})^\top(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}) + 2\mathrm{tr}(\boldsymbol{H})$$

over a logarithmic grid of λ values. Wood (2000) proposes an efficient Newton search procedure for minimising the Generalized Cross-Validation score (GCV) of Craven & Wahba

(1978) over λ where

$$\mathsf{GCV} = \frac{\frac{1}{n}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y})^\top(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y})}{(1 - \mathsf{tr}(\boldsymbol{H})/n)^2}$$

However, Hurvich & Tsai (1989) find evidence that AIC has a tendency to overfit due to bias when samples are small, and is particularly conspicuous when the dimensionality of the model approaches that of the sample size. Therefore Hurvich *et al.* (2002) propose a 'bias - corrected' AIC called AICc, defined as

$$\mathsf{AICc} = \log(\hat{\sigma}^2) + 1 + \frac{2(\mathsf{tr}(\boldsymbol{H}) + 1)}{n - \mathsf{tr}(\boldsymbol{H}) - 2} \tag{2.19}$$

where $\hat{\sigma}^2 = (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y})^\top(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y})/n$, and is designed to more strongly penalise complex models, and is thought to be more appropriate when overfitting can occur such as in semiparametric regression. A comparison of how these three criteria compare for a single set of data across a range of λ values is shown in Figure 2.5, where it can be seen that the strongest smooth (largest λ) has been selected using AICc and the weakest by AIC. Of course, many other measures of model performance exist such as the Unbiased Risk Estimator (UBRE) of Craven & Wahba (1978), or the Bayesian Information Criterion (BIC) of Schwarz (1978).

2.3.2 Calculating tr($\boldsymbol{H}$)

Regardless of the particular performance measure, the main hindrance to performing an effective selection procedure is computational, because each evaluation of the measure requires the calculation of $\mathsf{ED}(\boldsymbol{\lambda}) = \mathsf{tr}(\boldsymbol{H})$. Since $\boldsymbol{H}$ is $n \times n$, evaluation of its trace is burdensome for large n. Some relief may be found by rearranging Equation 2.13 by

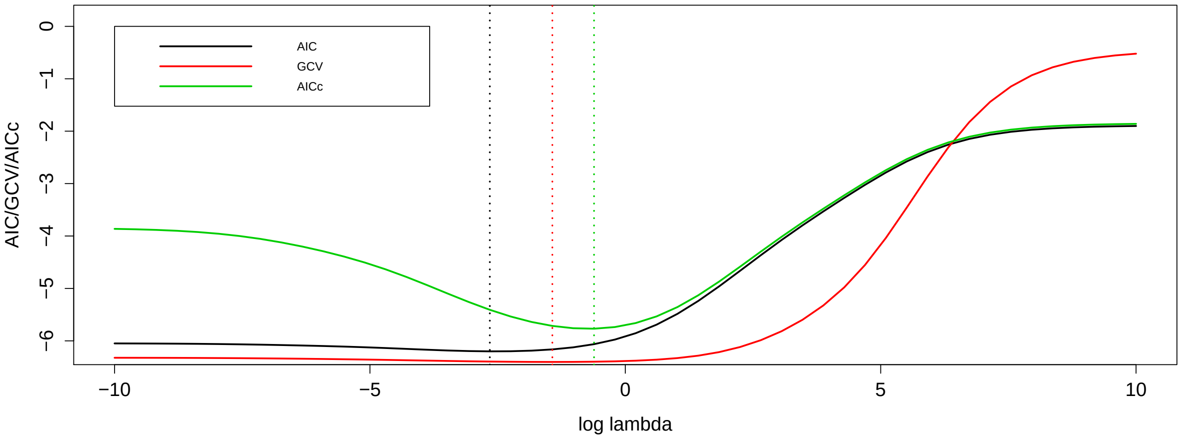


FIGURE 2.5: Profiles of AIC, GCV and AICc across a log grid of values of λ for the data shown in the right panel of Figure 2.1. Vertical dashed lines correspond to the minimum values for each profile: for AIC the minimum is achieved when $\lambda = 0.07$, for GCV the minimum is at $\lambda = 0.24$ and for AICc the minimising $\lambda = 0.54$.

recognising that the trace operation is invariant under permutation and so

$$\begin{aligned} \mathsf{tr}(\boldsymbol{H}) &= \mathsf{tr}\left[\boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B} + \lambda\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{B}^\top\right] \\ &= \mathsf{tr}\left[(\boldsymbol{B}^\top\boldsymbol{B} + \lambda\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{B}^\top\boldsymbol{B}\right] \\ &= \sum_i^p\sum_j^p \left[(\boldsymbol{B}^\top\boldsymbol{B} + \lambda\boldsymbol{D}^\top\boldsymbol{D})^{-1} \odot \boldsymbol{B}^\top\boldsymbol{B}\right]_{ij} \end{aligned} \tag{2.20}$$

and hence it is only required to manipulate $p \times p$ matrices. This reduces the computational overhead and is worthwhile when $p << n$, however in some settings, for example those arising in the river network modelling described in Chapters 4 and 5, both p and n may both be large enough that a single evaluation of $\mathsf{tr}(\boldsymbol{H})$ can be very expensive. In such cases, it can be helpful to find a rapid approximation to $\mathsf{tr}(\boldsymbol{H})$, provided that the associated approximation error is small relative to the features of the criterion surface over $\boldsymbol{\lambda}$.

The overall fit of each smooth component is robust to modest changes in the associated λ smoothing parameter, which is a feature that can be exploited to obtain faster evaluations of a performance criteria that depends on $\boldsymbol{\lambda}$. Therefore, a fruitful approach might lie in finding a way to approximate $\mathsf{tr}(\boldsymbol{H})$, assuming that the approximation can be obtained more cheaply than performing the exact calculations described in Equation 2.20. Two results due to Hutchinson (1989) mean that this is possible.

Lemma 1: (Hutchinson (1989)) Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix and let $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$ be a vector of n independent samples from a random variable U with mean zero and variance σ^2. Then

$$\mathbb{E}(\boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{u}) = \sigma^2 \mathsf{tr}(\boldsymbol{A}) \tag{2.21}$$

Proposition 1: (Hutchinson (1989)) Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix with non-zero trace. Let U be the discrete random variable which takes the values 1, -1, each with probability $\frac{1}{2}$ and let $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$ be a vector of n samples from U. Then $\boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{u}$ is an unbiased estimator of tr$(\boldsymbol{A})$ and

$$\mathsf{Var}(\boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{u}) = 2 \sum_{i \neq j} a_{ij}^2 \tag{2.22}$$

Moreover, U is the unique random variable amongst zero mean random variables for which $\boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{u}$ is a minimum variance, unbiased estimate of $\mathsf{tr}(\boldsymbol{A})$.

Lemma 1 and Equation 2.21 show how it is possible to obtain an unbiased stochastic estimate of the trace of a matrix using a random vector $\boldsymbol{u}$. Furthermore, Proposition 1 and Equation 2.22 show that by choosing for $\boldsymbol{u}$ the particular random vector composed of -1 and 1 each appearing with probability $\frac{1}{2}$, the estimator has minimum possible variance. These results suggest that the Monte Carlo estimator $\mathsf{tr}(\boldsymbol{H}) \approx \tilde{\mathsf{tr}}(\boldsymbol{H}) = \frac{1}{s} \sum_i^s \boldsymbol{u}_i^\top \boldsymbol{H} \boldsymbol{u}_i$ can be used for suitably large enough s. In practice, this can be evaluated by making use of the Choleski decomposition $\boldsymbol{L}$, which is defined as $\boldsymbol{L}\boldsymbol{L}^\top = (\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{D}^\top \boldsymbol{D})$. The Choleski decomposition is a square root matrix which has sparseness properties that inherit from

$(\boldsymbol{B}^\top\boldsymbol{B} + \lambda\boldsymbol{D}^\top\boldsymbol{D})$ and can therefore be obtained cheaply.

$$\begin{aligned}
\mathsf{tr}(\boldsymbol{H}) &= \mathsf{tr}\left[\boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B} + \lambda\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{B}^\top\right] \\
&= \mathsf{tr}\left[\boldsymbol{B}\boldsymbol{L}^{-1}\boldsymbol{L}^{-T}\boldsymbol{B}^\top\right] \\
&\approx \frac{1}{s}\sum_i^s \boldsymbol{u}_i^\top\boldsymbol{B}\boldsymbol{L}^{-1}\boldsymbol{L}^{-T}\boldsymbol{B}^\top\boldsymbol{u}_i \\
&= \frac{1}{s}\sum_i^s\sum_j^s \left[(\boldsymbol{U}^\top\boldsymbol{B}\boldsymbol{L}^{-1})^2\right]_{ij}
\end{aligned}$$

where $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s)$ is the matrix whose columns are the s random vectors drawn as described in Equation 2.22. When $\mathsf{tr}(\boldsymbol{H})$ is required for new λ it is necessary to recalculate $\boldsymbol{L}$, however, the sparsity pattern for $\boldsymbol{L}$ depends on $\boldsymbol{B}$ and $\boldsymbol{D}$ and not on λ and it is therefore only necessary to recalculate the entries that are non-zero. In many scenarios where large B-spline bases are required, $\boldsymbol{L}$ can be very sparse, and using the `update.chol.spam` function in the `R` library `spam`, the updating of these elements can be performed very quickly. A further saving can be made by using the *same* $\boldsymbol{U}$ matrix for each evaluation of $\mathsf{tr}(\boldsymbol{H})$ that is required: consequently $\boldsymbol{U}^\top\boldsymbol{B}$ only needs to be calculated once at the outset, and every subsequent evaluation of $\mathsf{tr}(\boldsymbol{H})$ depends only on solving the system of equations $\boldsymbol{U}^\top\boldsymbol{B} = \boldsymbol{L}$ for each new $\boldsymbol{L}$. Using a fixed set of random vectors $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s)$ makes sense if s can be chosen to be large enough so that the bias associated with the surface $\tilde{\mathsf{tr}}(\boldsymbol{H}(\lambda_1, \ldots, \lambda_q))$ is acceptably small. In Section 5.2.2, some simulations show that relatively small (and computationally cheap) s, $s \geq 50$, result in estimates of $\mathsf{tr}(\boldsymbol{H})$, AICc and subsequent minimising values of λ that are very close to that which would be obtained using exact expressions.

2.3.3 P-splines using Bayesian analysis

2.3.3.1 A link between penalised LS and a Bayesian approach

A connection can be made between the penalised least-squares criterion in Equation 2.7 and a Bayesian model formulation. The stochastic analogue of the difference penalties applied to the parameter vector $\boldsymbol{\alpha}$ is a Gaussian kernel of the form

$$\begin{aligned} p(\boldsymbol{\alpha}) &\propto \exp\left(-\frac{1}{2\tau^2}\left[\sum_{i=1}^{k-1}(\alpha_i - \alpha_{i+1})^2\right]\right) \\ &= \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\alpha}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\alpha}\right) \end{aligned} \tag{2.23}$$

where smoothness is controlled by the variance parameter τ. For fixed τ and assuming a Gaussian likelihood with fixed variance, the posterior density resulting from Equation 2.23 is given by

$$\begin{aligned} p(\boldsymbol{\alpha}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha})^\top(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha}) - \frac{1}{2\tau^2}\boldsymbol{\alpha}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\alpha}\right) \end{aligned}$$

and therefore up to a constant of proportionality,

$$\log\left[p(\boldsymbol{\alpha}|\boldsymbol{y})\right] = -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha})^\top(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha}) - \frac{1}{2\tau^2}\boldsymbol{\alpha}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\alpha}. \tag{2.24}$$

Equation 2.24 is the negative of the penalised least squares criterion, and therefore the posterior mode of $p(\boldsymbol{\alpha}|\boldsymbol{y})$ for fixed $\lambda = \frac{\sigma^2}{\tau^2}$ is the solution to Equation 2.7 namely, at $\hat{\boldsymbol{\alpha}} = (\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{D}^\top \boldsymbol{D})^{-1}\boldsymbol{B}^\top \boldsymbol{y}$.

2.3.3.2 Bayesian inference in semiparametric regression

Building on this idea, P-spline models can be fitted in a Bayesian framework by applying appropriate random walk priors over neighbourhoods of spline basis coefficients. This area has been developed particularly as a result of the work described by Lang & Brezger (2004) which is implemented in the general purpose software package `BayesX` (Brezger *et al.* (2005)). Parameter estimation and inference is usually achieved using posterior sampling based on MCMC algorithms such as Gibbs sampling or Metropolis-Hastings schemes. Although the details differ from the penalised-likelihood methods described so far, most of the matrices involved in the respective calculations are the same - for example the prior precision matrix for parameters $\boldsymbol{\alpha}$ is the cross-product $\boldsymbol{D}^\top \boldsymbol{D}$ of the differencing matrix. As a result, the Bayesian approach is equally able to exploit matrix sparsity to perform faster calculations, which is a key feature of the `BayesX` software.

The Bayesian approach is particularly attractive as it permits the fitting of a very rich class of models with complex hierarchical structures, for example enabling models with variable or adaptive smoothness, such as that described in Lang *et al.* (2002). Since smoothness is controlled by the variance parameter σ^2 of a random walk prior distribution, treating σ^2 as random means that uncertainty about optimal smoothing is included in the estimates for all other parameters. This is an important property, because in cases where the observed data may be equally well described by a wide range of smoothness strengths, the plug-in estimate obtained from performance criterion minimisation is unable to represent this source of model uncertainty.

2.3.4 Smoothing as a mixed model

Many authors have realised the connection between spline smoothing and linear mixed effects models (LMMs) with general and comprehensive summaries available in Ruppert *et al.* (2003) and Wood (2006) for example. The connection is attractive, as LMMs are supported by an extensive literature and well developed software, rendering models so formulated very easy to fit. In order to fit a LMM we need to be able to write down the model in the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon} \tag{2.25}$$

where $\boldsymbol{a}$ corresponds to a set of *fixed effects* parameters, and $\boldsymbol{b}$ to a set of *random effects*. In addition, we require $\boldsymbol{b} \sim N(0, \boldsymbol{\Sigma})$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$. Considering the case of a univariate P-spline smooth

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{2.26}$$

the stochastic analogue of the quadratic penalty defined in Equation 2.7 is a Gaussian prior for the set of $\boldsymbol{\alpha}$, proportional to $\exp\left(-\frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\alpha}\right)$. However, the matrix $\boldsymbol{D}^\top \boldsymbol{D}$ is not of full rank, and therefore its inverse $(\boldsymbol{D}^\top \boldsymbol{D})^{-1}$ is not defined. This is most easily illustrated by observing that $\boldsymbol{\alpha}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\alpha}$ is an expression of squared differences of $\boldsymbol{\alpha}$, and as a result is unchanged by the addition of a constant to $\boldsymbol{\alpha}$, i.e. $\boldsymbol{\alpha}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{\alpha} = (\boldsymbol{\alpha} + c)^\top \boldsymbol{D}^\top \boldsymbol{D}(\boldsymbol{\alpha} + c)$. Consequently, it is necessary to perform a reparameterisation so that the covariance matrix is defined. The spectral decomposition of the penalty or prior precision matrix $\boldsymbol{D}^\top \boldsymbol{D} = \boldsymbol{V}\boldsymbol{U}\boldsymbol{V}$ where $\boldsymbol{V}$ is an orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{D}^\top \boldsymbol{D}$, and $\boldsymbol{U}$ is a diagonal matrix whose diagonal elements are the corresponding eigenvalues. Rewriting Equation 2.26 as $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}$ where $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{V}$ and $\boldsymbol{b} = \boldsymbol{V}\boldsymbol{\alpha}$, we have that $\boldsymbol{b} \sim N(\boldsymbol{0}, \frac{1}{\lambda}\boldsymbol{U}^{-1})$. This is because

$$\begin{aligned}
\mathsf{Var}(\boldsymbol{b}) &= \mathsf{Var}(\boldsymbol{V}\boldsymbol{\alpha}) \\
&= \lambda^{-1}\boldsymbol{V}^\top(\boldsymbol{D}^\top \boldsymbol{D})^{-1}\boldsymbol{V} \\
&= \lambda^{-1}\boldsymbol{V}(\boldsymbol{V}\boldsymbol{U}\boldsymbol{V})^{-1}\boldsymbol{V} \\
&= \lambda^{-1}\boldsymbol{U}^{-1}.
\end{aligned}$$

Since $\boldsymbol{D}^\top\boldsymbol{D}$ has at least one zero eigenvalue, at least one element of $\boldsymbol{b}$ is unpenalised and is therefore implicitly a fixed effect. If there are p zero eigenvalues, then we can further define $\boldsymbol{V} = [\boldsymbol{V}_1|\boldsymbol{V}_2]$ where $\boldsymbol{V}_1$ is the $n \times (n-p)$ matrix of columns of $\boldsymbol{V}$ corresponding to the non-zero eigenvalues of $\boldsymbol{D}^\top\boldsymbol{D}$ and $\boldsymbol{V}_2$ is the $n \times p$ matrix of columns of $\boldsymbol{V}$ corresponding to the zero eigenvalues. Correspondingly, we can set $\boldsymbol{b} = (\boldsymbol{b}_1, \boldsymbol{b}_2)$, and finally suppose that $\boldsymbol{U}_1$ is the square diagonal matrix that contains only the positive eigenvalues of $\boldsymbol{D}^\top\boldsymbol{D}$ as its diagonal elements. Then, we can write

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{B\alpha} + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{b}_1 + \boldsymbol{Z}\boldsymbol{b}_2 + \boldsymbol{\varepsilon} \\
\boldsymbol{X} &= \boldsymbol{B}\boldsymbol{V}_1 \\
\boldsymbol{Z} &= \boldsymbol{B}\boldsymbol{V}_2 \\
\boldsymbol{b}_2 &\sim N\left(0, \frac{1}{\lambda}\boldsymbol{U}_1^{-1}\right)
\end{aligned}$$

which is in the form described by Equation 2.25. It is then relatively straightforward to fit the model described in Equation 2.27 using software containing highly efficient routines developed for linear mixed models, such as the packages `nlme` (Pinheiro *et al.* (2013)) and `lme4` (Bates *et al.* (2013)) written in the `R` language. In this section, the reparameterisation required to fit using mixed model software was illustrated with the example of a single univariate smooth model. Although this is restrictive, the ideas generalise to a very wide class of semiparametric models that include tensor product terms and non-Gaussian error processes. For a detailed discussion of such extensions see Lee (2010).

2.4 Chapter summary

This chapter has described the principles underlying smoothing with P-splines. An overview of constructing appropriately penalised smooth curves, additive models and varying coefficient models has been outlined. A discussion of the most common ways in which parameter estimates are obtained for semiparametric models was provided, which is more complex than in general linear models due to the presence of smoothness controlling parameters. The Bayesian approach described in Section 2.3.3 potentially offers the greatest flexibility and scope to fit models with complex hierarchical structures which would be very difficult for the other approaches. The mixed model representation outlined in Section 2.3.4 is particularly appealing, because in common with the Bayesian approach, the smoothing parameters can be represented as variance components estimated as part of the model fitting, avoiding the use of plug-in estimation used in Section 2.3.1. Once the model is represented as a mixed model, well developed software can be used for model fitting. However, in both the Bayesian representation of P-splines and that of the mixed model, the main disadvantage can be computation. In the Bayesian case, MCMC is usually required for inference and is very costly if many tens of thousands of samples are required to achieve convergence to the target distribution. For the mixed model, a single reparameterisation based on the spectral decomposition of the penalty matrix is required to allow the semi-parametric model to be written in the appropriate form. This approach has complexity $O(n^3)$ which is prohibitive for some of the river network models described in Chapter 4, but also results in the storage of an $n \times n$ orthogonal matrix of eigenvectors which can also be difficult when n is large.

Since the goal of this thesis lies in constructing appropriate models for environmental data, and is not intended to provide a comparison of techniques for parameter estimation, we adopt the conceptually simplest approach of those described in this chapter which is to minimise some performance criterion as summarised in Section 2.3.1. In the chapters that follow, the thesis seeks to exploit features of matrix sparsity that render fast model fitting under this framework, which is a primary concern for large data sets. However it is noted that under some circumstances, in particular when the errors are serially correlated, the performance of smoothing parameter selection is not always reliable (Wang (1998), Currie & Durban (2002)), and when this is thought to be occurring some discussion has been provided along with suggestions for future improvement.

Chapter 3

Distributed lag models for rainfall and stream flow data

This chapter addresses the problem of modelling the lagged temporal dependence between high frequency time series of river flow rates and rain gauge data, each collected at single locations in space. In particular, the models developed in this chapter are applied to data that was collected on the River Dee in the North East of Scotland.

3.1 Introduction

Modelling river flow has long been of interest to environmental scientists. In particular, relating river flow to covariates such as hill slope gradient, ground canopy coverage, rainfall and snow-melt has been an important goal, often forming the basis of large catchment-scale models known as distributed models (Beven (1985)). These models commonly make use of rich data sets including high resolution satellite imaging to estimate land usage or snow coverage in discrete areal units. Such data are costly and scarce, and often all that is readily available are average river flows and meteorological data observed at point locations. While large scale distributed models are unavailable in such situations, flexible statistical models may be invaluable in providing simplified approximations to the system of study. Our interest lies in capturing changes in the temporal dependence of river flow on rainfall using approximations based on flexible regression methods, that are particularly useful when covariates that would have allowed physically-based models to be constructed are absent.

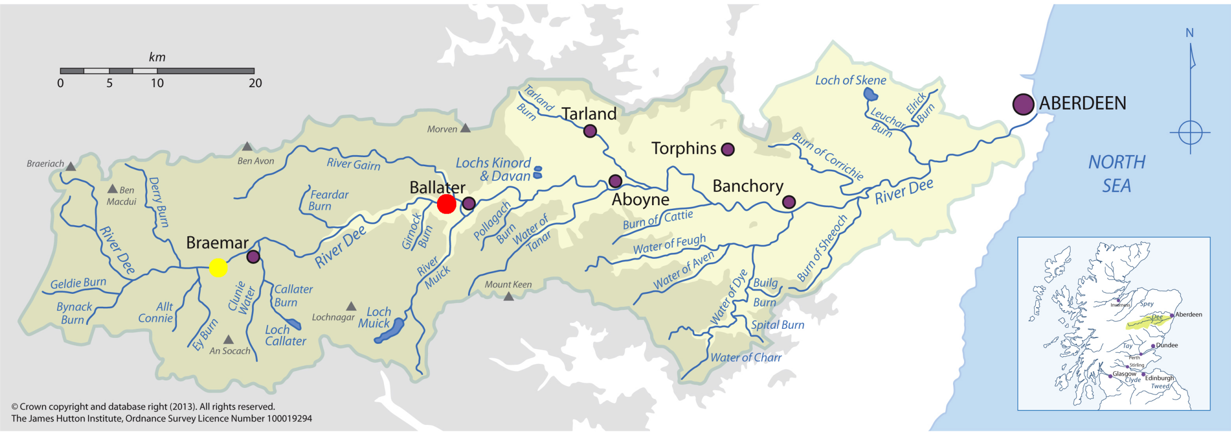
The rainfall-flow relationship is the ensemble of a number of interacting physical processes, most of which are unobserved. River flow is partly generated by a slow 'baseflow' process where infiltration of rainfall from surrounding land seeps out over long periods of time, in a manner which depends on the sponge-like water storage properties of surrounding ground strata (Shaw (2010)). Baseflow accounts for much of the river flow that persists during very dry summer months. In contrast, a faster responding 'runoff' process causes a more instantaneous response of flow to rainfall and accounts for much of the river flow during storms and prolonged rainy periods (Beven (1985)). Fast runoff arises when antecedent soil moisture increases to a level where rainfall can move more quickly near the soil surface without being absorbed, and can result in a more rapid increase in flow over periods of hours. Baseflow and runoff are for most catchments, the two most important drivers of variation in flow levels, with the influence of each determined by physical factors including soil and subsurface composition, surrounding land usage, evaporation and transpiration.

Accumulation and ablation of transient snow packs also form a key feature in the hydrology of many temperate and high altitude river systems, causing baseflow and runoff to decrease during winter periods and increase suddenly during warmer winter and early spring months. Snow deposition, as well as depth and density, are highly spatially heterogeneous and are less commonly and reliably measured than rainfall data, and in catchments prone to heavy snowfall and accumulation, modellers must be mindful of the increased uncertainty that this presents in rainfall-flow relationships during winter periods.

The dynamics underlying river flow generation are complex and are difficult to capture even in detailed physical models, and in addition, hydrologists are often interested in identifying when latent processes are most active, such as the influence of accumulation and melting of snow. Without detailed covariate data, we proceed by utilising flexible statistical methods with the aim of constructing a framework that allows us to approximate flow generating processes without attempting to identify the individual contributing components, that act over different timescales. The work described here is based on simple point-based rainfall data, but the wider modelling aim is to investigate methods by which complex environmental processes in both space and time can be approximated by semiparametric models.

3.1.1 Data

One particular aim is to develop a framework within which inference about the latent meteorological drivers of flow is possible. For example, it is speculated that recent climatic change has reduced the depth and duration of high altitude snow accumulation in some river networks, which changes the hydrology, and in turn the ecology of some river systems. Snow accumulation is not measured reliably and widely enough to directly detect such changes and so we hope to build towards an indirect method of detecting these types of patterns. Although the methods that are subsequently developed are applicable to any river system, we focus attention here to the River Dee in the North East of Scotland. The source of the River Dee is in the Cairngorm Mountains of Scotland, and it extends 141km before reaching the North Sea in Aberdeen with a total catchment area covering 2100km^2 (Baggaley *et al.* (2009)), see Figure 3.1. The River Dee is an important water resource, contributing around 50% of the total water supply to over 500,000 people for both drinking and industrial purposes, and is also of interest to environmental and conservation scientists with much of the river lying within reserved conservation areas (Langan *et al.* (1997)).



© Crown copyright and database right (2013). All rights reserved.
The James Hutton Institute, Ordnance Survey Licence Number 100019294

FIGURE 3.1: Geographical location and outline of the River Dee. Approximate locations of Polhollick flow monitoring site is identified by a red circle, location of the rain gauge at Braemar identified by a yellow circle (image provided by the James Hutton Institute).

The influence of different flow drivers is best illustrated with graphical summaries of hourly rainfall and flow data collected on the River Dee (that is later used in modelling): hourly rainfall accumulations (mm) collected to the nearest 0.1mm at Braemar and river discharge data (m^{3}s^{-1}) are collected from Polhollick, both located in the North East of Scotland and whose approximate locations are shown in Figure 3.1. Previous work on the River Dee showed that hourly flows are the highest resolution necessary to identify peak flow levels (Baggaley *et al.* (2009)).

The top left panel of Figure 3.2 displays a late winter period where little rainfall is observed but there is some flow variability remains, some small oscillation are also visible that occur on a daily cycle that could indicate the influence of melting snow. The top right panel displays a summer scenario with sparse rainfall, alongside low levels of river flow that appear to respond sluggishly to intermittent rain storms; this is typical of a period when baseflow dominates. The lower panels display a November period in which flow and rainfall are at high levels and a strong and immediate response to rainfall impulse is evident - a strong indication that runoff dominates during this period. The nature of the responsiveness is more easily seen in the bottom right panel of Figure 3.2 which shows a single week from its left-hand neighbour. It is evident from Figure 3.2 that the flow response to rainfall

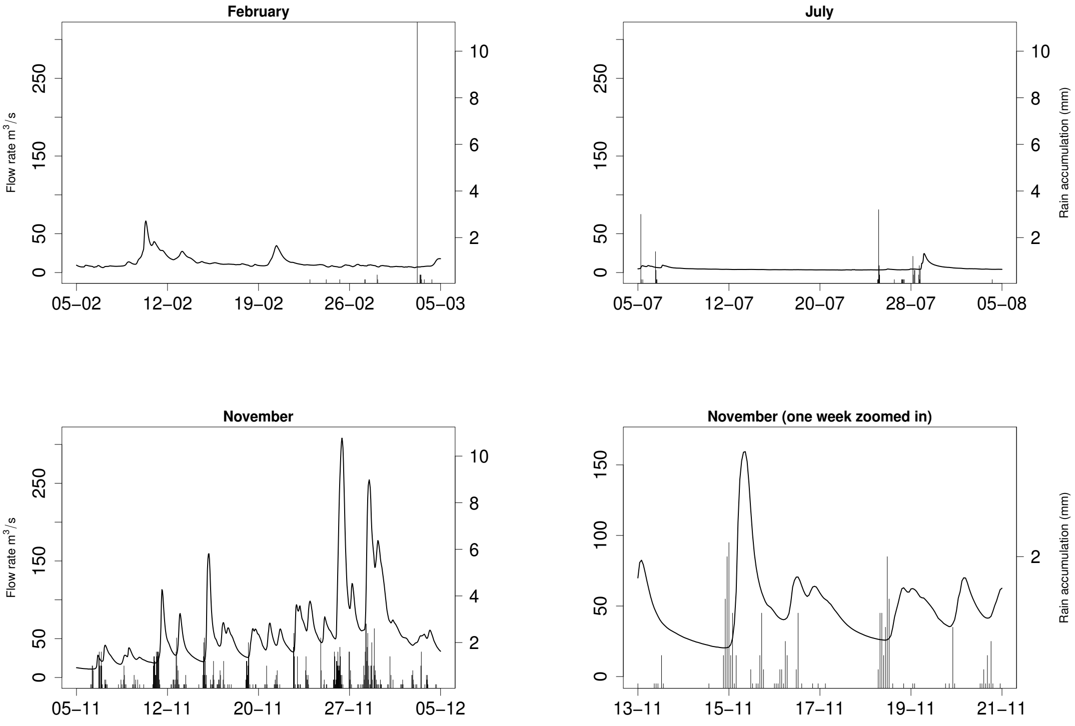


FIGURE 3.2: Rainfall and flow responses from the River Dee for four selected months in 2006. Continuous lines are flow rates (m^3s^{-1}); vertical line segments are hourly rainfall levels (mm).

varies throughout the year, in accordance with seasonal changes in rainfall patterns. It is also clear that the influence of a unit of rainfall is delayed and spread over time, caused by spatial separation (and intervening ground conditions) of rainfall across the catchment and flow gauges.

3.2 Modelling with distributed lag models

3.2.1 The distributed lag model in rainfall and flow models

Approaches to modelling the temporal dependence of flow on rainfall often assume that rainfall $r(t)$ and flow $f(t)$ are determined by the convolution

$$f(t) \quad = \quad \int_0^\infty h(s)r(t-s)ds$$

where t is a point in time, s is a lag variable and h is some response function. This is known as the *instantaneous unit hydrograph* (Nash (1957)), describing the impact over time that a unit of rainfall has on flow. Jakeman *et al.* (1990) suggested filtering rainfall data to first estimate 'effective runoff' before proceeding to estimate h. Direct approaches to modelling rainfall and flow include the nonlinear autoregressive moving average with exogeneous inputs (NARMAX) model of (Tabrizi *et al.* (1998)) which represents flow rates f as a degree l polynomial function g of rainfall and an error sequence:

$$\begin{aligned} f(t) \quad &= \quad g^l[f(t-1),\ldots,f(t-n_f),r(t),\ldots,r(t-n_r), \\ &\qquad \varepsilon(t-1),\ldots,\varepsilon(t-n_\varepsilon)]+\varepsilon \end{aligned}$$

Wong *et al.* (2007) propose 'functional coefficient modelling' in which river flow is modelled as flexible functions of rainfall and previous flow levels so that

$$\begin{aligned} f(t) \quad &= \quad g_1(r(t-1))+\ldots+g_l(r(t-l))+\beta_1(f(t-d))f(t-1) \\ &\qquad +\ldots+\beta_p(f(t-d))f(t-p)+\varepsilon(t) \end{aligned}$$

Where $\{g_1,\ldots,g_p,\beta_1,\ldots,\beta_p\}$ are unknown functions. It has been recognised that the form of the time dependence between flow and rainfall is an important model choice, and some authors have implemented polynomial constraints (Tabrizi *et al.* (1998)) on neighbouring lag variables or used local polynomial smoothers (Wong *et al.* (2007)). More generally, models of the form

$$E(y(t)) \;=\; \alpha + \beta_0 x(t) + \beta_1 x(t-1) + \ldots + \beta_l x(t-l)$$

where the impact of one time-dependent variable, $x(t)$, on another, $y(t)$, is spread over time, can be called a distributed lag model. We refer to the β_is as *lag coefficients*, and these can be considered as forming a discrete estimate, $\hat{h}$, of the underlying function h, which we term the *lag structure*. In many time series settings, multicollinearity emerges when a time-dependent variable is transformed to a set of l lagged covariates and care must be taken in estimation to avoid the highly variable estimates that result from an unconstrained regression. Typically some constraint is applied to the β_ls, a common choice being the Almon lag (Almon (1965)) in which the lag coefficients must lie on a polynomial of order p, $f^p(l)$, $l \in \{1, \ldots, L\}$, or the Koyck lag (Koyck (1954)) in which the lag coefficients are subject to a geometric decay constraint determined by the lag number.

3.2.2 The distributed lag model in air pollution

DLMs have seen much development (Zanobetti *et al.* (2000); Muggeo (2008); Welty *et al.* (2009); Gasparrini *et al.* (2010)) in the context of the delayed impact of urban air pollution on daily mortality counts. In this setting interest lies in specifying plausible shapes for DL curves and in particular the 'mortality displacement' effect, a phenomenon characterised by negative coefficients in the tail of the estimated lag structure, see Figure 3.3 for an illustration.

Zanobetti *et al.* (2000) propose a generalised model taking a penalised spline approach to modelling DL curves while Welty *et al.* (2009) discuss a Bayesian approach with penalties on parameters determined by carefully chosen priors. Others (Muggeo (2008); Gasparrini *et al.* (2010)) allow lag coefficients to change with temperature in addition to lying on a

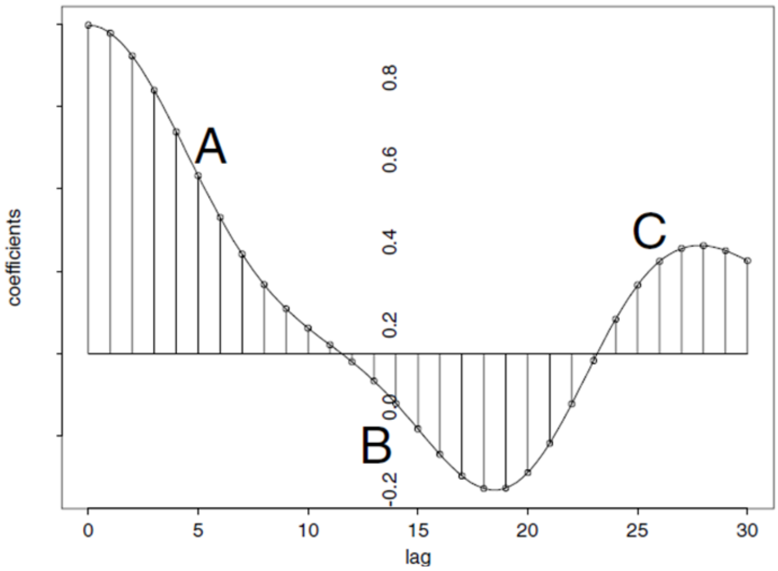


FIGURE 3.3: Example of a distributed lag curve showing the estimated impact of air pollution on mortality, where the y-axis scale shown in the middle of the plot represents the relative influence of each lagged days' air pollution on mortality. The 'mortality displacement' effect is characterised by the negative coefficients labelled by 'B' and the positive coefficients that are present at lower and higher lags, labelled by 'A' and 'C' respectively. Reproduced from Zanobetti *et al.* (2000).

low-rank smooth curve, so that a surface of lag coefficients results. Muggeo (2008) presents a framework where dependence of the DL curve on temperature is piecewise linear, with unknown breakpoints. Gasparrini *et al.* (2010) propose DL curves that lie on a bivariate surface parameterised by splines defined on lag index and temperature values.

Smoothing on model parameters rather than data, as is the case with DL curve estimation, is a situation where appropriate smoothness levels are not easily judged by visual inspection of the fitted model. For this reason, a P-splines approach (Eilers & Marx (1996)) as summarised in Section 2.2.2 is convenient and is adopted here, where a rich set of uniformly spaced B-spline basis functions, together with a roughness penalty on neighbouring basis functions yields a fitted function with the appropriate level of smoothness. The strength of the roughness penalty is most easily selected by minimising an information criterion such as AICc (Hurvich *et al.* (2002), Section 2.3). We proceed to construct a DL model for rainfall and river flow rates, relaxing the assumption of a fixed lag structure; the use of P-splines are found to facilitate specification of flexible models while maintaining a high level of computational efficiency by taking advantage of sparse model objects.

3.3 Time varying DLM

We set up a model for flow at time t, $f(t)$, in terms of a weighted sum of preceding upstream rainfall $(r(t-1), \dots, r(t-L))$ with weights $(\beta_1, \dots, \beta_L)$, subject to the constraint that the β_l lie on a spline constructed from a set of I degree 3 basis functions $\{B_1(\cdot), \dots, B_I(\cdot)\}$. The form of the model is

$$\begin{aligned} f(t) &= \alpha + \sum_{l=1}^{L} \beta_l r(t-l) + \epsilon(t) \quad \text{where} \quad \beta_l = \sum_{i=1}^{I} a_i B_i(l) \\ &= \alpha + \sum_{l=1}^{L} \sum_{i=1}^{I} a_i B_i(l) r(t-l) + \epsilon(t) \end{aligned}$$

where α is an intercept term and $\epsilon(t)$ is an IID error process. We further allow the relationship between each rainfall lag variable $r(t-l)$ and $f(t)$ to change smoothly with time, the form of which depends on a further set of J B-spline basis functions $\{B_1(\cdot), \dots, B_J(\cdot)\}$ so that $a_i = \sum_{j=1}^{J} b_{ij} B_j(t)$. This gives the representation

$$f(t) = \alpha + \sum_{l=1}^{L} \sum_{i=1}^{I} \sum_{j=1}^{J} b_{ij} B_j(t) B_i(l) r(t-l) + \epsilon(t).$$

In matrix notation,

$$\begin{aligned} \boldsymbol{f} &= \boldsymbol{Z\theta} + \boldsymbol{\epsilon} = [\mathbf{1}, \boldsymbol{X}]\, \boldsymbol{\theta} + \boldsymbol{\epsilon} = (f(t_1), \dots, f(t_n))^\top \\ \boldsymbol{X} &= \boldsymbol{B}_J \Box \boldsymbol{R}\boldsymbol{B}_I = (\boldsymbol{B}_J \otimes \mathbf{1}'_I) \odot (\mathbf{1}'_J \otimes \boldsymbol{R}\boldsymbol{B}_I) \\ \boldsymbol{\theta} &= (\alpha, \boldsymbol{b}) = (\alpha, b_{11}, b_{21}, \dots, b_{I1}, \dots, b_{1J}, b_{2J}, \dots, b_{IJ})^\top \end{aligned}$$

Where the i^{th} row of $\boldsymbol{B}_{\text{J}}$ is $\{B_1(t_i), \ldots, B_J(t_i)\}$, i^{th} row of $\boldsymbol{R}$ is $(r(t_i-1), \ldots, r(t_i-L))$ and $\Box$ is the Box product as used by Eilers *et al.* (2006). The intercept included in the specification represents flow rates after rainfall has not been observed for over L lags.

We wish to control the level of smoothness in the fitted coefficients in two ways: by how each rainfall lag variable $r(t-l)$ influences $f(t)$ as t changes; and by how different the influence of $r(t-l)$ and $r(t-l+1)$ is allowed to be at any time t. These constraints will be represented by two different roughness penalties. The first term, $\lambda_1 \boldsymbol{D}_1^\top \boldsymbol{D}_1$, penalises the 'wiggliness' of the β_is through time, and so $\boldsymbol{D}_1$ is a block matrix where each block is a quadratic difference matrix $\boldsymbol{P}_J$ with J columns so that

$$\boldsymbol{P}_J \boldsymbol{b} \;=\; \sum_{i=1}^{I} \sum_{j=1}^{J-2} \left(b_{i,j+2} - 2b_{i,j+1} + b_{i,j}\right)^2$$

and, in Kronecker notation, $\boldsymbol{D}_1 = \boldsymbol{P}_J \otimes \boldsymbol{I}_I$. The second penalty term, $\lambda_2 \boldsymbol{D}_2^\top \boldsymbol{D}_2$, controls differences between β_l and β_{l+1}, $l \in \{1, \ldots, L-1\}$ at any time t and this is achieved similarly by penalising differences between $b_{i,j}$, $b_{i,j+1}$ and $b_{i,j+2}$ for $i \in \{1, \ldots, I\}$ and $j \in \{1, \ldots, J-2\}$, so that $\boldsymbol{D}_2 = \boldsymbol{I}_J \otimes \boldsymbol{P}_I$. Combining the two penalties and for fixed values of λ_1 and λ_2, the parameter estimates $\hat{\boldsymbol{\theta}}$ are obtained by penalised least squares by

$$\hat{\boldsymbol{\theta}} = \boldsymbol{S}\boldsymbol{f} \;=\; \left(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda_1 \boldsymbol{D}_1^\top \boldsymbol{D}_1 + \lambda_2 \boldsymbol{D}_2^\top \boldsymbol{D}_2\right)^{-1} \boldsymbol{Z}^\top \boldsymbol{f}$$

with standard errors given by $\mathsf{s.e.}(\hat{\boldsymbol{\theta}}) = \sqrt{\mathsf{diag}(\boldsymbol{H}^\top \boldsymbol{H})}$ where $\boldsymbol{H} = \boldsymbol{Z}\boldsymbol{S}$

3.4 General specification

More generally, DLMs can be specified so that the lag structure varies with any set of covariates. For example, if the β_is are required to change smoothly and non-linearly with one additional covariate, a model matrix with one additional Box product, $\Box$, must be constructed.

Let $x_1(t), \ldots, x_r(t)$ be r n-length time-dependent covariates and $\boldsymbol{J}^1, \ldots, \boldsymbol{J}^r$ be marginal basis matrices defined on the $x_i()$ so that the m^{th} row of $\boldsymbol{J}^i$ is

$$[B_1(x_i(m)), \ldots, B_{J_i}(x_i(m))],$$

and J_i is the size of the basis set defined on the ith covariate. A general multidimensional DLM model matrix is defined as

$$\boldsymbol{X} = \boldsymbol{J}_1 \Box \boldsymbol{J}_2 \Box \ldots \Box \boldsymbol{J}_r \Box \boldsymbol{RB}_{\mathrm{I}}$$

where $\boldsymbol{RB}_{\mathrm{I}}$ is defined as in Section 3.3 and the corresponding $\boldsymbol{\theta}$ is a vector of spline coefficients and an intercept with length $1 + I \prod_{i=1}^{r} J_i$. Since the model sets up a smooth in $r+1$ dimensions (r for coefficient bases and 1 for lag structure basis) we require $r+1$ penalty terms, these can be expressed as a sequence of Kronecker products with identity matrices

$$\boldsymbol{D}_i = \left[\bigotimes_{j<i} \boldsymbol{I}_j\right] \otimes \boldsymbol{P}_i \otimes \left[\bigotimes_{j>i} \boldsymbol{I}_j\right]$$

where $\bigotimes_{j<i} \boldsymbol{I}_j = \boldsymbol{I}_1 \otimes \ldots \otimes \boldsymbol{I}_{i-1}$, and each $\boldsymbol{D}_i$ corresponds to a roughness penalty on the ith dimension of the tensor smooth defined by $\boldsymbol{X}$.

3.4.1 Computational aspects

The models described in Sections 3.3 and 3.4 potentially require the storage and manipulation of $n \times (1 + IJ)$ and $(1 + IJ) \times (1 + IJ)$ matrices which can be expensive. Currie *et al.* (2006) describe how, if model matrices arising in tensor-product type models can be factorised so that $\boldsymbol{X} = \boldsymbol{X}_1 \otimes \boldsymbol{X}_2$, then much of the computational and storage overhead can be bypassed. In the present case, $\boldsymbol{X}$ cannot be so factorised, due to the row-wise tensor product matrix structures. However, significant savings can be made by exploiting the sparseness properties of many of the model objects. Since a set of penalised B-splines is used, all basis matrices are sparse, and additionally their rows are defined on consecutive sequences of integers (time and lag indices here) and are therefore banded. Therefore $\boldsymbol{R}\boldsymbol{B}_I$ is a banded sparse matrix, and hence $\boldsymbol{X} = \boldsymbol{B}_J \Box \boldsymbol{R}\boldsymbol{B}_I$ is banded, and in turn, $\boldsymbol{Z}^\top \boldsymbol{Z}$ and $\left(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda_1 \boldsymbol{D}_1^\top \boldsymbol{D}_1 + \lambda_2 \boldsymbol{D}_2^\top \boldsymbol{D}_2\right)$ are banded and sparse. Hence we are required only to manipulate a banded sparse matrix object which is faster than the general sparse case, and dramatically reduces storage requirements. The sparseness properties are further enhanced by the zero-inflated distribution of hourly rainfall data. In `R` sparse matrix algebra is easily performed using the `Matrix` package (Bates & Maechler (2013)). It is also noted that recent work by Lee & Durbán (2011) develop the idea of 'nested' B-spline basis in order to reduce the computational complexity in tensor product smooths, and it should be a key component in further work to investigate its application in the present context.

In order to choose an appropriate 'rich' basis size when applying models of Sections 3.3 and 3.4, a short iterative process is required to determine a minimal basis size such that on application of different strength penalties, a broad range of smoothing strengths result, and in particular, when $\boldsymbol{\lambda} = \boldsymbol{0}$ the model overfits the data. Having selected a rich basis, the optimal penalties λ_i are found by searching a logarithmic grid for the values that minimise AICc.

3.5 Time varying DLM on simulated data

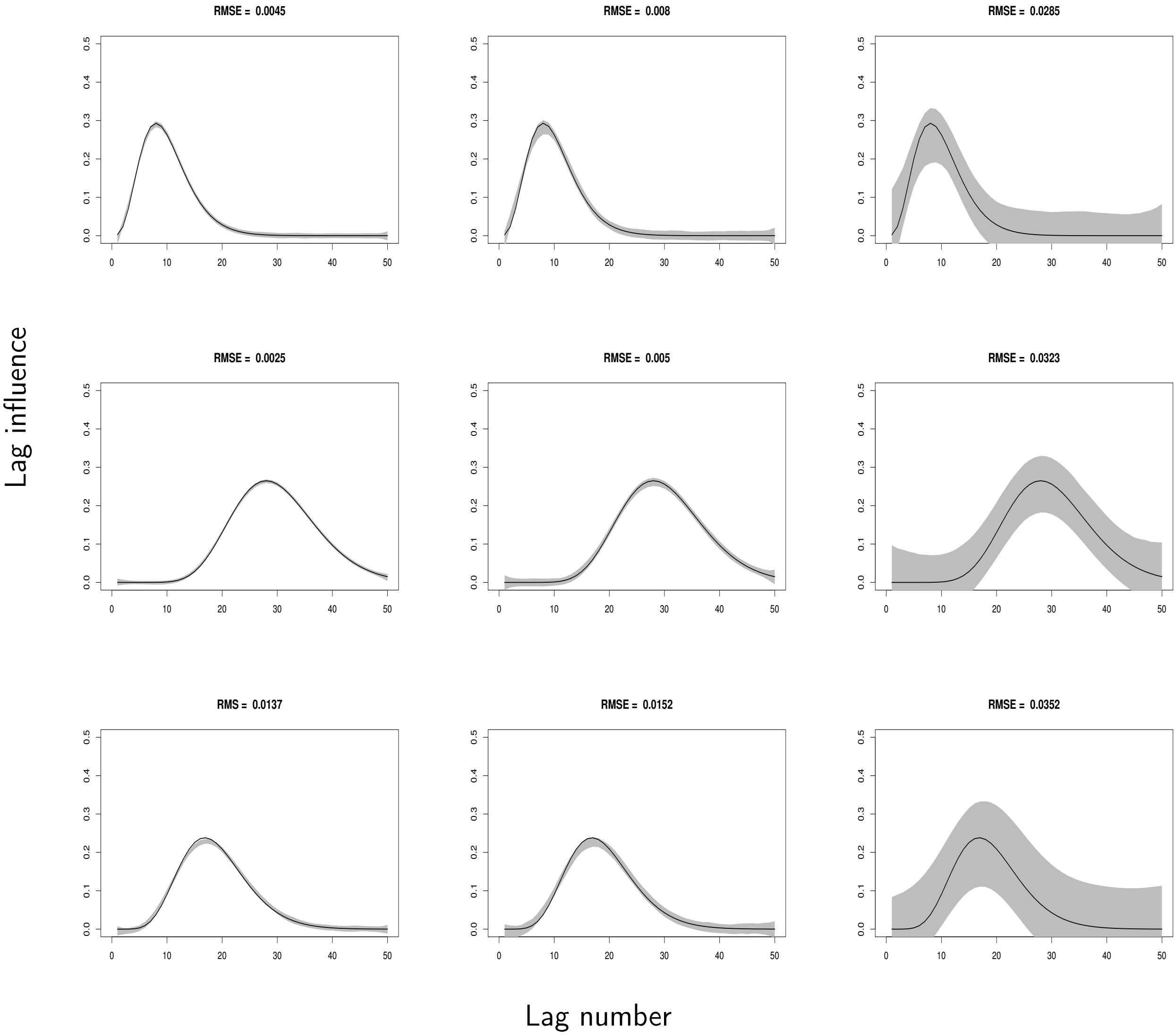
Before applying the models in Section 3.3 to river flow data, we first apply the methods to simulated data to examine how well the model captures different lag structures in the presence of different error processes, and to test the performance of AICc in selecting an appropriate level of smoothness for the fitted DL curves. Three time series of flow data were constructed by convoluting the 2006 Braemar hourly rainfall data described in Section 3.1.1, with three different DL curves defined up to 50 lags so that,

$$f_{ij}(t) \quad = \quad \sum_{l=1}^{L} \beta_{li} r(t-l) + \epsilon_j$$

where β_{L1} and β_{L2} are time invariant and are based on Gamma distribution functions (shown in Figure 3.4), and β_{L3} varies smoothly over time between the shapes of β_{L1} and β_{L2}. Furthermore, $\epsilon_1 \sim N(0, 0.04)$, $\epsilon_2 \sim N(0, 0.16)$ and ϵ_3 follows a normal random walk process through time with $\sigma = 0.01$ so that nine possible scenarios result. We then simulated 200 times from each of the nine scenarios and fitted the time varying model of Section 3.3, with $L = 50$ lags and moderate basis sizes of $I = J = 20$. Since our main interest is in recovering the underlying DL structures, the fitted DL curves can be compared against the true curve using 95% pointwise simulation envelopes and the root mean squared error (RMSE), both shown in Figure 3.4.

From Figure 3.4 it can be seen in all cases that the DLM recovers the underlying lag structure well, with the 95% envelope functions lying close to the true curve. Some detail is lost in some of the fitted functions at the 'peak' of the estimates, particularly where the peak is pointed, which is likely to be a result of the choice of basis size being slightly too small. The random walk process is included as an example of a strongly correlated error process that environmental data often exhibit, and while the model performs less favourably than with independent errors, it still succeeds in recovering the shape of the underlying function.

FIGURE 3.4: True DL curves with pointwise 95% simulation envelopes plotted in grey. Each row corresponds to simulations under differing DL shapes shown in black, and each column to differing error structures ϵ_1, ϵ_2 and ϵ_3, respectively. The third row shows a mid-July snapshot of the true DL surface and 95% simulation envelopes under the time varying DL scenario. Mean RMSE values are quoted at the top of each panel.



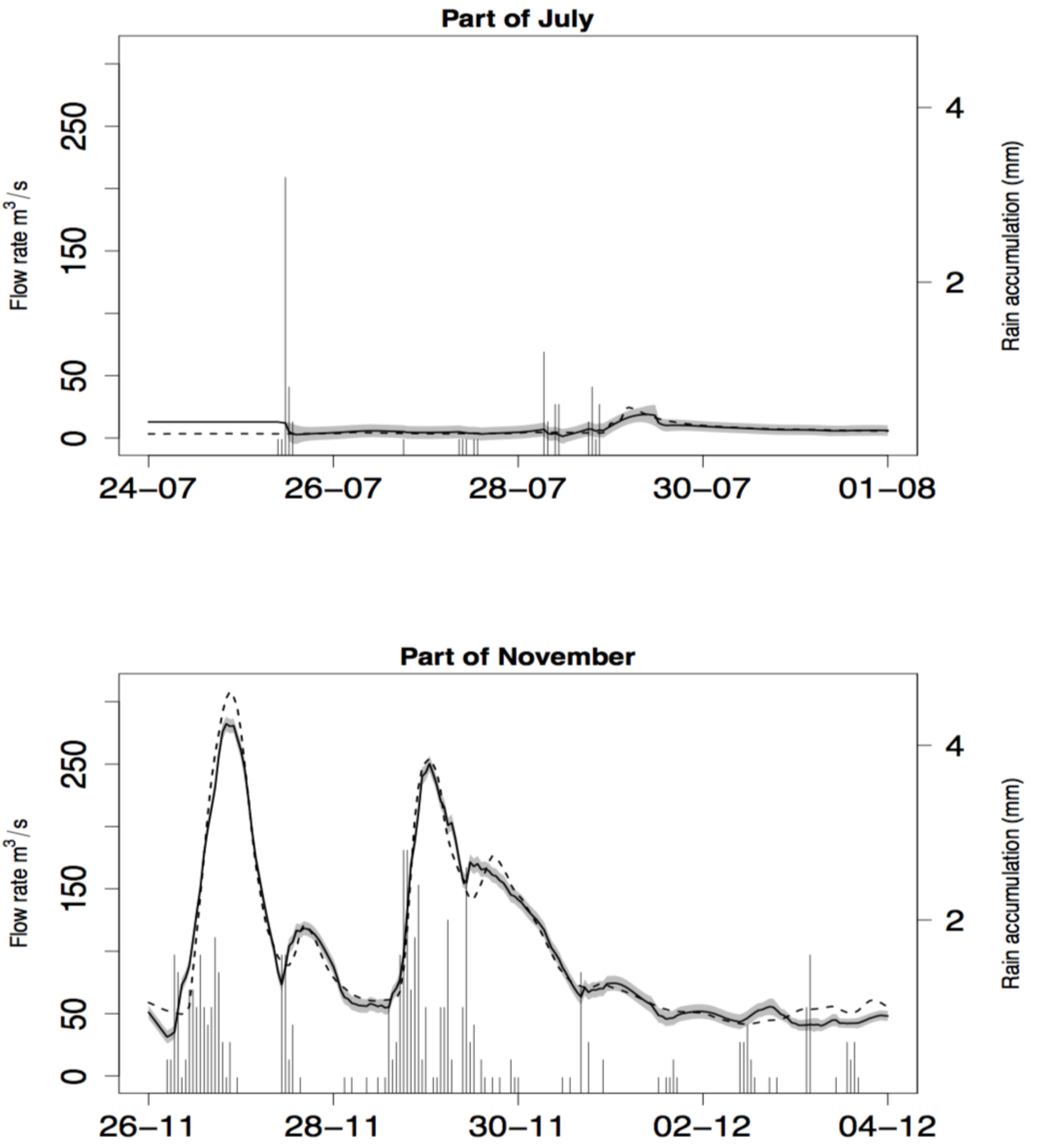
3.6 Application to the River Dee

The River Dee data described in Section 3.1.1 is now considered with the model of Section 3.3. High resolution data is relatively scarce and what follows has been fitted to the 8861 average hourly flows and rainfall for the year 2006 only; ideally several years data would be considered and adjustment made to account for seasonal and inter-annual variation.

A rich basis was first chosen, with $I = 50$ and $J = 100$ selected so that, without penalty terms (i.e.. $\lambda_1 = \lambda_2 = 0$), the model overfits the data. A large number ($L = 100$) of lags were chosen and the optimal λ_1 and λ_2 were found by the method described in Section 3.4.1. The fit of the model can be examined by inspecting plots of observed and fitted values during different parts of 2006, shown in Figure 3.5.

In the upper plot of Figure 3.5 we see a period of very low rainfall and low flow; the model performs poorly where rainfall has not been observed for more than $L = 100$ hours, with the intercept $\alpha = 12.7$ left to account for the remaining recession of flow levels. By contrast, the lower plot corresponds to a wet period and the model fits well, despite the extreme levels reached in river flow.

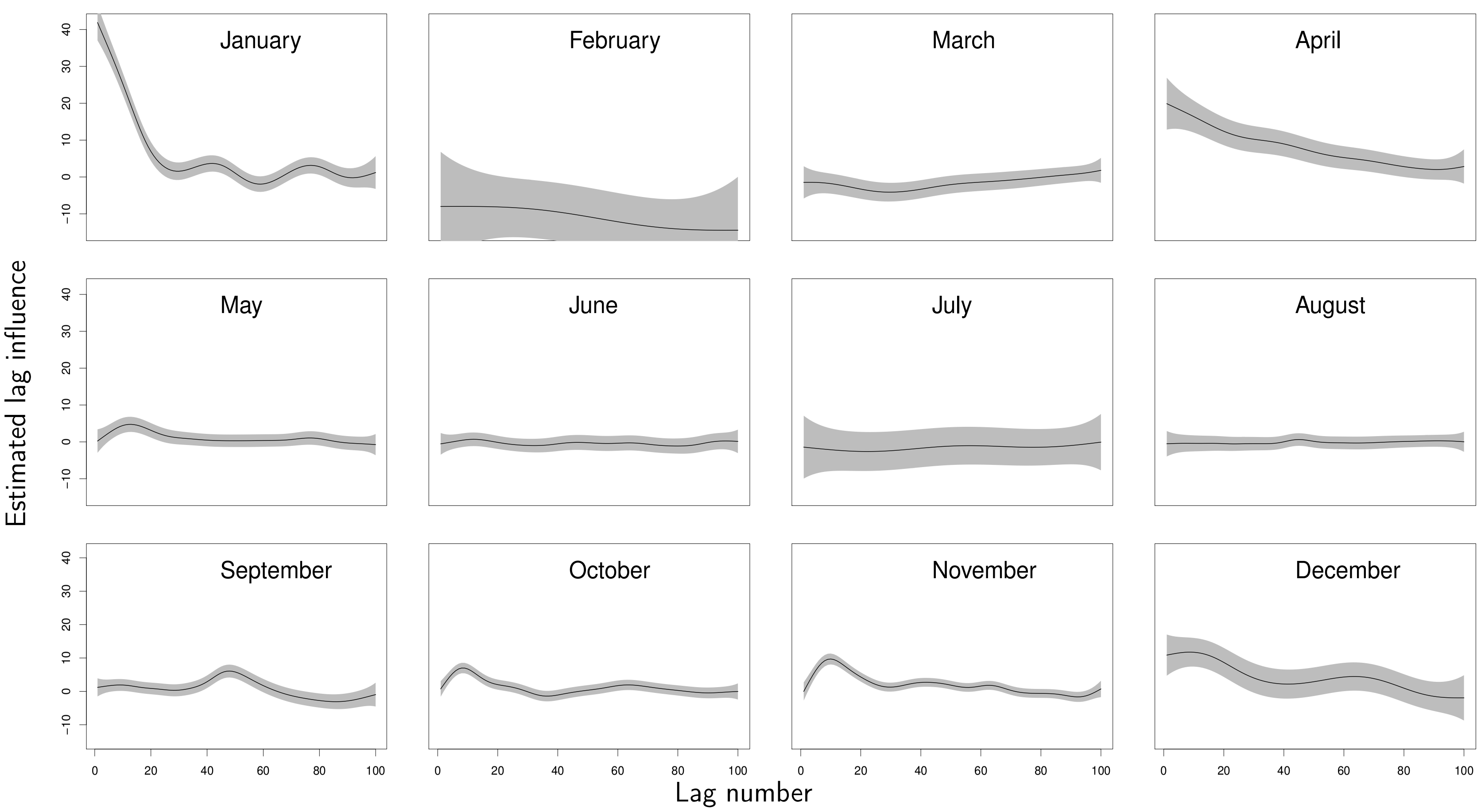
Figure 3.5: Fitted flows alongside flows observed on the River Dee: dashed lines represent observed flow levels, solid lines represent fitted flows and vertical line segments are hourly precipitation and grey shaded areas are 95% confidence regions.



We can also examine the fitted DL curves which are shown in Figure 3.6. There is clear evidence of differences in the estimated lag structures throughout the year: in summer months, lag structures are mostly flat, indicating a slow and delayed response, and during wet autumn months are sharply peaked and very tightly contained within their 95% confidence intervals. A strong and consistent responsiveness in flow levels when rainfall has been heavy or prolonged is visible, for example during November 2006, with a clear peak in lagged influence that most likely indicates the predominance of fast-moving runoff. At other times less consistent or interpretable response functions are estimated; for example in January 2006, shown in Figure 3.6, responses appear very high, suggesting extremely high influence of rainfall up to the most distant lags which is unlikely to be the case as snow is the most likely flow driver at this time. During periods in which freezing temperatures are common rainfall data can be unreliable as snow and ice accumulate in the measuring device until they melt, often much later. We therefore interpret the estimates for January and winter months with caution, and note that they indicate the presence of some effect yet to be accounted for.

In the final weeks of observation, a sustained period of heavy rain and an overall increase in flow with progressively more extreme peaks is observed. The lag structures within this period gradually increase in height, particularly in the ‘peak’ of influence at approximately a 10 hour lag. Such changes in lag structure are consistent with an increase in ground saturation causing a higher proportion of rainfall to convert to runoff, with flow levels subsequently appearing to be highly sensitive to new rainfall. It is therefore desirable to construct a model that attempts to account for temporal variation in lag structures during wet periods using information on long-term ground wetness.

FIGURE 3.6: Estimated mean DL curves with pointwise 95% confidence regions estimated from River Dee rainfall and flow data plotted at monthly ‘snapshots’; x-axes correspond to the lag numbers (between 1 and 100) of points on the response function



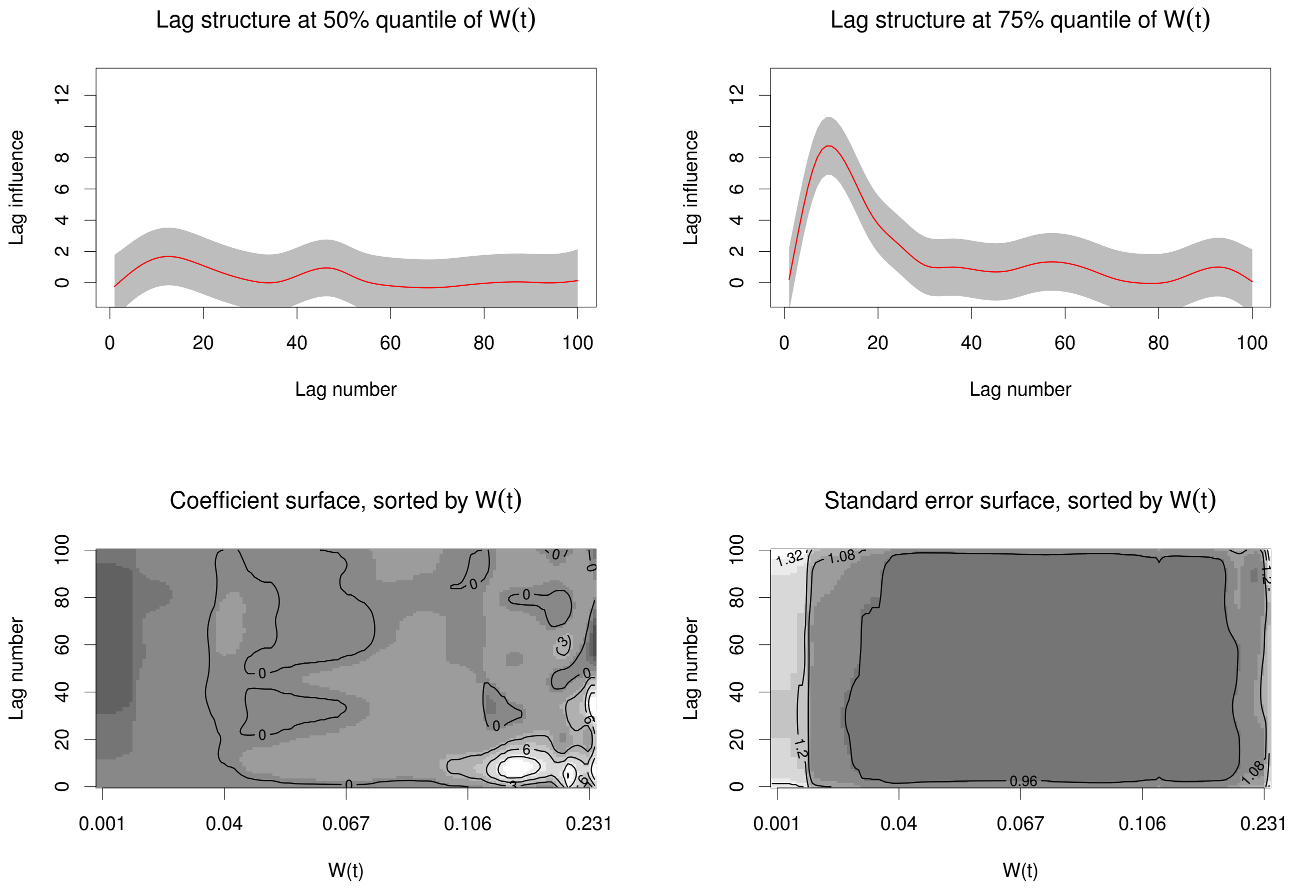
3.6.1 A 'ground-wetness' varying DLM on River Dee data

We now consider introducing a covariate representing unobserved antecedent ground wetness, for which a 30 day moving-window mean of observed hourly rainfall with exponentially decaying weights is constructed as a proxy, which we now call $W(t)$. The choice of 30 days represents the belief that variation in rainfall response is driven by a larger ensemble of precipitation outwith the largest lag of the DLM of Section 3.6, particularly during prolonged wet periods. A number of window widths were tried and the resulting model was not found to be sensitive to small changes. An alternative approach might make use of catchment-specific water residence time distributions, if known, to inform the weights and window widths in construction of such a proxy. In what follows, $W(t)$ is assumed to be the only modifying factor of the lag structure and is intended to account for much of the temporal variation in the β_i observed in Section 3.3 during wet periods. In similar notation to Section 3.3 the model is specified as

$$f(t) \quad = \quad \alpha + \sum_{l=1}^{L} \sum_{j=1}^{J} \sum_{m=1}^{M} c_{jm} B_m(W(t)) B_j(l) r(t-l) + \epsilon(t).$$

Estimation proceeds as in Section 3.3, where the coefficient vector $\boldsymbol{\theta} = (\alpha, c_{11}, \ldots\ldots, c_{JM})$. The model parameters were $L = 100$, $M = 50$, $J = 100$ again, representing an overfitted model when the penalty vector $\boldsymbol{\lambda} = \mathbf{0}$. It was found when selecting optimal $\boldsymbol{\lambda}$ with AICc that there was a tendency to choose undersmooth estimates for variation in the $W(t)$ dimension, hence it was decided to use the 'optimal' estimate as a lower bound on $\boldsymbol{\lambda}$. By visual inspection of the resulting parameter surface, a penalty was selected that was at least as strong as that chosen by minimising AICc in order to avoid overfitting. The intercept term was similar to that in Section 3.3 with an estimate of $\alpha = 12.2$. Interest lies in how the β_i respond to different levels of $W(t)$. The top panels of Figure 3.7 illustrate the changes in lag structure at different quantiles of the distribution of $W(t)$; at higher levels of $W(t)$ more peaked and overall larger lag structures are visible, particularly at the highest levels of $W(t)$. In the bottom panels of Figure 3.7, images illustrating the changes in lag structure across the range of $W(t)$, and through time, are given. An important feature here is the shift in peak influence from later lags to earlier lags which is visible as $W(t)$ increases. It is also notable that less dominant peaks later in the lag structure appear at the lowest and highest levels of $W(t)$.

FIGURE 3.7: Top: Lag structure at 50% and 75% quantiles of $W(t)$ indicating different rainfall response scenarios for different wetness conditions; the with grey band represents pointwise 95% confidence intervals around the fitted DL function shown by the solid red line. Bottom left: contour plot of lag structure estimates as they vary with increasing $W(t)$, corresponding surface of standard errors are shown in the bottom right panel.



3.7 Discussion

We have proposed flexible and computationally attractive DLMs with roughness penalties that are successful in capturing the dependence between river flow and a sequence of preceding rainfall measurements. In Section 3.6, a complex and time varying relationship between river flow and rainfall was identified, with Section 3.6.1 uncovering evidence that some of this variability arises through a complex interaction between slowly changing ground wetness and the time when rain falls. It was also found that the degree and location of peak influence in the lag structure can change dramatically, and that these were persistent features under the use of different strength penalties.

3.7.1 Residual autocorrelation

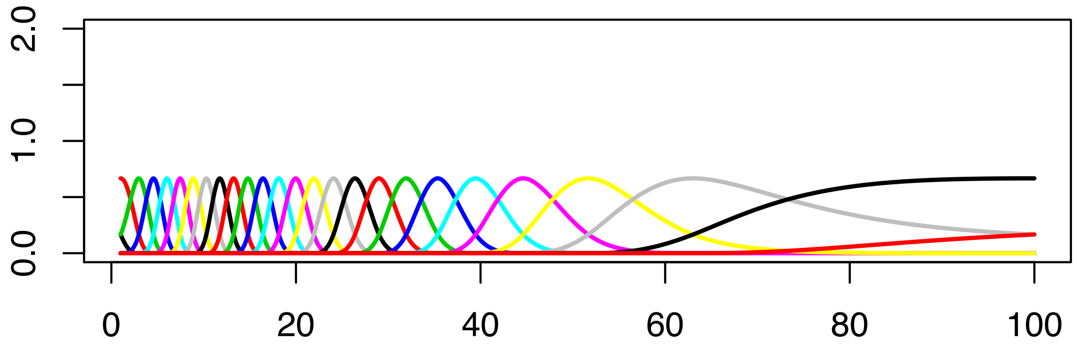
Figure 3.5 shows that fitted flow levels exhibit a weaving pattern around the observed data, which would indicate the presence of a serially correlated error sequence, and some simple plots of the autocorrelation function confirmed that this was indeed the case. Although the simulations in Section 3.6 show that the underlying DL curves could be recovered effectively under strongly correlated errors, the smoothing parameters are likely to have been underestimated as noted by Wang (1998) and the associated standard errors are likely also to be underestimated as a result of such autocorrelation. A possible adjustment might involve fitting a model to the residuals, and adjusting the hat matrix $\boldsymbol{H}$ by the estimated residuals variance matrix as Bowman *et al.* (2009) did in the context of spatio-temporal modelling. If required, approximate hypothesis tests could then be constructed, as in Bowman *et al.* (2009), to assess and compare aspects of competing models, in particular to determine whether a time varying model is required over a fixed lag structure.

It is also likely that some of the residual autocorrelation is induced by spatio-temporal heterogeneity in the rainfall process that can not be well represented by point-location rainfall data. Bias in lag structures can arise when the underlying weather is dynamic, for example, when rain storms occur near to the rain gauge but are not recorded. It is therefore the intention in future research to represent river flow as *both* a temporal and spatial ensemble of rainfall, using data containing information on the spatial position of rainfall events.

3.7.2 Model misspecification

The issue of biased estimation may arise if the temporal extent and influence of baseflow is not adequately characterised. In the current context an intercept term is all that accounts for the decay in flow rates after rainfall has been absent for L or more hours. More sophisticated approaches might treat L as unknown, or assume a very large L in order to fully incorporate baseflow response into the DL specification. Both approaches may require more structured penalties, for example imposing stronger penalties at higher lags than shorter ones so that models in which $\beta_i \to 0$ as $i \to \infty$ are preferred; see Muggeo (2008) for such an application using health data that incorporates an additional ridge penalty. More complex still, an *adaptive* smooth could be used in which the difference penalty is allowed to vary across the space spanned by the basis, either in some prespecified parametric way or in a way described by a further non-parametric surface. It is expected that rainfall occurring in the more recent past has a far higher influence on current river flow than older rainfall, and that the decay in lagged influence should be rapid. Therefore, the use of P-splines with a uniformly spaced B-spline basis could be an inefficient way to represent this curve, because allowing sufficient detail at short lags encourages too much flexibility at higher lags and requires a large number of basis functions. A direct approach to solving this might be to use a monotonic I-spline basis that would ensure monotonic decreasing behaviour after a certain lag, or simply increasing the strength of the penalty function as the lag number increases. However, a more elegant approach might be to relax the uniformity of the spacing of the B-spline basis, and allow a higher density spacing at short lags, as illustrated by Figure 3.8 where the knot spacing is determined by the quantiles of a Gamma distribution function chosen with a mode near shorter lags. This approach could significantly reduce the number of functions required to represent a lag structure in which the ‘tail’ of the function is very smooth; however, the interpretation of the difference penalty on such a basis can no longer be interpreted as a measure of global roughness of the function.

FIGURE 3.8: Example of unevenly spaced B-spline basis, where knots are placed at quantile of a Gamma distribution function with a mode at shorter lags and the x-axis represents the lag index.



A related issue was the tendency to encounter undersmoothing as a result of choosing smoothing parameters by grid search over AICc. As a practical measure, in Section 3.6.1 the selected parameters were treated as lower bounds on smoothing to maintain a realistic level of wiggliness. Some of this undersmoothing may result partially from the type of misspecification described above; however, an important component of further work will be to ensure that automatic smoothness selection can achieve reliably smooth estimates, particularly when the underlying AICc profile is very flat and the optimal selection based on a grid search may be very sensitive to small changes in the data. One way in which to deal with this issue would be to take a fully Bayesian approach to modelling, based on the ideas described in Section 2.3.3, using random walk priors on pairs of neighbouring coefficients to penalise the smooth distributed lag curve. An attractive property of this approach is that when there is high uncertainty about how smooth the fitted curve should be, this will be reflected by a large amount of spread in the corresponding posterior distribution for the random walk variance, and the posterior distributions for the spline coefficients incorporate this additional level of uncertainty.

3.7.3 Towards a spatial distributed lag model

As already suggested in Section 3.7.1, rainfall data collected at a single location in space limits the ability of the model to fit the flow rates observed, because the rain gauge data contains little information about the localised spatial structure of precipitation. For example, under scenarios in which localised rain storms occur nearby to, but not at the rain gauge site, the DLM fit to river flow levels will likely be very poor. Therefore, the most important extension that could be made to modelling river flow with distributed lag models would incorporate spatial rainfall data that can account for the influence of precipitation occurring at all locations in the upstream area that drains to the flow monitoring location. Given this spatial data, one possibility would be to extend the single-input distributed lag model to a set of DL curves indexed in space, each representing the influence and drainage speeds of a single sub region. This idea is more complex and is developed in more detail in Section 6.2.

Chapter 4

Flexible regression for river networks

This chapter attempts to address the problem of modelling data that arise on a stream network. The unique features of river networks are described in detail, alongside the difficulties these present for traditional spatial modelling. An approach to building spatial models for river network data using flexible regression techniques is explored, and is found to be particularly useful in order to characterise the complex spatial and temporal dependencies that are often exhibited by environmental data. The models developed are fitted to data collected on the River Tweed, a large river network located in the South East of Scotland.

4.1 Introduction

Large data sets arising on stream and river networks are increasingly common because of widespread environmental monitoring programs, with attributes such as dissolved pollutant concentrations, stream temperature and measures of biodiversity (i.e. counts of birds and insects) collected along the branches of many rivers. These data are often used to address vital questions pertaining to the effects of climate change on habitat and species distributions, as well as other anthropogenic impacts on in-stream habitat and aquatic pollution. It is therefore critical that appropriate statistical methods, which adequately account for the different sources of variability are used to make valid inferences from stream network data.

It is typical to find evidence of residual spatial autocorrelation in spatial modelling settings, where the dependent variable is spatially indexed, and spatial statistical models for stream networks are no different. This residual variation is usually the result of some unobserved confounding variables that are correlated in space and if left unaccounted for, can cause the model variance and parameters to be unreliably estimated, and partly as a result of this, spatial statistical modelling is a key area of statistical research for which a large literature exists. Most of the established literature assumes that the spatial region of interest is a simple subset of $\mathbb{R}^2$ and that Euclidean separation is the natural measure of spatial separation. The spatial association between two locations is often characterised as a function that decreases with increasing spatial separation. An example where these assumptions are not appropriate is the case of a set of connected stream segments that makes up a river network, where observations are made on the tributaries, lakes and other water bodies of which the river is comprised.

In standard spatial modelling applications with environmental data, *Tobler's first law of geography* is generally assumed to be true, namely that 'Everything is related to everything else, but near things are more related than distant things' Tobler (1970). In practice, such an assumption is justified when, after accounting for covariate data that is available using a linear or additive model, spatial correlation is conspicuous in the residuals. Geostatistical models typically then make use of a continuous covariance function whose argument is the Euclidean separation between two points, to account for the remaining spatial variation. This spatial variation is often imagined to arise from one or a number of underlying and unmeasured variables that drive the response of interest. For example, the prevalence of respiratory disease is primarily driven by smoking prevalence and social deprivation and to a lesser extent by other factors such as air pollution. For a variety of reasons including the difficulty in adequately measuring social deprivation, strong residual spatial structure remain when building simpler regression models, and a common remedy is to introduce a spatially smooth set of random effects to account for these spatially structured effects, for a recent example see Lee & Mitchell (2014).

In the air pollution scenario, it makes sense to think of how air pollutant concentrations decrease with distance from a point source, and one might imagine that the decline in concentration might be the same in any direction around the source at the same distance (assuming there is no wind). By contrast, the change in concentration of a dissolved pollutant in a stream is not described by the same type of process, because the mixing of water at confluences means that abrupt changes in concentration are likely as the path of the river is followed. Even after accounting for this mixing process, Euclidean distance separating two points on the river network is unlikely to well describe the correlation between them, particularly for pollutants that are driven by neighbouring land types and uses. For example, a meandering river might pick up large quantities of nitrate from surrounding farm land, and although the locations at which the meander starts and ends may be close in Euclidean separation, the water may have travelled some distance between them and exhibit strong dissimilarity in nitrate concentrations as a consequence. For these reasons, additional consideration is required for modelling data on river networks, and a brief history of developments in this area is now provided.

4.2 Literature review

There is an emerging literature that attempts to incorporate the unique features of a river network into an appropriate spatial model. The most important of these features is the inadequacy of Euclidean distance as an appropriate measure of spatial separation. Clement *et al.* (2006) approach this issue by treating river based monitoring sites as nodes on a directed acyclic graph (DAG) and measurements in time are modelled using an autoregressive process, computational details are described in Clement & Thas (2007).

Ver Hoef *et al.* (2006) replace Euclidean distance with *stream distance* which is defined as 'the shortest distance between two locations, where distance is only computed along the stream network' and is also used by Cressie *et al.* (2006) and Gardner *et al.* (2003). Ver Hoef *et al.* (2006) and Cressie *et al.* (2006) both show that substituting stream distance for Euclidean distance does not, in general, produce a valid spatial covariance model except when the exponential covariance model is used. Ver Hoef *et al.* (2006) and Cressie *et al.* (2006) subsequently develop a broad class of valid spatial covariance models that use stream distance in addition to weights that determine influences of inflows and outflows at confluence points as well as the flow connectedness of observation points. In particular, these models assign a correlation of zero to pairs of locations which are not flow-connected. Ver Hoef & Peterson (2010a) developed the theory further by defining 'tail-up' and 'tail-down' moving average constructions, that allow for correlation between pairs of locations which are not flow-connected. Subsequent applications of these developments include Peterson & Ver Hoef (2010), Garreta *et al.* (2010), Peterson *et al.* (2013),. Much of the methodology developed by Ver Hoef and Peterson is now implemented in the `R` package `SSN` (Hoef *et al.* (2014)) and uses a standardised `S4` object to store the essential features of a river network that are required for modelling.

Both of the approaches of Ver Hoef & Peterson (2010a) and Clement & Thas (2007) are powerful and allow broad classes of models to be fitted to river network data. However, neither approach easily allows the incorporation of smooth covariate effects, which is particularly important for capturing seasonal patterns and changes over longer periods of time. An attractive approach is then to place the emphasis on the direct modelling of these trends, using suitable forms of flexible regression. This line of thinking is also well developed, for example in the geoadditive models of Kammann & Wand (2003) and more generally in the semiparametric and additive modelling frameworks described, for example, by Ruppert *et al.* (2003) and Wood (2006). An attractive approach would satisfy the dual goals of capturing spatial variation on the path of the river network while respecting behaviour at confluence points *and* incorporate smooth temporal and other covariate effects.

The aim of this chapter then is to develop methods of flexible regression for data located on the branches of a river network. In Section 4.3 the River Tweed data are described that will subsequently be the focus of the work documented in this chapter, followed by some developments in Section 4.4 that allow network data to be modelled using the P-splines framework described in Chapter 2. In Section 4.5 a more general framework allowing smooth additive and spatial interaction components is described and developed, which is in turn applied to the River Tweed data. Section 4.6 closes the chapter with a discussion of the findings and points for further research.

4.3 The River Tweed data

In this chapter, the data that forms the focus and motivation behind the models that are developed come from the River Tweed in South East Scotland and have been supplied by the Scottish Environment Protection Agency (SEPA). SEPA is responsible for a broad range of environmental regulation and monitoring, and as such is involved in the collection and analysis of water quality data from rivers and lochs across Scotland. The importance of this is underlined by European Union policies such as the Nitrates Directive (European Parliament (1991)) and the Water Framework Directive (European Parliament (2000)), which set targets in terms of water quality and ecological status.

Data on the Tweed catchment are available from January 1987 to August 2011 for each of eighty three monitoring stations on the river, although the range and sampling frequency of observations varies across the stations. The River Tweed has been selected because it lies in a nitrate vulnerable zone and is highly heterogeneous in terms of nitrate pollutant levels. It is also a highly dendritic network, which makes it particularly challenging from a spatial modelling perspective. Figure 4.1 illustrates both of these properties by superimposing the main tributaries of the river over a physical map of the surrounding area and shows the mean nitrate pollutant measurement at each active station during February 2004. Stations which were inactive during February 2004 are represented by empty circles.

Although nitrate level is not the only measurement available over the River Tweed and other rivers monitored by SEPA, it will form the focus of the work in this chapter. High levels of nitrate can cause damaging algae build up called eutrophication that results in reduced dissolved oxygen levels in affected water bodies, which can severely impact endemic aquatic ecosystems. Drivers of nitrate pollution on the River Tweed are diffuse sources such as sewage effluent and runoff from fertiliser. Two different types of nitrate measurement are available, namely Nitrate (N) and Total Oxidised Nitrate (TON), both measured in milligrams per litre. TON is the sum of nitrate and nitrite concentrations, although the latter tends to be very small, so the two measures are treated as equivalent and this will be assumed in the analysis. In order to improve normality and stabilise the variance, nitrate level will be analysed on the log scale.

The analysis of concentrations of dissolved pollutants hinges critically on knowledge of river flow levels around the time of a pollutant concentration measurement. Unfortunately, reliable flow data are only available at a limited set of locations on the Tweed and consequently, SEPA uses a hydrological model to estimate relative flow levels for each of the 298 separate stream segments. This is an adequate representation because, in later analysis, only relative flow across the stream segments is required. In cases where there is

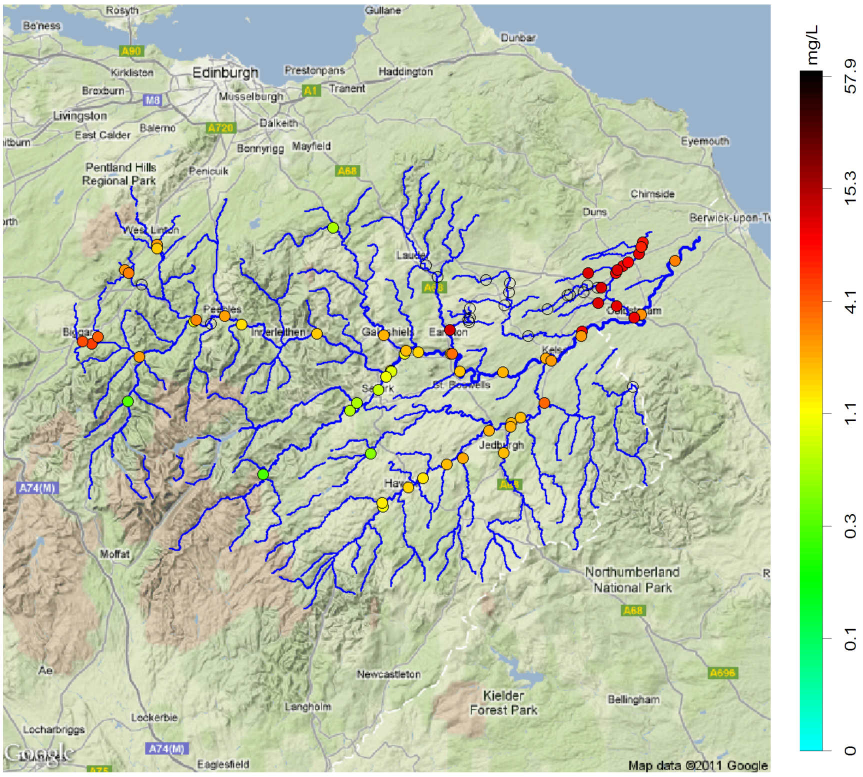


FIGURE 4.1: The River Tweed catchment, with sampling stations colour coded by nitrate level recorded in February 2004. Stations where no measurement was available at this time point are indicated by open circles. The map is based on Google maps, and can be produced using the `RgoogleMaps` package described by Loecher (2014).

concern about the quality of flow information, an alternative is to follow Ver Hoef *et al.* (2006) by using a proxy such as 'stream order', which indexes each stream segment by its location in the hierarchy of tributaries to the main river. However, flow rates and volumes represent a source of uncertainty in models for dynamic systems such as river networks that should ideally be accounted for. It is therefore a shortcoming of any procedure to assume as fixed such quantities without knowing how much variation is likely to underlie them - this point is discussed in more detail in Section 4.6.

4.4 Network smoothing

In keeping with Chapter 2, we wish to represent a river network as a structure over which smoothing using an appropriately constructed basis representation might allow insight into spatial trends and patterns in a computationally efficient manner.

4.4.1 Penalised splines

A desirable approach would involve constructing an appropriate B-spline basis over the branches of the network, but this faces the difficulty of constructing appropriate penalties across confluence points and matching the derivatives of the fitted values. One possibility that avoids this issue would be to partition the network into p homogeneous stream stretches, associate a parameter with each, and impose some type of roughness penalty over these. We can then enumerate these stream segments from 1 to p and associate with each segment a mean pollutant level ($\{\beta_1, \ldots, \beta_p\}$) so that

$$E(Z(x, y))) = \beta_i \tag{4.1}$$

where Z represents the spatial process at a location (x, y) on the i^{th} stream segment. Essentially, Equation 4.1 reduces the spatial network to a set of non-overlapping piecewise constants. Conveniently, the partitioning described above exists inherently in the way data are stored by SEPA, where every observation arises from one of a set of homogeneous 'stream units', although by the SEPA definition, there may be several stream units lying between two adjacent confluence points.

Since there are typically fewer observations than stream units, the estimation process is ill-defined and some additional information is required to describe how pollutant concentrations propagate through parts of the network that are not monitored. It is expected that there should be strong smoothness in the underlying mean surface across the network driven by many underlying factors such as the continuous nature of river flow, and the spatial homogeneity of surrounding land-types and land uses. This smoothness can be represented by application of some type of roughness penalty on adjacent stream segments.

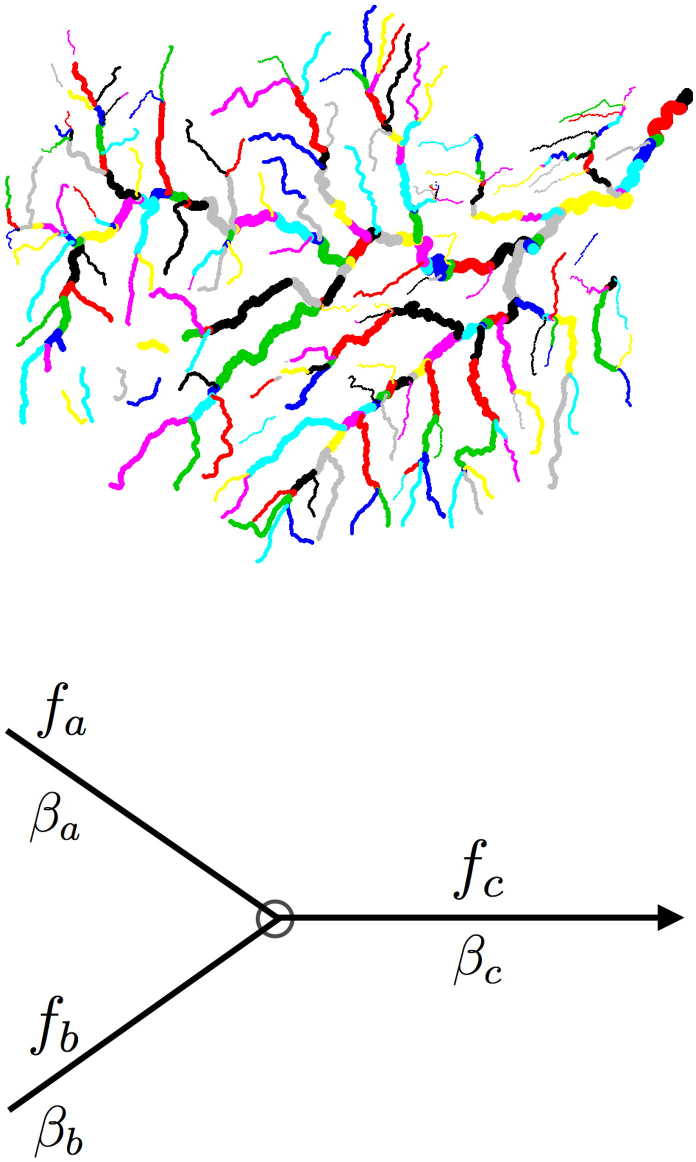


FIGURE 4.2: The top plot uses different colours to show the decomposition of the river network into a large number of small 'stream units'. The bottom plot gives a schematic representation of a confluence, with model parameters ($\boldsymbol{\beta}_a$, $\boldsymbol{\beta}_b$), flows (f_a, f_b) and the corresponding outgoing versions ($\boldsymbol{\beta}_c$, f_c).

The form of the roughness penalty that is used could take many forms, but a simple one may use the basic idea of mass balance, in which case it is helpful to consider the idealised network illustrated by Figure 4.2. The pollutant level at location c is driven by the flow levels upstream at b and a, but these contributions are clearly dependent on the relative levels of flow each contributes to c. It is therefore desirable that mean pollutant levels associated with each segment ought to reflect this property. In fact, it is also straightforward to incorporate information relating to flow levels, or some proxy, into a roughness penalty. If the flow levels are denoted by f_a, f_b, f_c, where a and b flow into c, for segments a, b and c, then we expect $f_c = f_a + f_b$ and the mixing of pollutants to be controlled by the relative flows of the inputs, $\frac{f_a}{f_c}$ and $\frac{f_b}{f_c}$. Following the principle of mass balance, the combined pollution input $\frac{f_a}{f_c}\boldsymbol{\beta}_a + \frac{f_b}{f_c}\boldsymbol{\beta}_b$ and the output $\boldsymbol{\beta}_c$ are identical if

$$\begin{aligned} \frac{f_a}{f_c}\boldsymbol{\beta}_a + \frac{f_b}{f_c}\boldsymbol{\beta}_b &= \boldsymbol{\beta}_c \\ \frac{f_a}{f_c}(\boldsymbol{\beta}_a - \boldsymbol{\beta}_c) + \frac{f_b}{f_c}(\boldsymbol{\beta}_b - \boldsymbol{\beta}_c) &= 0. \end{aligned} \tag{4.2}$$

In typical spline smoothing scenarios, a balance must be struck between an appropriate sum of squared differences between spline coefficients, representing roughness, and the least squares objective measuring model fit. Smoothness across a flow-directed confluence can therefore be measured by

$$\frac{f_a^2}{f_c^2}(\boldsymbol{\beta}_a - \boldsymbol{\beta}_c)^2 + \frac{f_b^2}{f_c^2}(\boldsymbol{\beta}_b - \boldsymbol{\beta}_c)^2. \tag{4.3}$$

In matrix notation the model described can be written as

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$, $\boldsymbol{\beta}$, and $\varepsilon \sim N(0, \sigma^2)$ denote the vectors of responses, parameters and errors respectively, and the design matrix $\boldsymbol{B}$ is simply an $n \times p$ indicator matrix whose ith row has the value 1 in the column corresponding to the stream unit of y_i and 0's elsewhere. The model is then fitted by minimising the penalised sum of squares

$$\begin{aligned} & (\boldsymbol{y}-\boldsymbol{B\beta})^\top(\boldsymbol{y}-\boldsymbol{B\beta}) + \lambda \sum_{i,j\sim k} \left(\frac{f_i^2}{f_k^2}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_k)^2 + \frac{f_j^2}{f_k^2}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k)^2 \right) \\ = \;& (\boldsymbol{y}-\boldsymbol{B\beta})^\top(\boldsymbol{y}-\boldsymbol{B\beta}) + \lambda\boldsymbol{\beta}^\top\boldsymbol{D}_s^\top\boldsymbol{D}_s\boldsymbol{\beta} \end{aligned} \tag{4.4}$$

where $i, j \sim k$ denotes that stream segments i and j are upstream neighbours of k joined through a confluence point. Here $\boldsymbol{D}_s\boldsymbol{\beta}$ expresses differences in adjacent stream segments and $\boldsymbol{D}_s$ is similar in structure to what is described in Equation 2.10. The rows of $\boldsymbol{D}_s$ consist of multiples of proportional flow contributions for the pair that the row corresponds to. The parameter λ controls smoothness by modulating the influence of the roughness measure in the minimisation. λ has an interpretation in terms of both the extent to which Equation 4.2 holds, and as a measure of spatial dependence not accounted for by an intercept term and covariates. For example, when $\lambda = 0$, the $\boldsymbol{\beta}$ are not identifiable where there is no data; for low and moderate λ, all of the $\boldsymbol{\beta}$ are identified and are relatively similar across confluences subject to Equation 4.2; when $\lambda \to \infty$, $\boldsymbol{\beta} \to 0$ and this represents a lack of spatial (network) structure in the data. This penalty describes a conditional independence model for stream units where measurements on the network are assumed to propagate via a mixing process at each confluence, rather than as a function of spatial location.

Although the penalty described in Equation 4.4 is based on the idea of two stream segments flowing into a single downstream segment, it can also be applied in situations where no confluence exists, for example if a break point is introduced in the middle of a segment resulting in one upstream segment and one downstream. The point at which the occurs is not a confluence, as it has one incoming stream segment, i, flowing into an outgoing segment k, but this can still be represented by Equation 4.4 using the proportions $\frac{f_i}{f_k} = 1$ and $\frac{f_j}{f_k} = 0$, (since there is only one upstream segment, j is treated as contributing no flow to the outgoing k).

As in previous chapters, a procedure is required to determine the optimal penalty parameter λ. Conditional on λ, the solution to this least squares problem is $\hat{\boldsymbol{\beta}} = (\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{D}^\top \boldsymbol{D})^{-1} \boldsymbol{B}^\top \boldsymbol{y}$, and associated standard errors, and degrees of freedom are readily calculated using Equations 2.12 and 2.13, respectively. For an illustration of the impact of using different values of λ to smooth the nitrate concentrations observed on the River Tweed in February 2004, see Figure 4.3.

Figure 4.4 shows the effects of smoothing on the average nitrate measurements from February 2004. The right hand panel shows the effects of P-spline smoothing while the left shows the use of standard two-dimensional Euclidean smooth, also using P-splines, both with 12 degrees of freedom. In the right panel, the network structure is represented through sharp changes at confluences that depend on flow estimates directly around each confluence point. For example, the relatively high concentrations of pollution exhibited by some of the tributaries in the northern periphery of the network are not immediately inherited by the larger and relatively unpolluted streams into which they flow, as a result of the weighted penalty described in Equation 4.4.

4.5 Spatiotemporal models for networks

If the aim is to estimate the levels of pollution over a network at a single time point, then the approach outlined in Section 4.4.1 may be sufficient. Where observations are available over time, the nature of spatial and temporal effects and their potential interactions becomes of considerable interest and, in contrast with approaches based on covariance functions, flexible regression methods can be extended relatively straightforwardly.

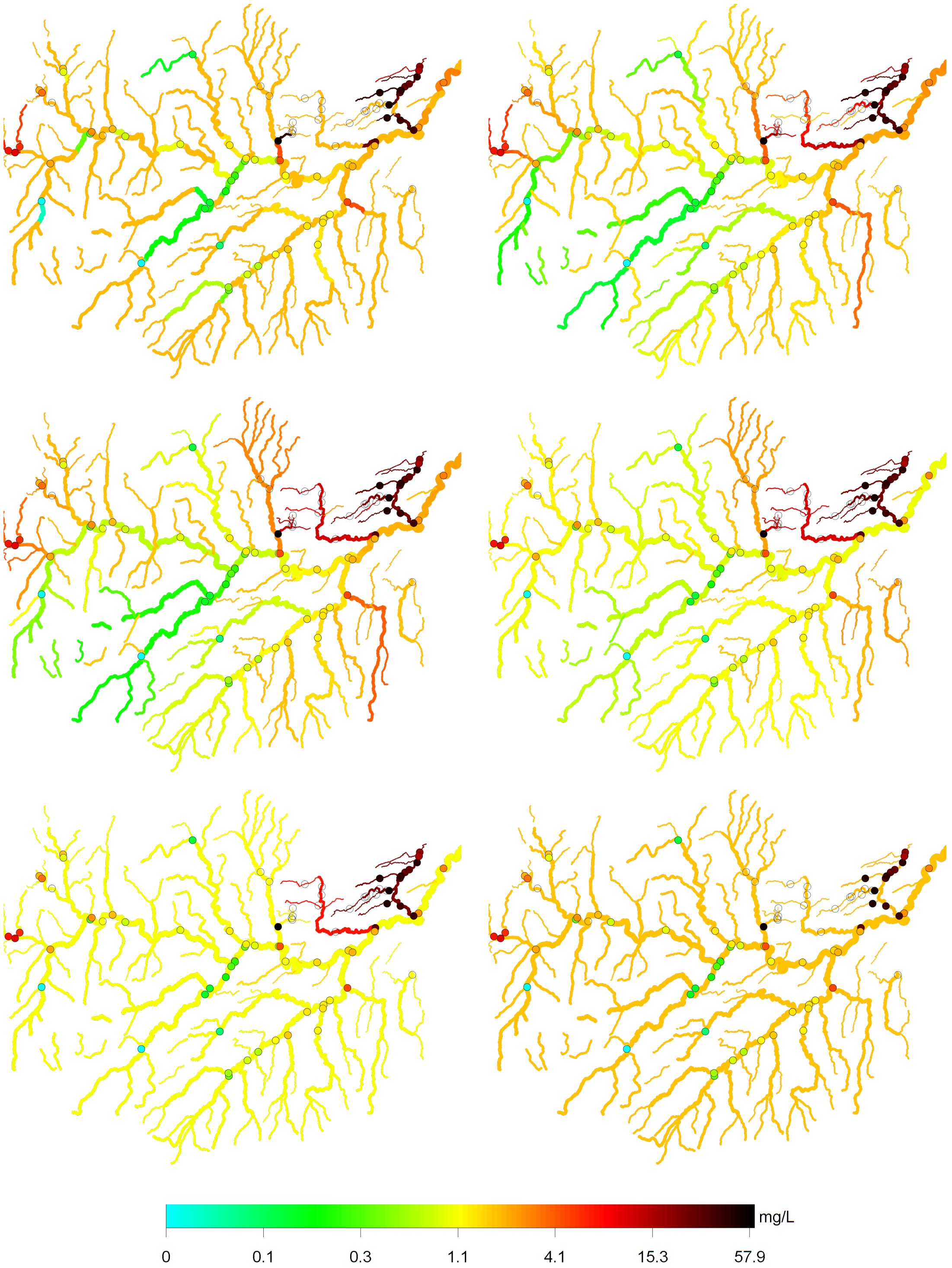


FIGURE 4.3: Plots showing the effect of different levels of spatial smoothing on the River Tweed, for nitrate data observed in February 2004. Top (from left to right) $\lambda = 0, 0.1$; middle (from left to right) $\lambda = 10, 100$; bottom (from left to right) $\lambda = 10000, 10^8$. Stream segment colours represent the estimated nitrate concentration under the particular smoothing scenario. Coloured points represent mean observed nitrate levels at each monitoring location.

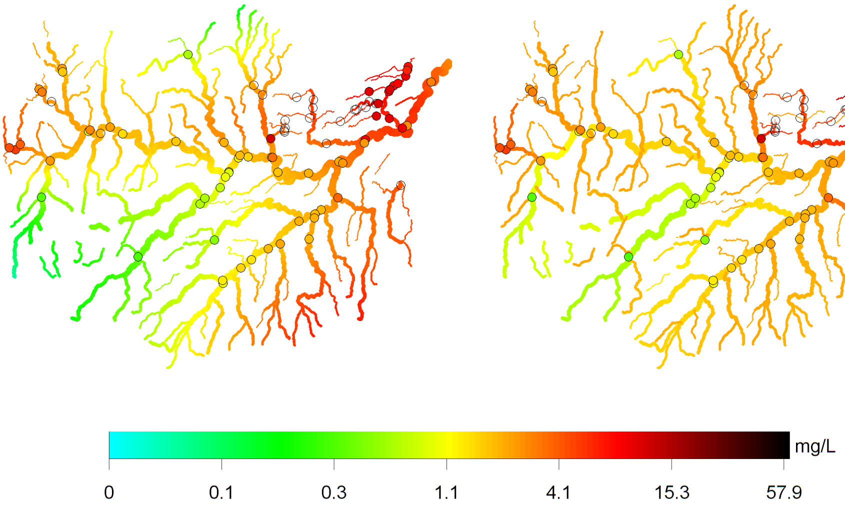


FIGURE 4.4: Left plot is a Euclidean distance smooth estimate. Right plot is a network smooth estimate.

There are many current examples of spatiotemporal models and Cressie & Wikle (2011) describe a modern introduction to a wide variety of modelling tools. Applications of space-time models are diverse and include disease mapping (MacNab & Dean (2002)), air pollution (Shaddick & Wakefield (2002)), house prices (Gelfand *et al.* (2004)) and rainfall (SansÃş & Guenni (1999)). Since long-term monitoring on a stream network is common, it makes sense to model stream data in the same framework; and some studies that have taken this approach include Cressie & Majure (1997), Clement & Thas (2007) and Militino *et al.* (2008). It is important to note that these examples do not consider the essential network features of river distance, flow connectedness and flow weighting and that as there is currently no general modelling framework for space-time stream network data. One example of a spatiotemporal model in this context is that of Money *et al.* (2009) who use a Bayesian maximum-entropy method to fit their models, however the low number of flow-connected locations in their data inhbit the use of more complex spatial statistical stream network models. Also, Gardner & McGlynn (2009) use a 'tail-up' model for nitrate data that accounts for flow-connectednessof the data locations, but the analysis is primarily spatial because only a small set of time points are available within a time interval of approximately one year.

There are three main variables which need to be accommodated in a spatiotemporal model for the River Tweed. One is space expressed through the river network locations (s_i), the second is time (t_i) measured on a scale of years to express long-term trends, and the third is time within the year (z_i) to express the seasonal changes which are often exhibited in environmental measurements. The P-spline additive models described in Chapter 2 are natural tools to consider, as they provide a framework within which flexible regression can be extended to a wide variety of data structures. In the present setting, a very simple additive model is

$$y_i = \mu + f_s(s_i) + f_t(t_i) + f_z(z_i) + \varepsilon_i \tag{4.5}$$

where the three functions f_s, f_t, f_z describe spatial, temporal and seasonal trends, y is the pollutant response and ε_i denotes error terms assumed to have a $N(0, \sigma^2)$ distribution. If each of the trend functions is estimated by B-splines then, following the ideas in Section 2.2.2, they can be represented as $\boldsymbol{B}_s\boldsymbol{\beta}_s$, $\boldsymbol{B}_t\boldsymbol{\beta}_t$, $\boldsymbol{B}_z\boldsymbol{\beta}_z$ where the columns of the design matrices evaluate each basis function at the observed values of the relevant covariate. The spatial network component f_s is represented by the piecewise constant structure described in Section 4.4.1, while cubic B-splines are good choices of basis functions for the temporal and seasonal effects, as it is expected that the estimated effects here will be simple smooth functions. The full model can be represented as $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\beta} + \varepsilon$, where $\boldsymbol{B}$ combines the columns of the individual design matrices, with an initial column of 1's.

It remains to construct suitable penalty terms to induce smoothness on the estimates of the trend functions. First-order differences were the natural choice for the spatial network parameters, as described in Section 4.4.1 and computed through a difference matrix. For cubic B-splines, second-order differencing of the parameter vector is the more standard choice. The smoothness penalty can then be expressed as

$$\boldsymbol{\beta}^\top \boldsymbol{P}\boldsymbol{\beta} = \lambda_s \boldsymbol{\beta}_s^\top \boldsymbol{D}_s^\top \boldsymbol{D}_s \boldsymbol{\beta}_s + \lambda_t \boldsymbol{\beta}_t^\top \boldsymbol{D}_t^\top \boldsymbol{D}_t \boldsymbol{\beta}_t + \lambda_z \boldsymbol{\beta}_z^\top \boldsymbol{D}_z^\top \boldsymbol{D}_z \boldsymbol{\beta}_z,$$

where the matrix $\boldsymbol{P}$ has block-diagonal form and combines the individual penalties so that

$$\boldsymbol{P} = \begin{pmatrix} 0 & & & \\ & \lambda_s \boldsymbol{D}_s^\top \boldsymbol{D}_s & & \\ & & \lambda_t \boldsymbol{D}_t^\top \boldsymbol{D}_t & \\ & & & \lambda_z \boldsymbol{D}_z^\top \boldsymbol{D}_z \end{pmatrix}.$$

and $\boldsymbol{\beta} = (\mu, \boldsymbol{\beta}_s, \boldsymbol{\beta}_t, \boldsymbol{\beta}_z)$. Cyclical behaviour in the seasonal term can be induced by requiring the coefficients of the first r basis functions to be identical with the last r basis functions. The penalty $\sum_{k=1}^{r} (\beta_{z,k} - \beta_{z,p+1-k})^2$ achieves this, with $r = 3$ for cubic splines and can be adopted into the definition of $\boldsymbol{D}_z$. The limit under strong smoothing of this penalty yields a constant function, which may not be appropriate when a very strong periodic signal is present; as an alternative, Eilers & Marx (2010) describe a special penalty that yields a sum of a sine and cosine as the limiting behaviour under strong smoothing.

In the presence of an overall mean parameter μ in Equation 4.5, the identifiability of each additive component can be achieved by the addition of a ridge penalty, as described by Eilers & Marx (2002). This corresponds to a penalty of the form $\boldsymbol{\beta}^\top \boldsymbol{Q}\boldsymbol{\beta}$, where $\boldsymbol{Q}$ is a diagonal matrix constructed from the vector $(0, \nu_s \mathbf{1}_s, \nu_t \mathbf{1}_t, \nu_z \mathbf{1}_z)$, with the ridge parameters denoted by ν_s, ν_t, ν_z and with $\mathbf{1}_a$ denoting a vector of 1's whose length is determined by the number of basis functions in the term denoted by a. The fitted model can then be expressed through the parameter estimates $\hat{\boldsymbol{\beta}} = (\boldsymbol{B}^\top \boldsymbol{B} + \boldsymbol{P} + \boldsymbol{Q})^{-1} \boldsymbol{B}^\top \boldsymbol{y}$. Denoting this as $\boldsymbol{H}\boldsymbol{y}$, standard errors for $\hat{\boldsymbol{\beta}}$, and so for fitted values, are available from the diagonal elements of $\boldsymbol{H}\boldsymbol{H}^\top$, multiplied by an estimate of the error variance which is constructed as

$\hat{\sigma}^2 = \mathsf{RSS}/(n - \mathsf{ED})$ where RSS denotes the residual sum-of-squares and ED the approximate degrees of freedom for the model, as defined in Equation 2.13. A penalised spline approach to (generalised) additive modelling is described by Marx & Eilers (1998), and many subsequent authors including Wood (2006), where further details are available.

The additive model described above is a natural starting point but it is implausible that the spatial pattern of pollution will change in exactly the same way over time, or throughout the year, at every location. It is therefore more appealing to consider an interaction model of the form

$$\begin{aligned} y_i &= \mu + f_s(s_i) + f_t(t_i) + f_z(z_i) + f_{s,t}(s_i, t_i) \\ &\quad + f_{s,z}(s_i, z_i) + f_{t,z}(t_i, z_i) + \varepsilon_i, \end{aligned}$$

where the functions $f_{s,t}$ and $f_{s,z}$ allow for differing temporal trends at different locations on the network. The term $f_{t,z}$ allows an adjustment to the overall seasonal component, allowing different patterns in different years. The interaction terms can also be conveniently represented in spline basis form, using a basis formed by all possible products of the spline basis functions on each separate variable, as described in Chapter 2. More precisely, we can write

$$f_{s,t}(s_i, t_i) = \sum_{j=1}^{p_1} \sum_{k=1}^{p_2} \boldsymbol{\beta}_{jk} B_j(s_i) B_k(t_i)$$

where $B_j(s_i)$ and $B_j(t_i)$ denote individual B-spline functions parameterising evolutions in space and time. This has the simple interpretation that the parameters associated with each stream unit are now allowed to evolve smoothly over time. Corresponding structures and interpretations can be adopted for the space-season and time-season interaction terms.

In matrix notation, the model matrix is

$$\boldsymbol{B} = \left[\mathbf{1} \,\middle|\, \boldsymbol{B}_s \,\middle|\, \boldsymbol{B}_t \,\middle|\, \boldsymbol{B}_z \,\middle|\, \boldsymbol{B}_s \Box \boldsymbol{B}_t \,\middle|\, \boldsymbol{B}_s \Box \boldsymbol{B}_z \,\middle|\, \boldsymbol{B}_t \Box \boldsymbol{B}_z \right]$$

where $\Box$ is the row-wise tensor product defined as $\boldsymbol{A} \Box \boldsymbol{Y} = (\boldsymbol{A} \otimes \mathbf{1}') \odot (\mathbf{1}' \otimes \boldsymbol{Y})$ and $\odot$ denotes the Hadamard (element-wise) product; see Eilers *et al.* (2006).

Smoothness in the model terms is induced by applying appropriate penalties, and corresponding penalty parameters λ_i for each term and allowing anistropic smooths for the 2-dimensional components $\boldsymbol{B}_s \Box \boldsymbol{B}_t$, $\boldsymbol{B}_s \Box \boldsymbol{B}_z$ and $\boldsymbol{B}_t \Box \boldsymbol{B}_z$. In the case of main effects, these are constructed through the difference matrices described above. Penalties for the interaction terms can be constructed by considering the coefficients $\{\boldsymbol{\beta}_{jk}\}$ in matrix form and applying smoothness penalties to both the rows and the columns. For example, space-time smoothness is induced by applying a first-order network penalty to the columns of the matrix $\{\boldsymbol{\beta}_{jk}\}$ and a second order difference penalty over the rows. As described above, identifiability is ensured, and ill-conditioning avoided, by adding a ridge penalty for each term in the model, expressed in a diagonal matrix $\boldsymbol{Q}$. Other constraints or penalties exist that could have achieved a similar effect, for example constraints that force the mean value of each component to be zero. However, the straightforward specification of the ridge penalty and the subsequent retention of sparseness of model objects makes this a convenient choice, as discussed in the subsection on computational details below. Each of the λ_i is estimated by a short search procedure to find values that minimise the corrected AIC (AICc) as defined in Section 2.3.1.

Tensor product spline smooths such as those specified in Section 3 rely on many basis parameters to represent each bivariate interaction and may therefore be intensive to fit. In the present case, the model matrix $\boldsymbol{B}$ is mostly composed of terms involving $\boldsymbol{B}_s$, an $n \times p_s$ matrix where p_s is the size of the network partition. $\boldsymbol{B}_s$ contains exactly n nonzero entries.

If $\boldsymbol{V}$ denotes the $n \times q$ matrix evaluating the bases for the other terms in the model, then $\boldsymbol{B}_s \square \boldsymbol{V}$ is at least $100(1 - \frac{1}{p_s})\%$ sparse and much more so if $\boldsymbol{V}$ is sparse. Sparse matrix algorithms can be used to decrease storage requirements and vastly increase performance. For example, the $\boldsymbol{B}$ that was fitted to the Tweed data, was 16000×5000 where $p_s = 298$ and was 99.8% sparse.

Figures 4.5 and 4.6 shows the results of fitting this interaction model to the Tweed data, using AICc to select all the penalty parameters. The top left hand plot, together with those in the second row, show estimates of the main effects for space, year and day of year. This highlights that areas of high pollution are present in the tributaries to the North East of the River Tweed. Across the years the overall pollution levels are relatively stable, but with some indication of a slight decreasing trend. The overall seasonal effect is strong, as expected, with a gentle decrease from February to August and a sharper rise at the end of the calendar year. The shaded bands in Figure 4.5 correspond to two standard errors on either side of the estimates and these indicate high precision, as a result of the substantial size of the data set. The top right hand plot shows the estimate of interaction between year and season. The values of the adjustments plotted here are small, indicating that the change in seasonal pattern over the years is modest. The two plots in Figure 4.6 show fitted values at two specific spatial locations, along with a comparison of a simple main effects model (green) and the interaction model (red). This shows clear improvement at some sites as a result of fitting the interaction terms.

4.5.1 Residual correlation

Having established a spatial interaction model for the River Tweed nitrate data that is appropriate for its network structure, it remains to check the assumption of independence made of the residuals. Informal plotting of the variogram clouds (Diblasi & Bowman (2001)) of the residuals uncovered some strong evidence of the presence of residual temporal

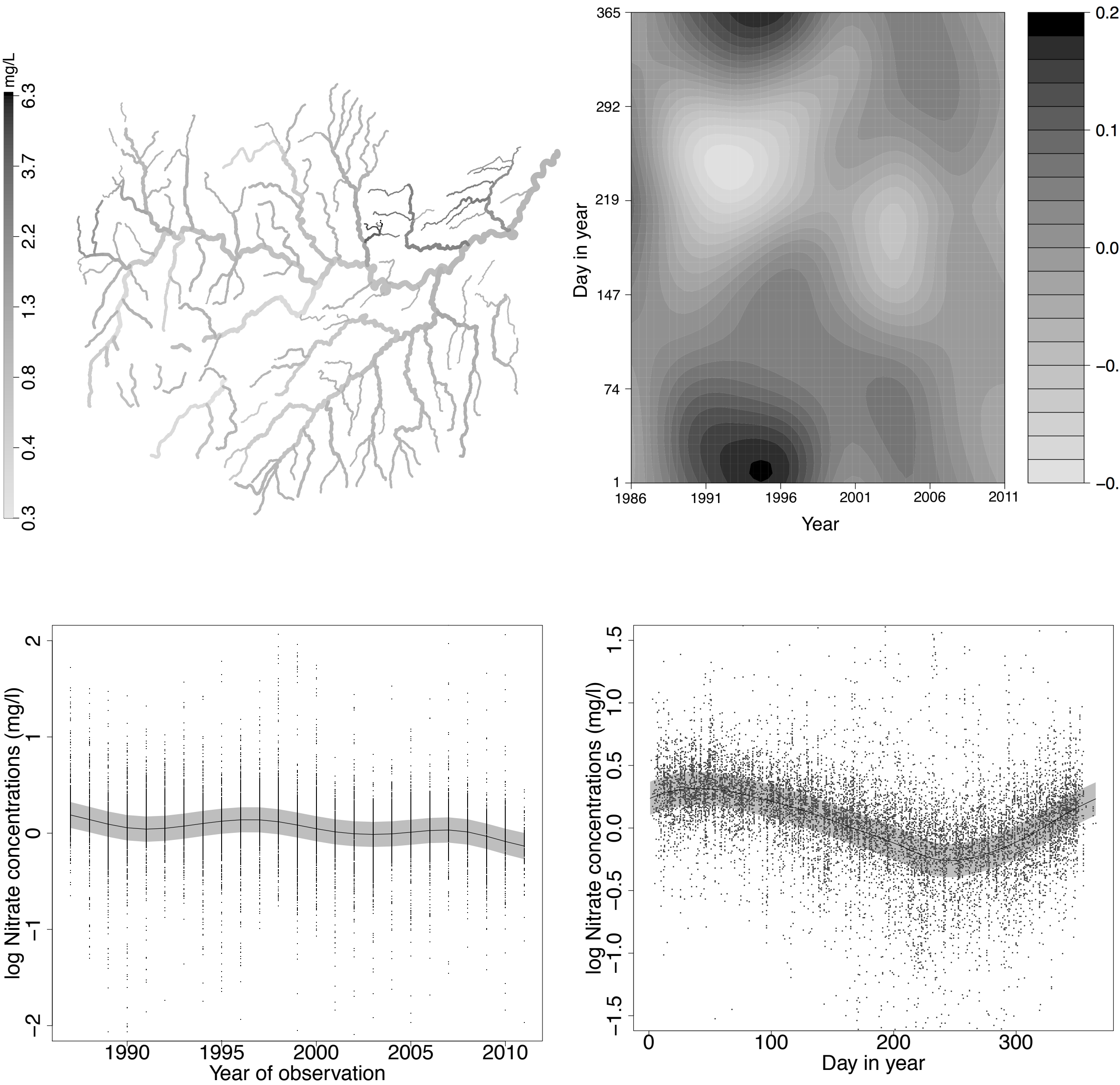


FIGURE 4.5: Clockwise from top left: main effect of space where darker grey corresponds to higher mean nitrate levels; image plot interaction showing interaction between the seasonal pattern and longer-term trend; plot showing estimated seasonal pattern (black line), with 95% confidence bands (grey shaded region), and partial residuals; plot showing long-term trend (black line) with 95% confidence bands (grey shaded region) and partial residuals.

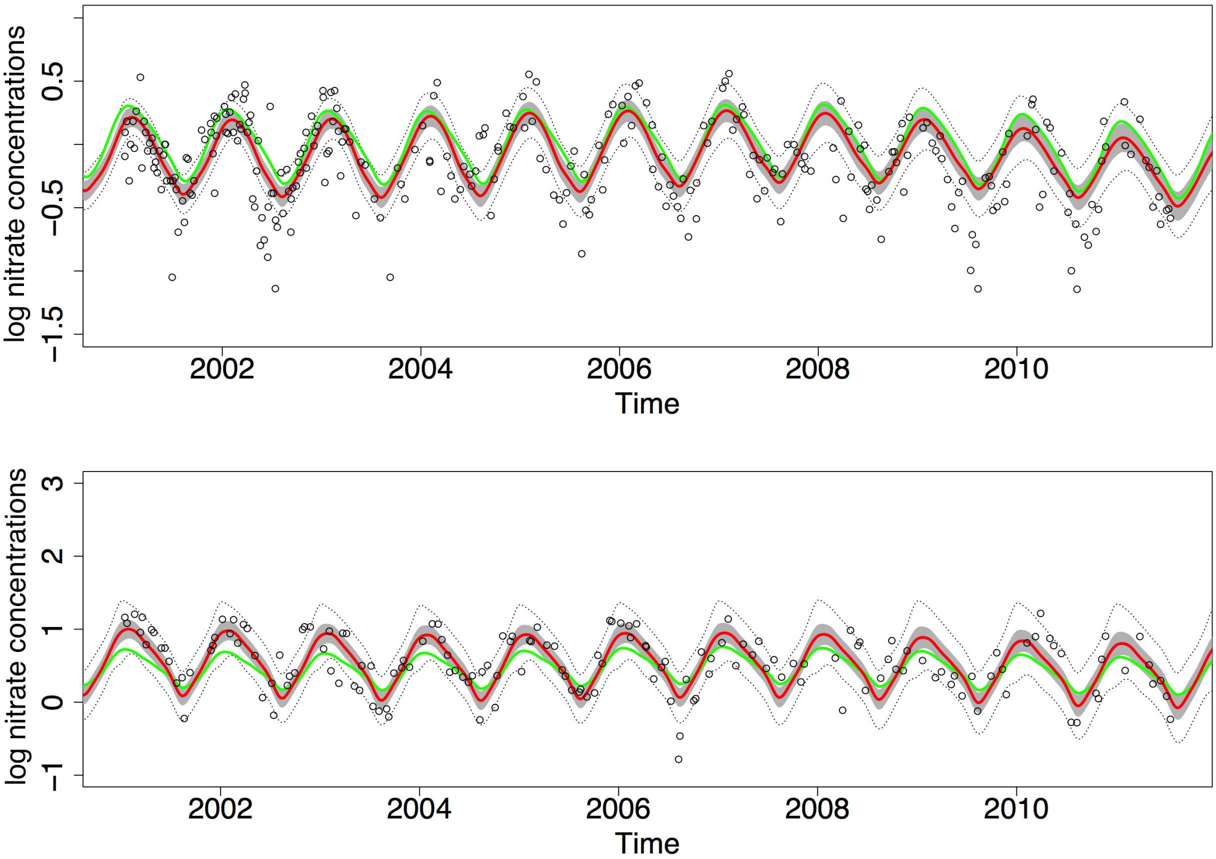


FIGURE 4.6: Comparison of fitted values for simple main effects model without interaction terms (green line) and the full model (red) with 95% confidence bands (grey shaded region), confidence bands after correction for residual correlation (dotted line) and data points (plotted as circles). Top shows fitted values at Gala Water Foot monitoring station; bottom shows those for Norham Gauging Station.

autocorrelation, and to a lesser extent some spatial autocorrelation. Evidence of residual temporal correlation at short time lags is to be expected, particularly as the model accounts for temporal structure only over longer time periods. Under the assumption of independent errors, all standard error estimates are likely to be underestimated when the underlying error process is correlated, so they must be adjusted appropriately. As a conservative measure, it was decided to fit a separable spatiotemporal model to the errors so that

$$\hat{\Sigma}_{ij} = Cov(\epsilon_i, \epsilon_j) = \omega_{ij}\, \sigma^2 \, \exp\left\{ -\frac{d_{ij}}{\rho} - \frac{|t_i - t_j|}{\psi} \right\}$$

where ω_{ij} is the product of the flow proportions contributed by each upstream segment that lies between segments i and j. The spatial and temporal correlation in the error process is assumed to depend on $t_i - t_j$, the time lag, and d_{ij}, the network separation measured in numbers of stream units. The correlation model was fitted by weighted least squares.

Having obtained an estimate for $\hat{\Sigma}$, the standard errors for the fitted values were then adjusted by

$$\mathsf{s.e.}\{\hat{\boldsymbol{y}}\} = \sqrt{\mathsf{var}\left\{\hat{\boldsymbol{H}}\boldsymbol{y}\right\}} \quad = \quad \sqrt{\mathsf{diag}(\hat{\boldsymbol{H}}\hat{\Sigma}\hat{\boldsymbol{H}}^\top)}$$

where $\hat{\boldsymbol{H}}$ is the projection or hat matrix given by $\boldsymbol{B}(\boldsymbol{B}^\top\boldsymbol{B}+\boldsymbol{P}+\boldsymbol{Q})^{-1}\boldsymbol{B}^\top$. The estimated parameters in the correlation model were $\rho = 8.3$ and $\psi = 27.4$ which represents moderate residual temporal correlation and (after adjusting with weights) weak residual spatial correlation. These parameters refer to a spatial scale in miles relative to a catchment diameter of approximately 70 miles, and a temporal scale in days, relative to a span of 26 years for the whole data set. The overall variance parameter σ^2 was estimated as 0.1554, which is very close to the estimate under an independence assumption (0.1442). The corresponding adjustments to standard errors are displayed in Figure 4.6 as dashed lines, from which it is clear that the increases in width over the independence model are not sufficiently large to lead to any substantive change in conclusions.

It would be possible to consider incorporating the correlation structure into the fitting process for the model. This would, however, considerably increase the complexity of the computations, particularly as sparsity would be compromised. The post-fitting adjustment approach combines computational efficiency with an effective first-order approximation to the correlation structure, which has been used to good effect in similar settings, as discussed by Giannitrapani *et al.* (2011).

4.5.2 Visualisation

While simple spatial terms can be plotted in map or network form, interactions with spatial components are more problematic to view. Figure 4.6 shows temporal effects at particular point locations on the River Tween network. An alternative illustrated in Figure 4.7 is to display the estimated spatial effects at different time points, here at three different months (January, June and November) in 2005. This helpfully focuses attention on the spatial areas where seasonal change is strongest. However, changes in colour alone can be difficult to assess, especially where those changes are modest. The plots shown in Figure 4.7 represent the values over the network as 'nodes', plotted approximately in the geographical midpoint of each stream unit. In addition to colour code, each node has radius proportional to the estimated nitrate pollutant level (on the original rather than log scale). This form of display is particularly effective at illustrating changes over time as small changes in size are more easily identifiable than small changes in colour.

A more satisfactory solution involves animation of the spatial pattern across time. This kind of effect can be achieved with graphical tools such as those provided by the `rpanel` package (Bowman *et al.*, 2007) for `R` (R Development Core Team, 2011). This allows the time setting for the spatial display to be controlled through a slider. In a similar manner, sliders can also be used to control the degrees of smoothing through interactive selection of values for the approximate degrees of freedom. Since visualising and understanding spatiotemporal model fits is challenging from static printed plots, two animations of the fitted models are provided online at `http://vimeo.com/46476977` and `http://vimeo.com/46321492`. These illustrate spatial and temporal variation in the fitted mean nitrate levels in both network and node form. The effect of the spatial penalty across neighbouring stream units is more evident in the first, while the degree of pollution and change through time is arguably better represented by the second.

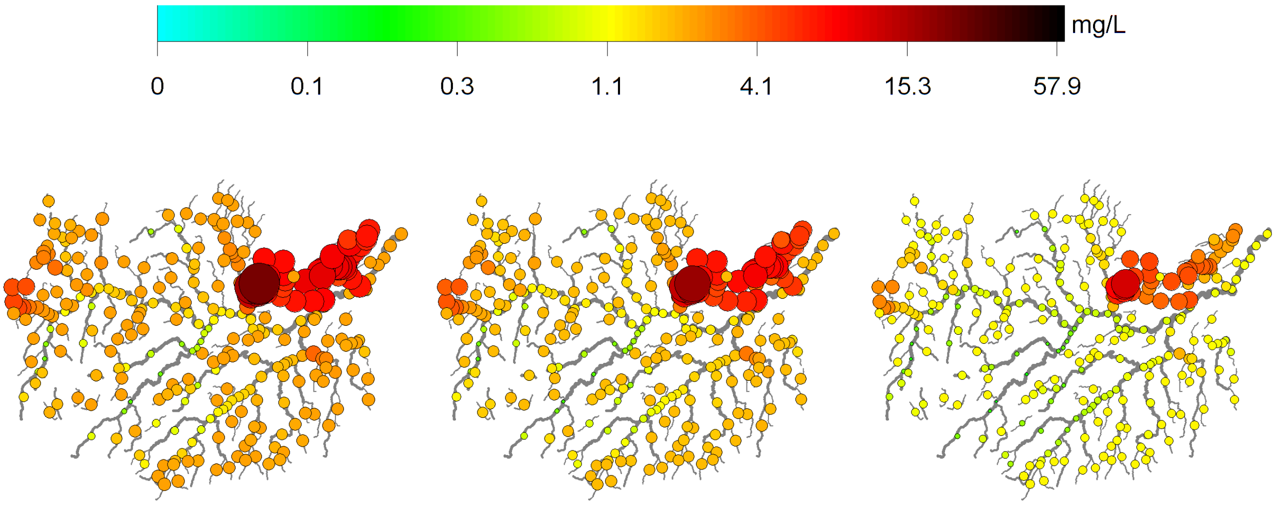


FIGURE 4.7: Estimated spatial effects at three different months (January, May and October) in 2005, indicated by colour and scaling of 'nodes' located at the stream units.

4.6 Discussion

In this chapter, flexible regression models were proposed that respect the unique spatial structure of data arising on river networks that can be easily extended to build spatio-temporal models for capturing complex spatial change. The results of fitting the models to the River Tweed nitrate data revealed evidence of strong overall seasonal patterns, but only a small degree of overall change in nitrate levels over the 26 year period. Perhaps of greater interest is the insight gained from interaction terms between space and time components, which permitted spatially-varying effects that do so while respecting the flow-driven structure of the stream network. It is clear from Figure 4.6 that local differences in seasonal and long-term effects account for a substantial proportion of variation at a specific site over time. This strongly supports the argument in favour of building and retaining well designed long-term monitoring networks, in order to gain insights into variation across a network as a whole which in turn can help to identify the effects of local environmental damage.

Attention has focused on the estimation of model terms and their standard errors, as these give clear and interpretable insight into the structure of the data. More formal methods of model comparison can be implemented for example through the approximate F-tests described by Bowman *et al.* (2009) in the spatiotemporal setting.

4.6.1 Model mispecification

The penalised models described use a discrete approximation to represent what is a continuous spatial process defined along the path of a set of stream segments. This approximation is useful because it allows the influence of confluences to be easily represented by assuming conditional independence of upstream and downstream segments. In addition to providing the necessary model structure to allow network-indexed variables to be estimated on unmonitored segments, the particular choice of penalisation renders all of the matrices involved in model fitting very sparse and therefore computationally straightforward to store and manipulate. The sparseness property allows complex network models to be constructed and further covariates to be included without the computational limitations which can hamper other approaches such as those based on Gaussian processes and covariance functions. The issue of computational speed is particularly relevant to the method used to select optimal smoothing parameters, especially because data sets exist with numbers of stream segments of the order 10^3 and 10^4, and with similar numbers of data points. For these reasons, the models developed in this chapter are particularly appropriate for very large or densely sampled networks.

Despite these advantages, the most obvious problem with the model thus specified is that variability that occurs at small scales and within stream segments cannot be captured. If this is found to occur, it will be necessary to perform the type of adjustment described in Section 4.5.1 in order to obtain more realistic standard error estimates. However, this is not an elegant solution, and a desirable but more complicated approach would involve the construction of a higher order B-spline basis on the network, that can represent smooth changes *within* stream segments. Since this is a much more complex problem, Section 6.3 is dedicated to sketching a potential solution to this issue.

It is also important to note that the log transformation performed on the nitrate data at the outset of the study means that the additive mass-balance argument made to motivate the spatial penalty described in Section 4.4.1 should be treated as representing an approximation of the true mixing process of pollutants at confluences. Although this does not diminish the generality of the methodology developed, a further refinement that would remove the approximation would involve modelling the nitrate data on its original scale and assume some positive and non-normal error structure, using for example a Gamma or log-Normal distribution.

4.6.2 Link to Gaussian Markov Random Fields

The penalised model specification can be translated into a Bayesian hierarchical model using a similar argument to that in Section 2.3.3. The stochastic analogue of the quadratic difference penalty in the penalised least squares criterion described in Equation 4.4, is the prior

$$\boldsymbol{\beta} \sim N\left(0, \frac{1}{\tau}(\boldsymbol{D}^\top \boldsymbol{D})^{-1}\right). \tag{4.6}$$

Equation 4.6 describes a special case of a Gaussian Markov Random Field (GMRF), Rue & Held (2005)) prior for $\boldsymbol{\beta}$, where a GMRF is a multivariate Gaussian distribution in which a least one pair of the $\boldsymbol{\beta}_i$ are conditionally independent of each other. This condition is easily seen to be satisfied by considering the result (Rue & Held (2005)) that if $\boldsymbol{x} \sim N(\boldsymbol{a}, \boldsymbol{Q}^{-1})$, then x_i and x_j are conditionally independent iff $Q_{ij} = Q_{ji} = 0$; as noted previously, $\boldsymbol{D}^\top \boldsymbol{D}$ is sparse and therefore Equation 4.6 describes a GMRF prior. There is a particularly strong connection with conditional autoregressive (CAR) models, which are special cases of GMRFs and are often used in applications such as disease mapping, and where spatial dependence is defined in a similar manner through a first order neighbourhood relationship. A Bayesian approach has attractive features, particularly because sampling from the posterior distribution allows uncertainty associated with smoothing parameters to be integrated out. At present, smoothing parameters are selected through a grid search over a range of candidate values, which can be cumbersome for multidimensional smooths or many covariates. In addition, for MCMC updates, the conditional independence structure of the random effects $\boldsymbol{\beta}_i$ lends itself to the efficient block updating for GMRFs described by Fahrmeir & Lang (2001). It is noted however, that significant computational demands can be made by the fitting of Bayesian space-time models in which anisotropic smoothing is required due to the presence of multiple smoothing parameters that must each be updated using a Metroplis-Hastings sampler. An important element of future work might seek to compare the current penalised least squares approach with this Bayesian framework.

4.6.3 Comments on the use of flow data

Regardless of estimation procedure used, some measure of proportional flow volumes are required for each stream unit in the network partition so that the idea of mass balance can be used across confluences. Flow data used here for the River Tweed was not observed, and instead came from modelled values used by SEPA. Ideally, observed flow data, where available, could allow a model to adapt to different flow settings over time, as observed by Cressie & O'Donnell (2010). Alternatives to using modelled flow data might be to use some proxy based on a notional stream order, such as Shreve order (Shreve (1966)). Of course, flow levels are highly variable in time, and so any flow measure is limited in its ability to inform flow mixing across confluences at anything other than the broadest time scales. Since it is important that this uncertainty is accounted for by the model, a more robust approach might treat the flow rates as random quantities that have prior distributions that might be informed by whatever data are available. It is highly likely that this would require a Bayesian hierarchical approach to model fitting.

Chapter 5

Validation of river network models

5.1 Introduction

Chapter 4 proposes a new approach to modelling stream network data that is capable of capturing complex and smoothly-varying changes across space and also smooth space-time interaction, and was subsequently able to provide new insights into the River Tweed data. In order for the models to be widely adopted, it is important that practitioners can be confident of the conditions under which they will perform well, and that they are appropriate tools for investigating a particular problem. An important consideration is that additive models with stream network components and multidimensional smooths, as described in Chapter 4, are complex to implement and it is necessary to remove this barrier. The goal of this chapter is therefore to address three related questions:

1) Can general purpose software be written that fits a broad class of river network models, and if so, how should the data be formatted?
2) Does the performance of the models described in Chapter 4 depend on the particular spatial covariance structure of the data, and if so to what extent?

3) Do the models perform well in comparison with other approaches that are already established in the stream network modelling literature?

To address 1), an `R` (R Development Core Team (2011)) package was developed that implements the stream network models outlined in Chapter 4. The package fits additive models to stream network data and automatically selects smoothing parameters, all through a user-friendly high-level interface. Using this software, 2) can be investigated by carrying out a simulation study aimed at assessing the empirical model fit performance of the penalised models described in Chapter 4 and O'Donnell *et al.* (2013). The fitting of stream network models is often motivated by a desire to make predictions at unmonitored spatial locations, and consequently the study emphasises predictive performance and discovering conditions that give rise to different levels of accuracy and interval coverage.

Although Chapter 4 described new models for stream network data, other techniques exist in the literature, and in particular those of Ver Hoef *et al.* (2006) and Ver Hoef & Peterson (2010a) are supported by well-developed software and have been used in a growing number of studies. Their approach is to build random variables whose spatial covariance is appropriate for stream networks; in particular, they build a Gaussian process where the covariance matrix is a function of separation defined by stream distance along the path of the stream network and incorporates the influence of relative flow levels and confluences. We therefore aim to address 3) by comparing the performance of the models of Chapter 4 and O'Donnell *et al.* (2013) with those of Ver Hoef *et al.* (2006) as part of the same simulation study. In order to distinguish between the two approaches we henceforth abbreviate and refer to the approach of Ver Hoef *et al.* (2006) as VHPT, and those detailed in O'Donnell *et al.* (2013) as OD. The overall aim of this comparison is to provide clear practical guidance about the relative costs and benefits associated with each modelling approach.

This chapter has the following structure: Section 5.2 discusses issues around designing user-friendly software to implement the models of Chapter 4 in a computationally convenient manner. Section 5.3 provides a summary of some of the theoretical aspects of the modelling framework proposed by Ver Hoef & Peterson (2010a), while Section 5.4 describes the design and implementation of a comprehensive simulation study that will enable comparison between the relative performances of the two approaches across a variety of realistic simulated stream network data. In Section 5.5 the relative performance of each model is evaluated based on the results of the simulation study, and the main differences are summarised. Finally, some discussion follows in Section 5.6 about the implications of the results for practitioners, and for future development of both techniques.

5.2 Software for penalised network models

In order to enable other users to use the methodology described in Chapter 4, an `R` package was developed that allows the general specification of additive models for data arising on stream networks. Importantly, the class of models described in Ver Hoef & Peterson (2010a) are well supported by the `R` language R Development Core Team (2011) software library `SSN` (`http://cran.r-project.org/web/packages/SSN/index.html`), and to ease the fitting of such models, they define a new standard `S4` object (Chambers (2008)) which stores all of the attributes of a stream network that are required for modelling and visualisation, and these are in turn obtained from a Geographical Information System (GIS). In writing software for the models of O'Donnell *et al.* (2013), it makes sense to adopt this framework in order to allow users to use both approaches on their data with little additional effort. This section discusses the design and use of such software, and in particular Section 5.2.1 discusses issues associated with selecting optimal smoothness parameters using general purpose optimisation; Section 5.2.2 investigates the accuracy and computational speed associated with using the trace approximation algorithm that was introduced in Section 2.3.2; Section 5.2 concludes the section by describing the software that was developed for fitting additive models for spatial stream networks, and its basic functionality.

5.2.1 Smoothing parameter selection

To automate the selection of smoothing parameters required for fitting the models described in O'Donnell *et al.* (2013), there are a number of issues that must be addressed. The use of the bias-corrected AIC (AICc, 2.19, Hurvich *et al.* (2002)),

$$\mathsf{AICc} \quad = \quad \log(\hat{\sigma}^2) + 1 + \frac{2(\mathsf{tr}(\boldsymbol{H}) + 1)}{n - \mathsf{tr}(\boldsymbol{H}) - 2},$$

as a way of achieving bias-variance trade-off across a range of values of the smoothing parameters can be problematic when the number of parameters to be smoothed is greater than the number of observed data. In these cases, it is possible for low values of λ to cause $\mathsf{tr}(\boldsymbol{H}) - 2$ to approach n and as a result, AICc exhibits very sharp local maximum and minimum points in this region. Since these typically occur when λ is close to zero, and these are unlikely to be realistic values for the smoothing parameter that achieves a parsimonious fit to the data, it is necessary to avoid these regions of the AICc surface when attempting to find a minimising value for λ. In order to reduce the number of evaluations required to find a minimum and avoid the use of a grid search procedure, it would be attractive to use a general-purpose optimiser. As a consequence of the issue arising from $\mathsf{tr}(\boldsymbol{H}) - 2 \to n$, the optimiser should be steered away from small values of λ that cause this behaviour, and it is for this reason that a box-constrained Nelder-Mead optimisation algorithm was chosen for this purpose, implemented in the R package `dfoptim` (Varadhan *et al.* (2011)).

Although general-purpose optimisers are very efficient, many repeated evaluations of AICc are still required which can be very computationally expensive for large n or $n_{\texttt{seg}}$. As discussed in Section 2.3.1 this is due to the need to calculate the effective dimension $\mathsf{tr}(\boldsymbol{H})$

for many different configurations of $\boldsymbol{\lambda}$. Therefore, we resort to the trace approximation that is described in Section 2.3.2, where a Monte-Carlo estimate of the trace operation can be obtained by

$$\mathsf{tr}(\boldsymbol{H}) \approx \frac{1}{s}\sum_{i}^{s}\sum_{j}^{s}\left[(\boldsymbol{U}^{\top}\boldsymbol{B}\boldsymbol{L}^{-1})^{2}\right]_{ij}$$

where $\boldsymbol{U}$ is a matrix of s random column vectors composed of uniformly generated random sequences of 1 and -1, $\boldsymbol{L}$ is the (sparse) Choleski decomposition of $(\boldsymbol{X}^{\top}\boldsymbol{X}+\boldsymbol{D})$ that depends on the vector $\boldsymbol{\lambda}$. Using the precalculation steps described in Section 2.3.2 in combination with the sparse matrix algorithms, we now show that the computational burden of obtaining optimal smoothness parameters is greatly reduced, without any substantial loss of accuracy.

5.2.2 Validation of the trace approximation

The approximation of $\mathsf{tr}(\boldsymbol{H})$ that is discussed in Sections 2.3.2 and 5.2.1 must first be checked for accuracy and computational feasibility before being used in the simulation study that follows. To do this, three features of the approximation will be tested: 1) the percentage error associated with selecting a smoothing parameter using the approximation compared to performing the exact calculation of the trace; 2) the time taken to select a smoothing parameter as the number of samples s involved in making the approximation increases; 3) the difference in computation time taken to select a smoothing parameter for different sized stream networks.

In order to test these three features, it will be necessary to generate data from different realisations of a spatial process on stream networks with different numbers of stream segments. As will be discussed in greater detail in Section 5.4.1, the software package `SSN` contains functionality that makes this straightforward to do. Data were generated 500 times from moderately smooth tail-up processes (also described in further detail in Section 5.3) on stream networks with 500 stream segments, and for each of these, stream network models were fitted to select the optimal smoothing parameter based on 10, 20, 50, 100, 500, and 1000 samples of the trace matrix. The results associated with this simulation experiment are shown in the top panel of Figure 5.1, where the distribution of percentage differences between the optimal λ selected using the trace approximation and the exact trace are plotted against the number of samples involved in the approximation. It should be noted that the percentage differences were calculated on the original scale of the smoothing parameter λ, and not on the log transformed scale, and therefore even in the least accurate scenario in which only $s = 10$ samples were used, more than 50% of the estimated differences are within the range of $(-7\%, 7\%)$ which are all deviations small enough to only have a very small impact on the resulting parameter estimates. As the number of samples increase, the error rate quickly becomes negligible, and values beyond $s = 100$ are likely to be unnecessary. It is therefore clear that for even small values of s the trace approximation is accurate enough for use in smoothing parameter selection.

In addition to comparing relative errors in the selected smoothing parameter, it is possible to evaluate 2) by comparing the relative computation times associated with different numbers of random samples, s, used in the trace approximation. In the middle panel of Figure 5.1, the ratio of the time taken to select a smoothing parameter using the approximation to $\mathsf{tr}(\boldsymbol{H})$ to using the exact calculation is shown. For $s \leq 100$, the computation time associated with the trace approximation is shown to be much faster cutting the time by between 30% and 50%. For $s = 500$ and above, the approximation is much slower, but as mentioned above, the increased accuracy that this buys is not likely to be necessary to obtain an optimal smoothing parameter under the Nelder-Mead optimisation

Finally, it is important to test the relative computational cost associated with using the trace approximation compared to using the exact trace calculation, for each of a range of stream network sizes in order to address 3). 500 simulations of spatial data were generated on large stream networks with 100, 500, 1000 and 2000 stream segments, and the time taken to obtain an optimal value for the smoothing parameter was recorded for models using the trace approximation, and for models using exact evaluations of the hat matrix trace. For each of these scenarios, the number of samples s was fixed at 100. The results of this simulation study are shown in the bottom panel of Figure 5.1, where the ratio of computer times associated with using the trace approximation compared to using the exact trace is plotted alongside the number of stream segments involved in each scenario. It is clear that on average, the trace approximation yields much faster selection, and is slower only occasionally on the smallest of the networks. It is also interesting to note that as the stream network increases, the ratio of the speed of the approximation relative to the exact calculation also increases, which makes this approximation particularly useful for large river networks.

5.2.3 `smoothnetwork` software and data format

The techniques described in Chapter 4 and smoothness selection as discussed in Section 5.2.1 have been integrated into the `R` software package `smoothnetwork`. A key feature of this package is the ability to fit additive models based on P-spline smoothing, and river network spatial effects using an interface that is similar to that used in the generalised additive model fitting software `mgcv` (Wood (2006)). In particular, specification of smooth functions of covariates is made straightforward, by adopting a function similar to the `s` function used in `mgcv`. To avoid conflicts with `mgcv`, we use a function called `m` for specifying smooth functions within the top level function `smnet`. For example, given a response vector $\boldsymbol{y}$, and covariate vector $\boldsymbol{x}$ each of length n, to fit the univariate additive

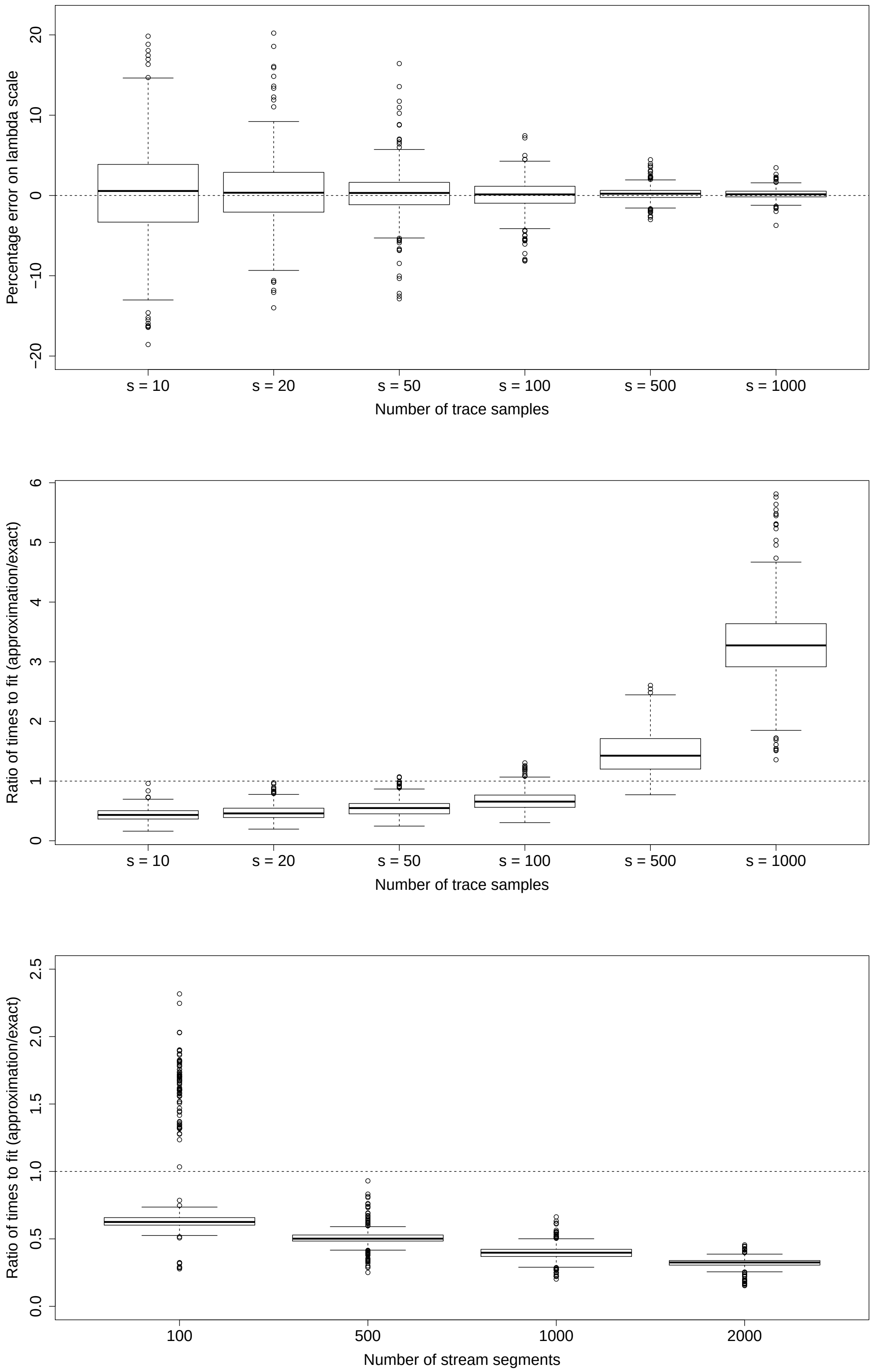


FIGURE 5.1: Top: Percentage difference in estimated smoothing parameter resulting from exact and approximate hat matrix traces for a range of approximation accuracy levels. Middle: Time taken to select an optimal smoothing parameter across a range of approximation accuracies. Bottom: Time difference (s) between smoothing parameter selection using exact and approximate matrix traces across a range of stream network sizes.

model

$$\begin{aligned} y_i &= \alpha + m(x_i) + \varepsilon \\ &= \alpha + \sum_{j=1}^{k} b_j B_j(x_i) + \varepsilon \end{aligned}$$

where k is the number of B-spline basis functions that are to be used to represent the smooth m, the statement

```
smnet(y ∼ m(x, k = k), data = df).
```

is used. The data object containing the covariate `x` and the response variable `y` is the data frame `df`. More complex models can also be fitted, for example if a response $\boldsymbol{y}$ is available at a set of geographical locations, $(\texttt{easting}, \texttt{northing})$. Suppose in addition that a vector of stream segments $\boldsymbol{s}$ is available that describes the network location upon which each observation arises, where $s_i \in \{1, \ldots, p\}$. Then the more complex spatial river network model

$$\begin{aligned} y_i &= \alpha + m(\texttt{easting}_i, \texttt{northing}_i) + \beta_{s_i} + \varepsilon \\ &= \alpha + \sum_{j=1}^{k} \sum_{l=1}^{k} b_{jl} B_j(\texttt{easting}_i) B_l(\texttt{northing}_i) + \beta_{s_i} + \varepsilon, \end{aligned} \tag{5.1}$$

where β_{s_i} is a stream segment specific parameter for observation i, defined with an associated penalty as described in Section 4.4.1. Equation 5.1 describes a model that allows spatial variation both in Euclidean space as indexed by the eastings and northings of each observations, and spatial variation that respects the network structure. In order to accommodate the latter network component within the `smoothnetwork` software, the additional operator `network` is used, and the model is specified as follows

```
smnet(y  ~  m(easting, northing, k = 10)
     + network(adj = adj, wgt = wgt, locs = s),
     data = df)
```

where `adj` is a $p \times p$ binary adjacency matrix whose i^{th} row describes the stream segments directly upstream of segment i. `wgt` is the vector of length k describing the proportion of flow each stream segment contributes to its downstream neighbour and `s` is a length k vector of integers describing the stream segment from which each observation y_i was taken.

5.3 Models based on covariance functions

5.3.1 Moving average construction

As discussed in Section 5.1, Ver Hoef *et al.* (2006) show that the use of a standard geostatistical model, using as a separation metric distance along the paths of the stream network, does not in general result in valid covariance structures, except when an exponential covariance function is used. As a result, Ver Hoef *et al.* (2006) and Cressie *et al.* (2006) seek to build appropriate covariance models through the use of moving average constructions. This construction states that a random variable Z can be defined as the convolution of a moving average function g and a white-noise process W so that

$$Z(s|\boldsymbol{\theta}) = \int_{-\infty}^{\infty} g(x - s|\boldsymbol{\theta}) dW(x), \tag{5.2}$$

where x and s are locations on the real line. Z then has covariance defined in terms of the choice of moving average function g:

$$\text{Cov}(Z(s), Z(s+h)) = C(h|\boldsymbol{\theta}) = \int_{-\infty}^{\infty} g(x|\boldsymbol{\theta})g(x-h|\boldsymbol{\theta})dx. \tag{5.3}$$

Since the argument x is defined on the real line in Equations 5.2 and 5.3, some additional work is required to adapt the process Z and the moving average function g to the context of a stream network, which is a set of connected line segments embedded in R^2. This is a more complex domain and so it is necessary to first establish some terminology that defines the topology, locations, and distance metrics that make up a stream network, and we adopt the nomenclature used in Ver Hoef *et al.* (2006).

5.3.2 Terminology

First, an enumeration of each of the segments of a stream network $i \in I = \{1, 2, \dots, n_{\text{seg}}\}$ is defined, where a segment is defined as the stretch of stream between branching or confluence points on the network (Figure 5.2). Locations on different segments may share the same upstream distance, defined as the distance to the most downstream location on the network, and so each location is uniquely expressed by x_i where x is the distance upstream of a point on the i^{th} stream segment. The most downstream location on the i^{th} stream segment is called l_i, while the most upstream location is called u_i, an example of this naming structure is given in the middle panel of Figure 5.2. It is then useful to define the set of stream segments upstream of, and including a location i, which we call $U_i \subseteq I$, while the set that *excludes* i is called $U_i^* \subseteq I$. In a similar fashion, the set of stream segments downstream of, and including i are called $D_i \subseteq I$, and $D_i^* \subseteq I$ if the set excludes i. This notation is necessary to formally define the concept of flow-connectivity, namely that $U_i \cap U_j \neq \emptyset$ implies that the stream segments i and j are *flow-connected*, conversely if $U_i \cap U_j = \emptyset$ then i and j are *flow-unconnected*. These definitions are also required to specify the valid covariance models over the spatial stream-network that are described next in Section 5.3.3.

5.3.3 The tail-up model

A stream network can be represented as a collection of connected line segments, and therefore to adapt Equation 5.3 to this setting, the integral must be performed in a piecewise manner over these line segments. One of the most important features of a stream network is the influence of flow direction on spatial dependence and as a consequence Ver Hoef *et al.* (2006) chose moving average functions $g(x|\boldsymbol{\theta})$ that are defined only upstream of a given location x. Ver Hoef *et al.* (2006) also recognise that spatial dependence on a stream network is determined by the relative flow volumes that meet at confluence points, a feature that is incorporated into the definition of the moving average function by dividing g at confluences relative to the proportions associated with the stream segments involved. This results in a scaling of the new segment-wise integral by a weight ω_k

$$Z(s_i|\boldsymbol{\theta}) = \int_{s_i}^{u_i} g(x_i - s_i|\boldsymbol{\theta})dW(x_i) + \sum_{j \in U_i^*} \left(\prod_{k \in B_{i,j}} \sqrt{\omega_k} \right) \int_{l_j}^{u_j} g(x_j - s_i|\boldsymbol{\theta})dW(x_j), \quad (5.4)$$

where $B_{i,j} = D_j \cap D_i$. If a confluence upstream of segment i has upstream segments j and k, then $0 \leq \omega_j, \omega_k \leq 1$ and $\omega_j + \omega_k = 1$. A visual representation of a moving average function is provided in Figure 5.2 which shows how the moving average function 'splits' at confluences, and how locations further upstream have little influence on s_1. To avoid truncating the moving average function, terminal stream segments (those furthest upstream) are treated as having infinite length. The construction in Equation 5.4 implies non-zero covariance between $Z(s)$ and $Z(s + h)$ if they are flow-connected, which is particularly desirable when the observed data are strongly dependent on flow, as may be the case with the concentrations of dissolved pollutants. Ver Hoef *et al.* (2006) show that Equation 5.4 implies a covariance between a pair of locations (r_i, s_j) that is defined as

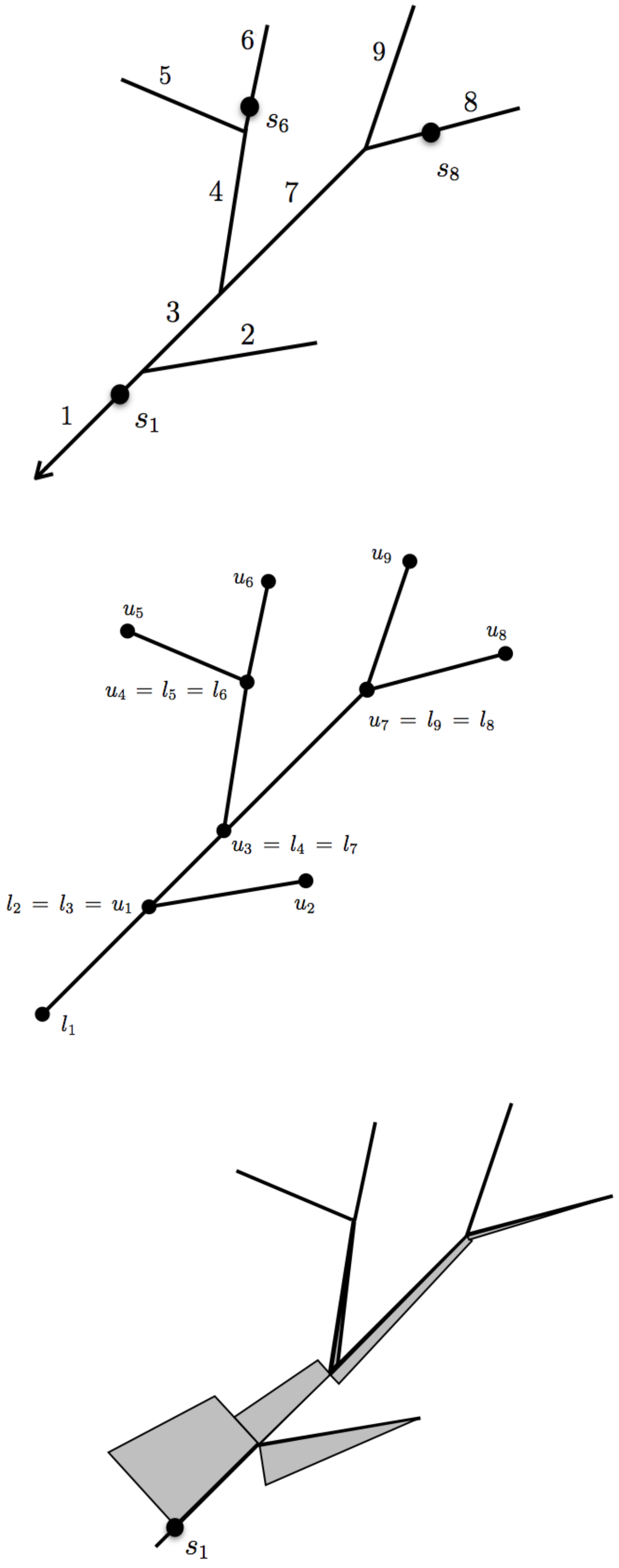


Figure 5.2: From top: example of an enumeration of the stream segments in a stream-network, with locations shown on stream segments using the notation $s_\bullet$, direction of flow is indicated by the arrowhead; illustration of the naming conventions for the beginning and end points for each stream segment; visual representation of a moving average function where the height of each rhombus corresponds to the relative size of the function, and decreases with distance from the example location s_1.

$$C_u(r_i, s_j|\boldsymbol{\theta}) = \begin{cases} \pi_{i,j} C_t(h|\boldsymbol{\theta}) & \text{if } r_i < s_j \text{ are flow-connected,} \\ 0 & \text{if } r_i \text{ and } s_j \text{ are flow-unconnected,} \end{cases} \tag{5.5}$$

where $\pi_{i,j} = \prod_{k \in B_{i,j}} \sqrt{\omega_k} \in [0, 1]$ represents the influence of the intervening flow weights on the covariance between flow connected r_i and s_j. Equations 5.5 and 5.4 define a model which is referred to as a 'Tail-up' model by Ver Hoef & Peterson (2010b), in order to distinguish it from moving average constructions that permit non-zero covariance between flow unconnected locations.

Different choices are available for the moving average function g, resulting in a process Z with different covariance properties. For example the exponential moving average function is defined as

$$g(x|\boldsymbol{\theta}) = \theta_1 \exp(-x/\theta_r) I(0 \leq x),$$

where θ_r is a range parameter. The moving average function in Equation 5.6 yields the unweighted covariance function

$$C_t(h|\boldsymbol{\theta}) = \theta_v \exp(-h/\theta_r),$$

where θ_v is the 'partial sill' parameter and is a function of θ_1 and θ_r. Other choices for the moving average function and their associated covariance functions are described in Ver Hoef & Peterson (2010b).

5.4 Outline of the simulation study

5.4.1 Simulating network spatial structure

In order to compare the two stream network modelling methodologies, it is important to simulate data with structures similar to those found in typical stream networks. Since real stream networks are often very large, with sparse data coverage, these are features that are likely to have a substantial impact on the performance of stream network models, and must form a key component of the study. In addition, realistic dendritic network structures do not permit divergence or bisection of stream segments, and confluences are restricted to having at most two upstream tributaries. The `SSN` software package has functionality that makes constructing the desired network structures of a specified number of stream segments straightforward, and as a result the impact of both small $n_{\texttt{seg}} = 100$ and large $n_{\texttt{seg}} = 2000$ is investigated. However, constructing valid dendritic networks with several hundred or more branches is a highly computationally intensive procedure, and so this operation was performed once at the outset of the study, and these fixed network structures were used to generate new realisations of spatial processes at each stage of the study. The two particular networks that were used are shown in Figure 5.3, with the smaller network of 100 segments shown in the left panel and the larger with 2000 segments on the right.

It is of interest to investigate the impact of the number of observed data points assumed for each simulation on the relative performance of each model. Therefore, observations were assumed to be made at each of $n = 100$ and $n = 500$ locations, where the locations were sampled without replacement from a large set of 10000 randomly distributed points that were generated once to reduce the computational burden. A key component of the study is in comparing predictive performance, and so an additional 1000 prediction locations were also generated on each network structure, which were used to assess out-of-sample predictive accuracy.

5.4.2 Spatial correlation structure of simulated data

In order to build appropriate spatial covariance into data that is simulated at the locations shown in Figure 5.3, the spatial process defined by the exponential Tail-up model (TU) described in 5.3 was used. Other covariance functions could have been used within the

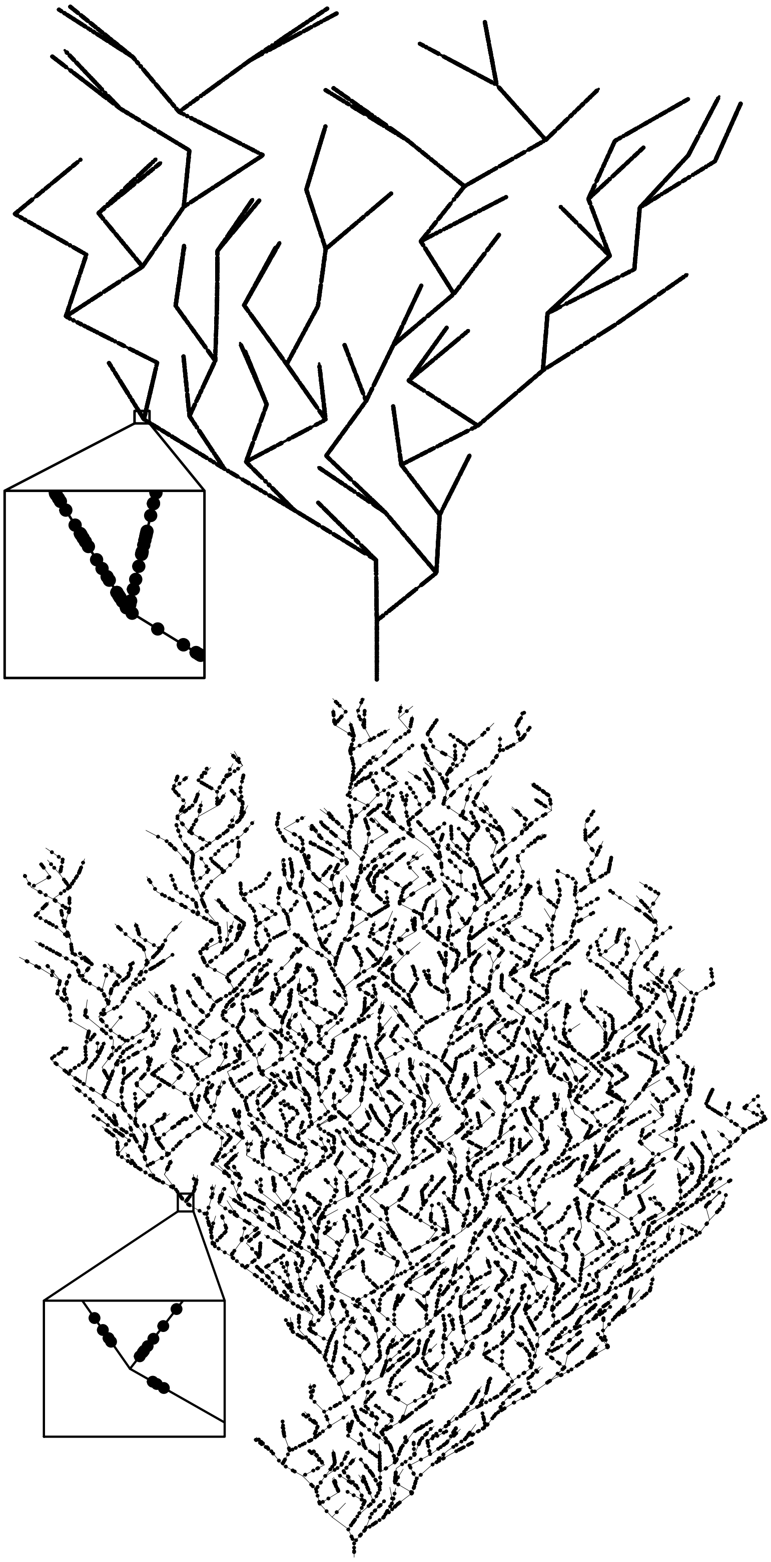


FIGURE 5.3: Plots depicting the two particular network structures on which all of the simulated data were generated: on the top is the smaller network with $n_{\mathsf{seg}} = 100$ stream segments and on the bottom is the large network with $n_{\mathsf{seg}} = 2000$ stream segments. The prediction locations are shown by black points on each network, magnified examples of which are shown in the insets.

class of TU models, but to maintain a feasible number of different factors in the simulation experiment, the effect of generating data from different covariance functions was not investigated. The impact of response variables that exhibit Euclidean spatial dependence as well as TU-type dependence across the stream-network was also investigated; this is important because spatial structure in stream-network data can arise as the result of a mixture of processes, some of which occur on the network and others in the terrestrial landscape within which the network is embedded. For a fixed set of network locations $\boldsymbol{s} = (s_1, \ldots, s_n)^\top$ with corresponding Cartesian coordinates $\boldsymbol{C} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]$, the process Y, was simulated as Gaussian with mean depending on covariates $\{X_1, X_2, X_3\}$, and a set of spatial processes $\{Z_1, Z_2\}$, representing TU and Euclidean structures, respectively. Y can be expressed as

$$\begin{aligned}
Y(\boldsymbol{s})|Z_1, Z_2 &\sim N\left(\beta_0 \mathbf{1}^\top + k_1 \sum_{i=1}^{3} \beta_i X_i(\boldsymbol{s}) + Z_1(\boldsymbol{s}) + k_2 Z_2(\boldsymbol{s}), \sigma^2 \mathbf{I}\right), \\
Z_1(\boldsymbol{s}) &\sim N(\mathbf{0}, \boldsymbol{\Phi}), && (5.6) \\
Z_2(\boldsymbol{s}) &\sim N(\mathbf{0}, \boldsymbol{\Psi}), && (5.7) \\
\boldsymbol{\Phi}_{ij} &= \begin{cases} \pi_{ij} \exp\left(-\frac{|s_i - s_j|}{r_\Phi}\right) & \text{if } s_i \text{ and } s_j \text{ are flow-connected,} \\ 0 & \text{otherwise} \end{cases} \\
\boldsymbol{\Psi}_{ij} &= \exp\left(-\frac{||\boldsymbol{c}_i - \boldsymbol{c}_j||}{r_\Psi}\right)
\end{aligned}$$

where $|s_i - s_j|$ was the stream distance between locations s_i and s_j, r_Φ and r_Ψ were range parameters for each spatial process, and π_{ij} was a set of weights determined by the number and influence of branches between locations s_i and s_j. $\boldsymbol{c}_i$ represents the ith row of $\boldsymbol{C}$. The processes $\{X_1\}$, $\{X_2\}$, and $\{X_3\}$ involved in the linear component $\beta_0 \mathbf{1}^\top + \sum_{i=1}^{3} \beta_i X_i(\boldsymbol{s})$, were each based on moderately smooth TU spatial structures, to simulate spatially patterned covariate effects. To simulate the effect of observing a variable that is not related to the response variable, we set $\beta_3 = 0$. The remaining variables $\{X_1\}$ and $\{X_2\}$ were set to have a significant association with Y with coefficients $\beta_0 = \beta_1 = \beta_2 = 1$. In order to simulate unobserved confounding, $X_2(\boldsymbol{s})$ was assumed unobserved, and was not included

Parameter	Levels	Interpretation
k_1	$\{0.1, 1\}$	Strong/weak linear effect
k_2	$\{0, 1\}$	Presence/absence of Euclidean structure
θ_v	$\{0.3, 1\}$	Long/short range TU structure
n	$\{50, 500\}$	Small/large number of observations
n_{seg}	$\{100, 2000\}$	Small/large network

Table 5.1: Summary of parameters and associated levels involved in the simulation study design.

in model fitting.

We varied the spatial components (TU and Euclidean), fixed effects, and covariance parameters, as well as the number of network segments and observations (Table 5.1)to generate a total of 32 different simulation scenarios. The Euclidean component, Z_2, was specified by a fixed range of $r_\Psi = 0.3v$ where v was the maximum separation between points on the network, with partial sill of 1. The binary control parameter k_2 denoted the presence or absence of Euclidean spatial structure in the data-generating process for a given simulation scenario. For the TU component, Z_1 the range parameter r_Φ was set to take two possible values, $\{0.3v, v\}$, in order to simulate spatial network structures with both long and short range dependence. The parameter k_1 scales the strength of the spatial component relative to the linear component and was given the values $\{0.1, 1\}$. All TU and Euclidean components require nugget and partial sill parameters that we fixed at 0.1 and 1, respectively.

The choices of spatial structure detailed here combined with the spatial structures of the observation locations outlined in Section 5.4.1 give rise to a total of 32 different simulation scenarios. A summary of the parameter choices for each of the scenarios is shown in Table 5.1.

5.4.3 Model fitting and measuring performance

In addition to fitting the models of VHPT and OD, it was also desirable to compare each of the model's relative performance to some baseline model. This model could take a number of forms, although for simplicity, we used a standard linear regression model that disregards residual spatial dependence. The three models that were fitted to the vector of observations $Y(\boldsymbol{s})$, when only the TU spatial structure was simulated were

$$\begin{aligned} &\textsf{VHPT} \quad \beta_0\mathbf{1}^\top + \beta_1 X_1(\boldsymbol{s}) + \beta_3 X_3(\boldsymbol{s}) + Z_1(\boldsymbol{s}) + \boldsymbol{\epsilon}, && (5.8)\\ &\textsf{OD} \quad \beta_0\mathbf{1}^\top + \beta_1 X_1(\boldsymbol{s}) + \beta_3 X_3(\boldsymbol{s}) + \boldsymbol{\beta_S} + \boldsymbol{\epsilon}, && (5.9)\\ &\textsf{Linear} \quad \beta_0\mathbf{1}^\top + \beta_1 X_1(\boldsymbol{s}) + \beta_3 X_3(\boldsymbol{s}) + \boldsymbol{\epsilon}, && (5.10)\end{aligned}$$

where $\boldsymbol{\epsilon}$ is independent $N(\mathbf{0}, \sigma^2\boldsymbol{I})$ and Z_1 is a TU spatial process with unknown sill (θ) and range parameters (r) as described in Section 5.3. The spatial component $\boldsymbol{\beta_S}$ in (5.9) was constructed from an $n \times n_{\mathsf{seg}}$ binary stream segment membership matrix and vector of n_{seg} spatial parameters. When Euclidean spatial dependence was present in addition to TU, the following appended models were fitted

$$\begin{aligned} &\textsf{VHPT} \quad \beta_0\mathbf{1}^\top + \beta_1 X_1(\boldsymbol{s}) + \beta_3 X_3(\boldsymbol{s}) + Z_1(\boldsymbol{s}) + Z_2(\boldsymbol{s}) + \boldsymbol{\epsilon}, && (5.11)\\ &\textsf{OD} \quad \beta_0\mathbf{1}^\top + \beta_1 X_1(\boldsymbol{s}) + \beta_3 X_3(\boldsymbol{s}) + \boldsymbol{\beta_S} + m(\mathbf{x}, \mathbf{y}) + \boldsymbol{\epsilon}, && (5.12)\\ &\textsf{Linear} \quad \beta_0\mathbf{1}^\top + \beta_1 X_1(\boldsymbol{s}) + \beta_3 X_3(\boldsymbol{s}) + \boldsymbol{\epsilon}, && (5.13)\end{aligned}$$

where the VHPT model in Equation 5.11 includes an additional Euclidean spatial Z_2 process with exponential covariance function and unknown range and sill parameters. Similarly, the OD model in Equation includes a bivariate smooth term $m(\mathbf{x}, \mathbf{y}) = \mathbf{B}(\mathbf{x}, \mathbf{y})\gamma' = (\mathbf{B}(\mathbf{x}) \otimes \mathbf{1}_k) \odot (\mathbf{1}_k \otimes \mathbf{B}(\mathbf{y}))$ where $\mathbf{B}(\mathbf{y})$ and $\mathbf{B}(\mathbf{x})$ were B-spline basis matrices each with k knots; $[\boldsymbol{x}^\top, \boldsymbol{y}^\top]$ correspond to the vector of network locations $\boldsymbol{s}$ transformed back to Cartesian coordinates, and γ was a vector of basis coefficients also estimated by penalised least squares, where smoothness was controlled by a single control parameter.

5.5 Results

The performance measures resulting from summarising the model fits under each of the 3 models have been divided into 4 sets of plots. The first two show prediction performance - Figure 5.4 displays summaries of predictive performance for models (5.8), (5.9) and (5.10) fit to data with TU structure, while Figure 5.5 summarizes predictive performance for models (5.11), (5.4.3) and (5.13) fit to data with TU and Euclidean mixture covariance structures. Switching to estimation of fixed effects, Figure 5.6 shows performance summaries of models (5.8), (5.9) and (5.10) fitted to TU data, and Figure 5.7 shows performance summaries of TU/Euclidean mixtures fitted by models (5.11), (5.4.3) and (5.13).

For Figures 5.4 and 5.5, RMSPE, bias and 90% prediction interval coverage are plotted across the different spatial structures considered, which are identified on the x-axis, while results corresponding to sample sizes are shown by differing points and line types within each panel. Figures 5.6 and 5.7 differ from 5.4 and 5.5 only in that they illustrate performance associated with estimating fixed effects, and therefore the top two panels show RMSE rather than RMSPE. In this Section, the main findings regarding predictive performance are first described for all scenarios, followed by those for fixed effects estimation.

5.5.1 Predictive performance

Under all scenarios, the spatial models outperformed the linear model, and of these, VHPT performed slightly better than OD (Figure 5.4). Under the $n_{\text{seg}} = 2000$ and $n = 100$

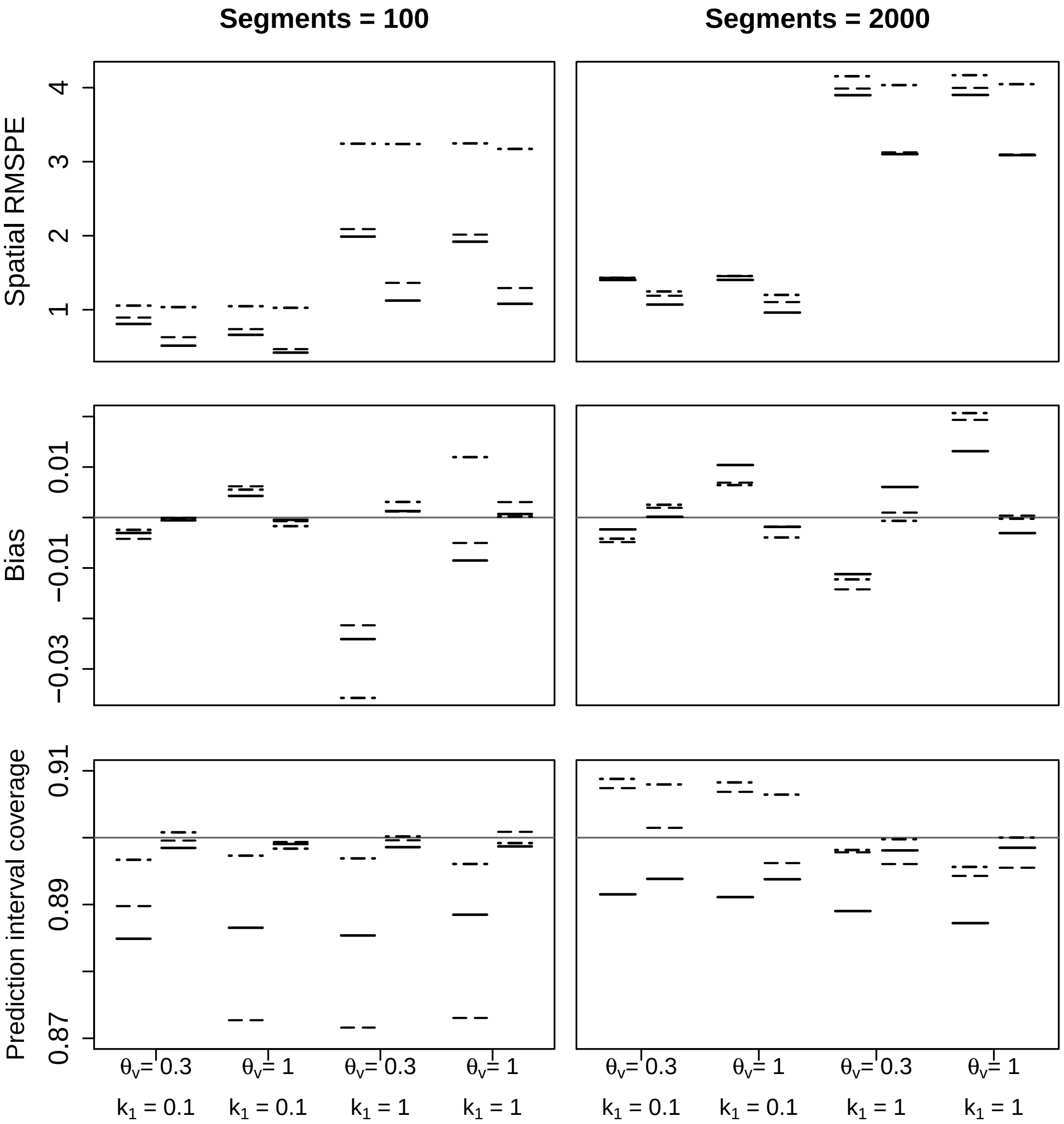


FIGURE 5.4: Predictive performance summaries for VHPT (solid lines), OD (long dashed lines) and linear models (dot-dashed lines) fit to data with TU spatial structure. The top panel shows the relative RMSPE for each technique, the middle shows the bias and the bottom shows the prediction interval coverage. The four x-axes index the different choices made in fixing the spatial structure of the simulated data. Values for $n = 100$ are shown to the left and $n = 500$ to the right above each parameter combination.

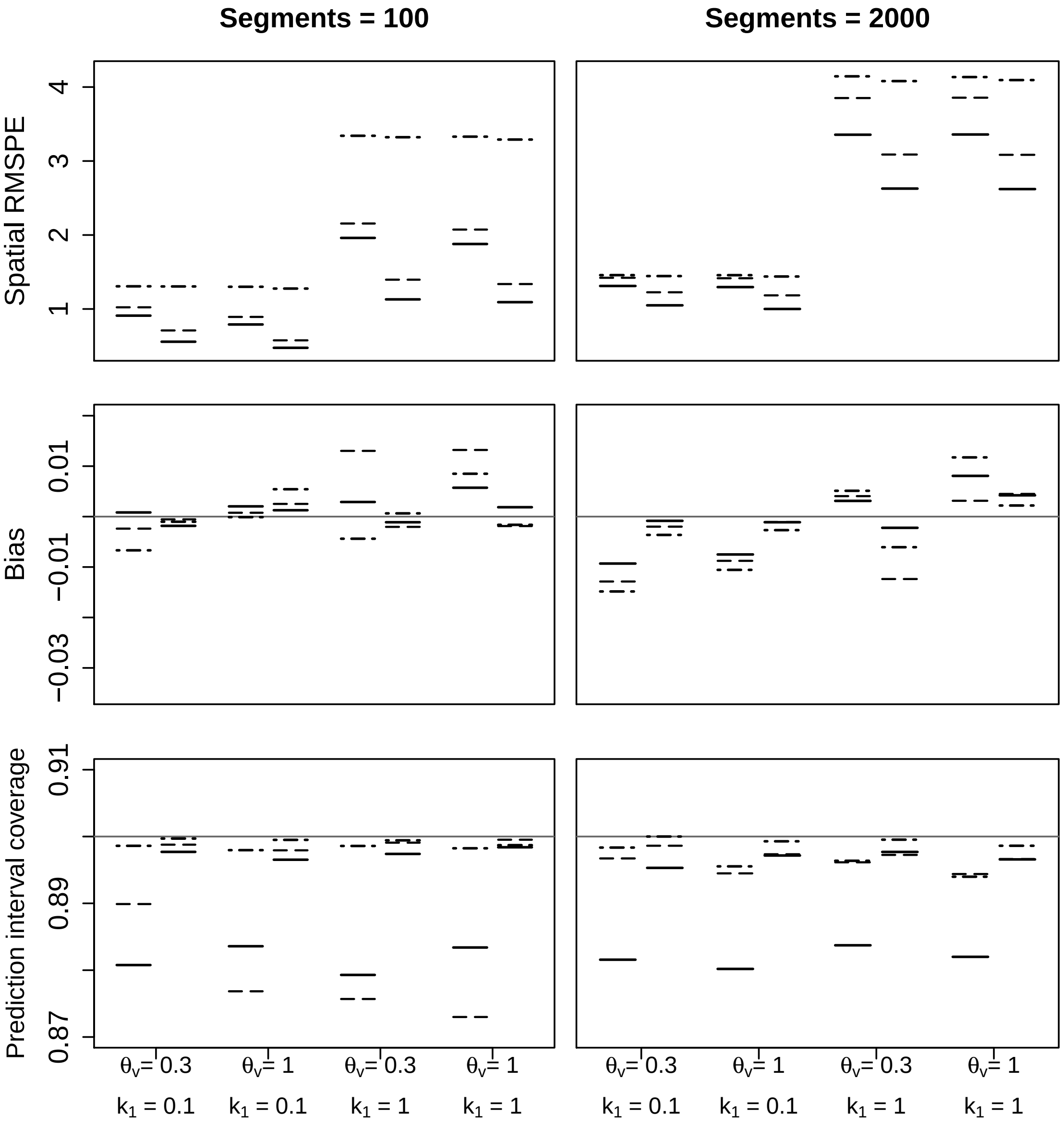


FIGURE 5.5: Predictive performance summaries for VHPT (solid lines), OD (long dashed lines) and linear models (dot-dashed lines) fit to data with TU and Euclidean mixture spatial structure. The top panel shows the relative RMSPE for each technique, the middle shows the bias and the bottom shows the prediction interval coverage. The four x-axes index the different choices made in fixing the spatial structure of the simulated data. Values for $n = 100$ are shown to the left and $n = 500$ to the right above each parameter combination.

scenario, the prediction error was almost equivalent for each of the models. Despite some differences in scale, broadly similar patterns are visible across the different simulated spatial structures; the lowest RMSPE is associated with data exhibiting a long spatial range and a weak linear component, while the highest RMSPE is associated with short spatial range and a dominant linear component.

In general, prediction bias is highest overall under the smallest sample size of $n = 100$ (Figure 5.4). Similar patterns are visible for all three models in the TU only scenario where the linear model performs only slightly more poorly than the others. When the Euclidean component is present, the relative performances (Figure 5.5) are less easily interpreted, although VHPT clearly achieves a lower level of bias than OD and the linear model. However, the bias is relatively small compared to RMSPE, and so all estimates can be considered unbiased.

When $n = 500$ most of the empirical interval coverages lie very close to the nominal level of 90% (Figures 5.4 and 5.5). However, when $n = 100$ and $n_{\texttt{seg}} = 100$, VHPT and OD achieve slightly lower coverages of between 0.87 and 0.89. It is notable that for the model of VHPT, this feature is only visible when $n_{\texttt{seg}} = 2000$.

5.5.2 Fixed effects estimation

Similar patterns in RMSE to those in RMSPE were found across each simulated spatial scenario (Figures 5.6 and 5.7), where VHPT performs the best, closely followed by OD and the linear model performs relatively poorly. An exception occurs when $n = 100$ and $n_{\texttt{seg}} = 2000$ where the error rates are essentially equivalent. Bias (middle panels of Figures 5.6 and 5.7) is low overall, and is roughly similar for each of the three models. As with prediction bias, the estimation bias is low relative to RMSE and indicates that these are

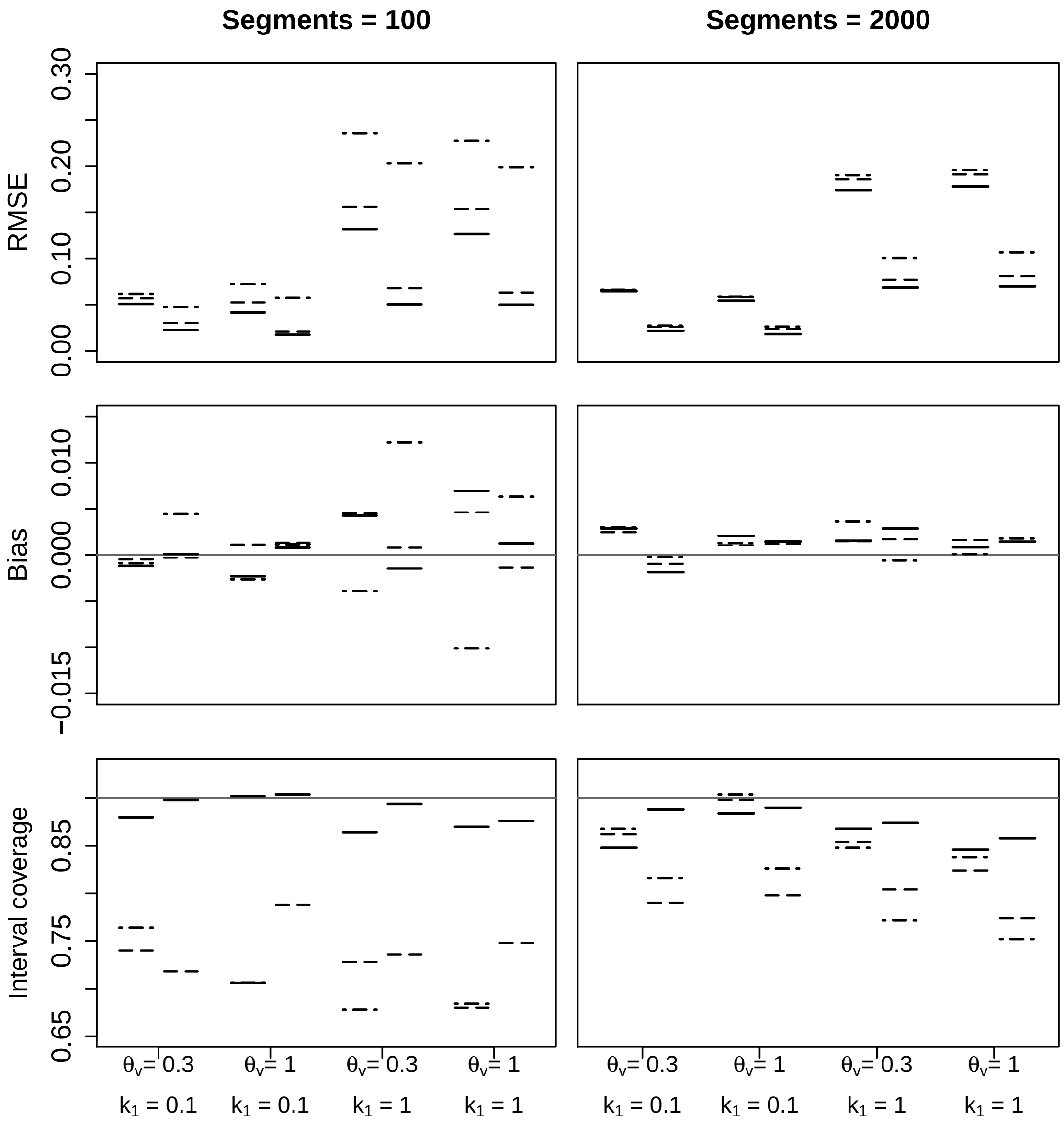


FIGURE 5.6: Summaries of estimation performance for VHPT (solid lines), OD (long dashed lines) and linear models (dot-dashed lines) fit to data with TU spatial structure. The top panel shows the relative RMSE for each technique in estimating the linear parameters, the middle shows the bias and the bottom shows the confidence interval coverage. The four x-axes index the different choices made in fixing the spatial structure of the simulated data. Values for $n = 100$ are shown to the left and $n = 500$ to the right above each parameter combination.

unbiased estimates.

For networks with $n_{\texttt{seg}} = 100$ (bottom left panels of Figures 5.6 and 5.7), empirical coverages for VHPT and OD approaches differ markedly: regardless of spatial structure and n, VHPT achieves the nominal 90%, while OD coverages are between 70% and 80%. For larger networks, the coverages occupy a narrower range of values around the target and VHPT are closest to the 90%. Interestingly, when $n = 100$, OD performs better than when additional data are present.

5.5.3 Computation

In addition to the empirical properties of estimation and prediction, we compared the relative computation times for each of the spatial models. The time taken for the VHPT model increases rapidly with n, whereas larger models can be fit with OD relatively quickly (top panel of Figure 5.8). The time taken to fit the model of VHPT remains constant across $n_{\texttt{seg}}$, while for the model of OD, the time to fit increases with network size, although the time taken is still less than under VHPT (bottom panel of Figure 5.8).

5.6 Discussion

In this chapter, two different approaches to modelling stream networks have been compared across a wide variety of simulated data. Across most scenarios, it was found that the the models of VHPT are most flexible and attain the highest levels of predictive accuracy while maintaining good estimation of the fixed effects terms, although in certain cases, OD perform comparably and for a much lower computational cost.

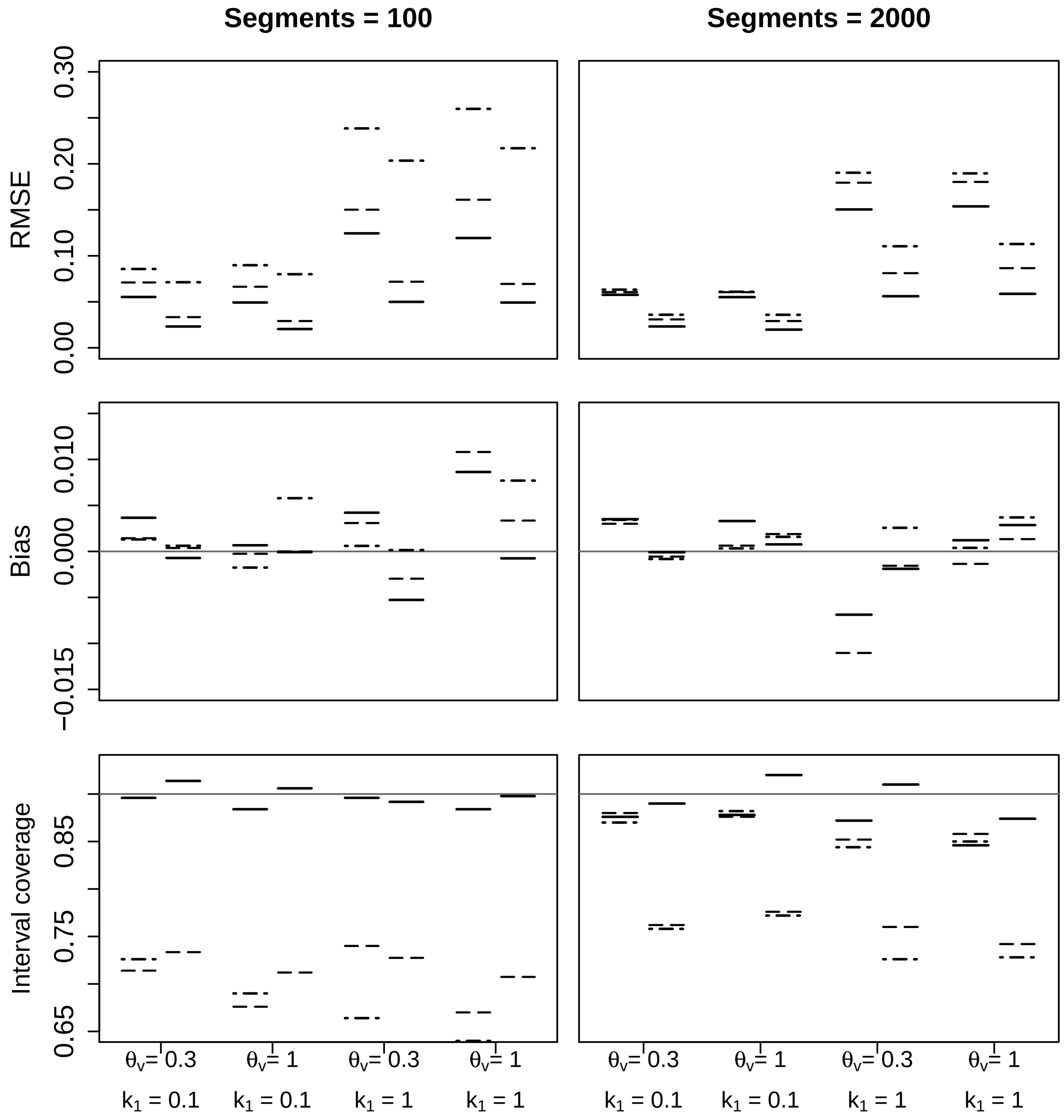


Figure 5.7: Summaries of estimation performance for VHPT (solid lines), OD (long dashed lines) and linear models (dot-dashed lines) fit to data with TU and Euclidean mixture spatial structure. The top panel shows the relative RMSE for each technique in estimating the linear parameters, the middle shows the bias and the bottom shows the confidence interval coverage. The four x-axes index the different choices made in fixing the spatial structure of the simulated data. Values for $n = 100$ are shown to the left and $n = 500$ to the right above each parameter combination.

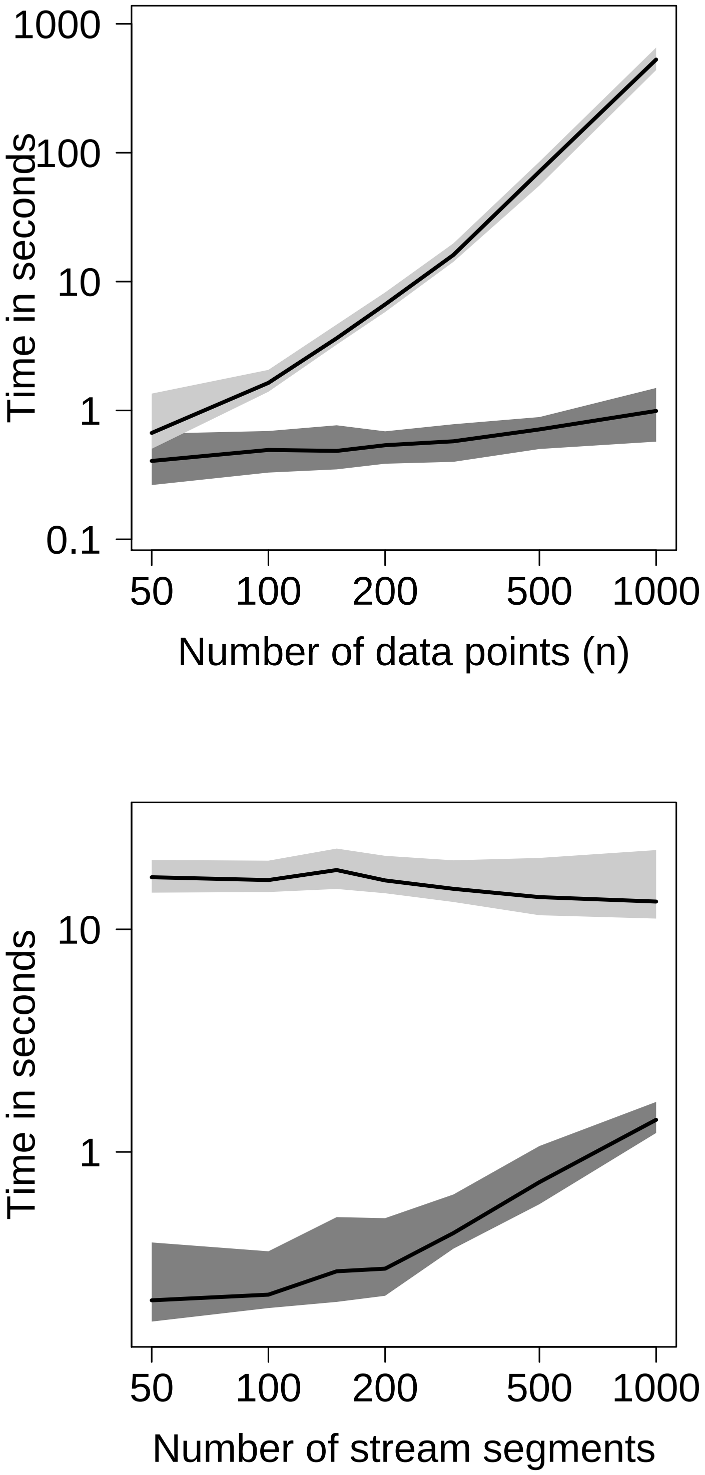


Figure 5.8: Plots showing how computation time scales with number of data points n (top panel) and with number of stream segments n_{seg} (bottom panel). The shaded regions contain the computation times between the upper 95% percentile and the lower 5% percentile where the dark grey regions correspond to the models of OD and the light grey regions to VHPT.

From the outset we aimed to answer the question, *when is each model appropriate*? The answer to this question as illustrated by the size of the study presented and from the results, cannot be straightforward and must depend on the particular data under consideration and the goals of the modeller. The interpretation of the results of the study presented here is nuanced because the tail-up model of VHPT under comparison is the true model for the simulated data, and therefore the relative performances of the three models may not generalise to all real applications. The decision to simulate from this particular covariance model reflects the type of spatial structures that are typical in practice, and recognises that the SSN software is a convenient way in which these otherwise highly complex structures can be readily simulated.

Despite this, some general guidance can be drawn from the study, for example networks in which more information is available about dependence *within* stream segments, for example, scenarios in which lots of data are available or the network is not very large, the moving average approach offers a more flexible and realistic description of the spatial structure of the stream network. This contrasts with the situation in which the ratio of stream segments to data points is much larger, so that little information on *within* stream segment dependence is available, and the predictive accuracy of both models is almost the same. Under these types of scenarios in the study, spatial dependence is driven entirely by the confluence mixing and associated flow weights, and such structures form the conceptual underpinning of the models of O'Donnell *et al.* (2013) and so it is therefore to be expected that the two models are roughly equivalent. As might be expected, when the data consist of few data points on a very large network, good estimation and prediction is strongly inhibited, and this property is reflected by the near-equivalence of parameter estimates and prediction errors for each of the models that were fitted under these conditions.

Although differences in methodology drive the differences in performances observed in this work, it is also the flexibility of the supporting software that can make particular approaches more or less attractive for modelling. It is therefore important to emphasise that the comparison presented is necessarily partial, and that the two frameworks admit desirable extensions, some of which are exclusive of one another. For example, `SSN` allows the fitting of *generalised* spatial models with a variety of commonly used link functions, and also permits the fitting of mixtures of spatial covariance structures including the recently developed 'Tail-down' construction. On the other hand, the methodology described by O'Donnell *et al.* (2013) was primarily concerned with capturing non-linear functional relationships between covariates, and the software, `smoothnetwork`, allows the fitting of additive models based on P-splines, in addition to the network construction described here. This feature can be particularly useful for capturing spatial and spatio-temporal effects.

Chapter 6

Conclusions and extensions

6.1 Summary of thesis contributions

This thesis has been primarily concerned with developing more flexible ways to represent the complex hydrological and meteorological relationships that underpin rainfall, flow and dissolved pollution observed within a river catchment. A common feature of the relationships that have been investigated is that they do not appear to adhere to any *a priori* functional form, and this thesis illustrates the utility of smoothing techniques in such contexts. Semiparametric models, which are summarised in Chapter 2, are supported by a well established literature and are powerful devices for exploring non-linearity between variables in a regression context. The work documented promotes the use of P-splines as a framework for fitting flexible functions, because of the simplicity involved in setting up a uniformly spaced basis, the conceptual intuitiveness of applying difference penalties on this basis to attain appropriate smoothness, and the high level of computational convenience that can often be exploited. In particular, this thesis emphasises computational simplicity and therefore presents a suite of methods that are appropriate for exploring the large data sets that are increasingly common in environmental monitoring and research.

6.1.1 Distributed lag models for hydrological data

The work of Chapter 3 focuses on the problem of representing the time-lagged dependence between high frequency rainfall and flow time series. Previous approaches to modelling can be roughly separated into two types: the first involves the use of complex deterministic representations of the physical environment that affect river flow, while the second constructs statistical representations of the data generating process, drawing on ideas from traditional time series analysis. Time series approaches can be computationally convenient and are designed to represent uncertainty appropriately, both of which are often lacking from a mechanistic modelling approach. However, time series approaches suffer from being less powerful, because the level of detail required to capture the dynamics of flow generation are unlikely to be present in a small set of time series. Section 3.6 demonstrated how a novel statistical specification using flexibly structured, time-varying coefficient models can be successful in capturing hidden temporal structure in the relationship between rainfall and river flow. As a result of this flexible model, it was possible to identify that the degree of responsiveness of flow rates to preceding rainfall is noticeably amplified when the rainfall occurs during or after a period in which rainfall has been persistent. This effect is well known, and is driven by the changing proportion of rainfall that is converted to runoff and baseflow, which in turn is a complex function of the saturation of the surrounding land. This led to a refined distributed lag model which was constructed in Section 3.6.1 and was able to attribute some of the time variation in the rainfall-flow relationship to the impact of a slowly varying proxy variable, used to represent latent ground saturation processes.

Although the work of this chapter begins from a statistical viewpoint as others have, the models it describes are successful in accommodating a much wider range of temporal dependence structures than have been previously. The work highlights a novel application and extension of distributed lag models that are more commonly seen in short-term air pollution studies and often assume a fixed lag structure. An interesting possibility for extending distributed lag models fitted to hydrological data would be to synthesise a model in which deterministic non-linear structure could be combined with the statistical representation outlined in this thesis.

6.1.2 Flexible regression for river networks

Chapter 4 is devoted to the problem of constructing appropriate spatial models for data sampled on a river network. In particular, the work described how to make predictions at locations in space given a very sparse coverage of sampling points across the spatial stream network, and how to capture non-linear relationships between covariates and the measured variable of interest. The solution presented respects the importance of flow-connectivity and the influence of confluences and relative flow volumes, which are fundamental requirements for representing spatial dependence in stream network data. The spatial dependence was represented in the new model by imposing a quadratic penalty on neighbourhoods of parameters defined for each stream segment. The penalty incorporated all available information about flow connectedness, the locations of confluences and relative flow levels. Chapter 4 successfully described how to exploit the sparse matrices implied by the design and penalty matrices specification which results in very fast computation and low storage costs. The models developed were applied to data from the River Tweed, a very large dendritic network in South East Scotland, where the presence of complex spatio-temporal structure was identified.

The contribution made by Chapter 4 is important primarily because it is the first major attempt to embed a valid representation for spatial dependence over a stream network within an additive model, which is essential for capturing non-linear dependence that is present in environmental systems. The spatial component itself lends itself to fast computation compared to other approaches that are based on, for example, Gaussian processes which require the construction, multiplication and inversion of $n \times n$ covariance matrices, rendering models with more than $n = 10^3$ data points difficult to fit. Finally, as was shown in Section 4.5, having defined a representation of the spatial structure of the network, it is relatively straightforward to introduce non-separable space time structure using the appropriate tensor-product basis.

Despite the methodological contributions made by Chapter 4, the utility of the new model can only be assessed by a comparison with other existing techniques. In particular, the work described first in the pioneering paper by Ver Hoef *et al.* (2006) has been particularly important in the field of stream network modelling, and instrumental to the success of this approach has been the development of software that eases the process of extracting and processing data from a geographical information system into an appropriate format for modelling. Their software permits the fitting of a suite of Gaussian process models developed for stream network data, where carefully constructed covariance functions are adapted to the purpose of incorporating flow connectedness, direction and measuring spatial separation along the path of the river instead of standard Euclidean separations.

In recognition of the capability of the existing Gaussian process approach, Chapter 5 provides a direct comparison of the semiparametric model described in Chapter 4 with those of Ver Hoef *et al.* (2006). The comparison was made primarily by a very large empirical study of the relative performances of the models on some realistic simulated stream network data, constituting a model validation study that has not previously been undertaken for either framework. As a result of the comparison, Chapter 5 also describes user friendly software that was developed for implementing the models of Chapter 4, which is also critical for their adoption by a wider community of researchers. The chapter provides clear insight into the performance differences of the models, while also contributing an overall view of the individual performance of each model across a wide variety of data types. The comparison highlighted differences between the two approaches, most noticeably when the network is small and the data are large, in which the model of VHPT performs best. This outcome is as expected, since the penalised model approach does not permit variation within stream segments, and therefore has limited flexibility in cases where sampling has a dense enough spatial coverage that this can be detected.

6.2 A spatial distributed lag model for grid rainfall data

6.2.1 Rainfall RADAR

The work of Chapter 3 has shown that a suitably parameterised distributed lag function, can act as a discrete approximation to the true underlying time-lagged relationship that exists between a single time-series of rainfall at a single location in space and downstream river flow levels. Section 3.7, highlighted that this relationship is strongly influenced by spatial heterogeneity in the rainfall process, which cannot be represented by data obtained from a rain gauge at a single spatial location. Very sparse spatial sampling of precipitation is also typical in most regions in the UK, which means that the use of multiple rain gauges is unlikely to provide the required level of detail. Fortunately, other methods exist of measuring rainfall patterns at high spatial resolutions exist, such as that obtained using RADAR imaging.

Radio detection and ranging (RADAR) works by measuring the intensity and time delay in receiving a reflected pulse of radio waves and as a result is able to accurately estimate the location of objects. RADAR systems are widely used by the UK Meteorological Office to detect the location and intensity of rainfall events at a fine spatial and temporal resolution (UK Meteorological Office (2013)), and UK-wide data are available as part of the UK Meteorological Office NIMROD data set (UK Meteorological Office (2003)) at 1km^2 spatial resolution at 5 minute intervals. An illustration is provided in Figure 6.1 which displays part of a Meteorological Office rainfall RADAR image for a 5 minute snapshot over the

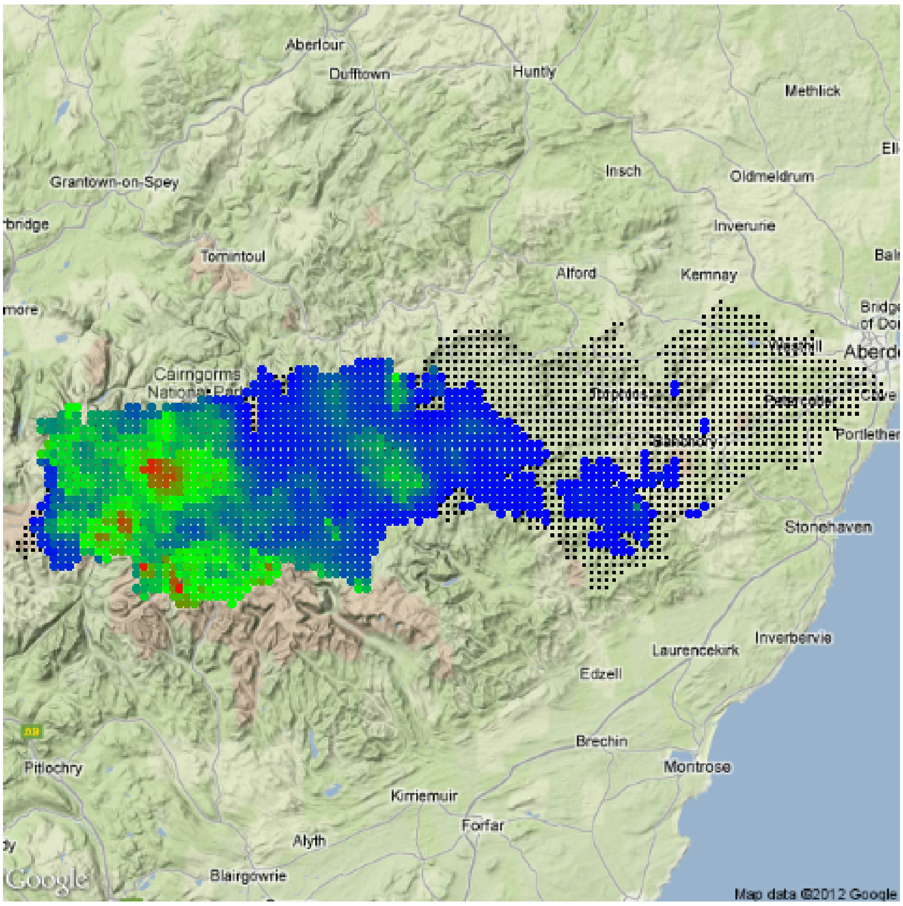


FIGURE 6.1: Rainfall RADAR snapshot over the River Dee catchment at 0015 on October 10th 2004. Black points represent 1km^2 cells across the River Dee where the RADAR observed no rainfall, while elsewhere the corresponding cells are coloured according to the observed intensity. Red points represent the highest intensity rainfall, green points represent moderate intensity and blue points represent low intesnity rainfall. Note that the mountainous west of the catchment experiences much higher precipitation than the east.

River Dee catchment in Scotland. Although by comparison to individual rain gauges, RADAR images appear to provide a very rich source of spatial and temporal information, there are a number of issues with such data which must be considered if these data are to be used for rainfall flow analysis.

1) RADAR measures reflectivity of layers of the atmosphere, and although efforts are made to calibrate the observed reflectivity against rain gauge data, the complexity of atmospheric meteorology means that there is no guarantee that the RADAR measures correspond to ground observations.

2) RADAR depends on clear line of sight in order to construct an accurate image, obstacles such as buildings and mountain regions inhibit this.

3) Rainfall and snowfall are not easily distinguished, and moderate snowfall has high reflectivity and appears as a heavy rainfall event.

6.2.2 A DLM for spatial rainfall data

An ideal model would allow non-linear, time-lagged dependence between each individual pixel from each of a set of rainfall RADAR images and a sequence of river flow measurements at a single location. With this type of model structure it could be possible to account for sudden flow events with functions of the intensity and distance of storms observed in the RADAR data. This level of spatial detail was not possible using the approach in Chapter 3, and is vital to build a realistic picture of the rain and flow dynamics in the area: for example, heavy rainfall occurring at points high in the River Dee catchment (shown in the left side of Figure 6.1) should take longer to appear as a flow increase downstream, than would a storm that is close to the river flow monitoring location. As a result, a better but more complex model would also allow *spatially-varying* time-lagged dependencies, to better account for spatial heterogeneity in water transit time across the catchment.

In order to allow greater flexibility in model specification, it is likely that a Bayesian analysis would be an appropriate tool to fit the complex models needed to represent the responsiveness of river flow to rainfall grid data. However, the models described here could still be fitted using the more traditional smoothing framework described throughout this thesis. Letting the time series of n flow observations be denoted by $\boldsymbol{f} = (f_1, \ldots, f_n)^\top$ and the pixels of the spatial rainfall grid by $\boldsymbol{r}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t}) = (r(x_1, y_1, t), \ldots, r(x_k, y_k, t))$ where $r(x_i, y_i, t)$ represents the rainfall intensity at grid cell with centroid (x_i, y_i), at time t. Following the ideas in the previous paragraph and directly extending the distributed lag models of Chapter 3, the expected flow levels $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ could be represented by a weighted sum of rainfall in space and time, up to a maximum time lag of L, such that

$$\begin{aligned} \boldsymbol{f}|\boldsymbol{\mu},\sigma^2 &\sim N(\boldsymbol{\mu},\boldsymbol{I}\sigma^2) \\ \mu_t &= \sum_{i=1}^{k}\sum_{l=1}^{L} r(x_i,y_i,t-l)\beta(x_i,y_i,l) \end{aligned} \tag{6.1}$$

where the k grid cells are those that lie in the area of land that drains to the point at which $\boldsymbol{f}$ is measured. Equation 6.1 describes flow levels as a linear combination of rainfall, where the weights $\beta(x_i, y_i, l)$ vary in space. Beven (2001) notes that this type of approach has been attempted, but that correlation between rainfall in space and time means it is difficult to identify the parameters associated with multiple transfer functions in space; however, with an appropriate penalisation included in the model specification using the flexible regression framework used throughout this thesis, it is possible that some progress could be made. As in Chapter 3, the dimensionality of the model implied by this specification could be prohibitive, requiring the estimation of potentially many β parameters. By reducing this large set of parameters to a smaller set of coefficients that are associated with a modest B-spline basis, the computations could be made feasible. Smoothness across adjacent β seems justified, because underlying properties of the drainage area in neighbouring grid cells are likely to share similar land uses, altitudes and gradients, and are therefore more likely to exhibit similar lagged rainfall-flow relationships, given the same rainfall input. To achieve this smoothness and reduce the number of parameters to estimate, $\beta(x_i, y_i, l)$ is represented by a 3-dimensional B-spline basis, defined in space and across lags so that

$$\ln[\beta(x_i,y_i,l)] = \sum_{a=1}^{k_1}\sum_{b=1}^{k_2}\sum_{c=1}^{k_3}\theta_{a,b,c}B_a(x_i)B_b(y_i)B_c(l) \tag{6.2}$$

where $B_a()$, $B_b()$ and $B_c()$ represent sets of evenly spaced B-spline basis functions of dimension k_1,k_2 and k_3, respectively. It is important to maintain the interpretation of a positive response function $\beta(x_i, y_i, l)$ existing at each location i, Equation 6.2 represents smoothness for the set of β on its log scale in order to maintain an interpretation in terms of a positive transfer function, i.e. it does not make sense to allow rainfall to *reduce* levels of river flow. Using the above reparametrisation, the problem has been reduced from estimating $k \times L$ parameters, to $k_1 \times\ k_2 \times k_3$. In order to ensure that the appropriate level of smoothing is chosen for $\boldsymbol{\theta} = (\theta_{1,1,1}, \ldots, \theta_{k_1,k_2,k_3})^\top$ and in turn β, a smoothness-inducing prior distribution should also be used:

$$\begin{aligned}
\boldsymbol{\theta} &\sim N(\mathbf{0}, \boldsymbol{Q}^{-1}) \\
\boldsymbol{Q} &= \lambda_1 \boldsymbol{D}_{k_1}^\top \boldsymbol{D}_{k_1} \otimes \boldsymbol{I}_{k_2 \times k_3} \\
&\quad + \lambda_2 \boldsymbol{I}_{k_1} \otimes \boldsymbol{D}_{k_2}^\top \boldsymbol{D}_{k_2} \otimes \boldsymbol{I}_{k_3} \\
&\quad + \lambda_3 \boldsymbol{I}_{k_1 \times k_2} \otimes \boldsymbol{D}_{k_3}^\top \boldsymbol{D}_{k_3}
\end{aligned}$$

where the prior precision matrix $\boldsymbol{Q}$ is constructed from Kronecker products of marginal penalty matrices $\boldsymbol{D}_{k_1}^\top \boldsymbol{D}_{k_1}$, $\boldsymbol{D}_{k_2}^\top \boldsymbol{D}_{k_2}$ and $\boldsymbol{D}_{k_3}^\top \boldsymbol{D}_{k_3}$ as described in Equation 3.1.

6.2.3 Issues and considerations

Computationally, the model described in Section 6.2.2 could still be very challenging to fit even after reducing the space and lag parameters to a set of basis parameters, and requires the estimation of $k_1 \times k_2 \times k_3$ parameters associated with this 3-dimensional basis. The model would be fitted using Bayesian inference, allowing more flexibility in how the model is specified, but means that some stochastic simulation algorithms such as Gibbs sampling will be required to draw samples from the relevant posterior distributions. On the other hand, some of the features that make the models in Chapter 3 easier to fit apply in this spatial context too, for example spatial rainfall is dominated by an abundance of 0 observations and so the data can be represented using sparse matrix algorithms.

Conceptually, the spatial DLM is a straightforward extension of the univariate DLM fitted to the River Dee data, but there are some reasons that the model proposed in Section 6.2.2 is naive. The spatial model assumes that, rainfall is converted to river flow according to a fixed response function specific to a particular point in space. Although some account is taken of the similarity between response functions at two closely located points, there is no representation of how these might interact. For example, it is likely that following heavy rainfall at high altitude the response functions at lower altitudes nearby are directly changed as a result of water running downhill and subsequently saturating low lying areas. Therefore, a feature of this future work should involve allowing for local interaction in addition to the smoothness already specified.

6.3 A continuous functional representation for river networks

Chapter 4 dealt with representing spatial smoothness on a river network through a set of penalised piecewise constants that join at confluence points. However, the study of model fit performance documented in Chapter 5 showed that in settings where high frequency spatial structure was present and enough data exists to detect it, the models of Ver Hoef *et al.* (2006) had much better confidence interval coverage and out-of-sample prediction than the penalised model, where a piecewise constant network structure was assumed. It would therefore be a highly desirable extension to Chapter 4 to build a model using a continuous basis representation of the spatial network, permitting the expected value of the spatial response at different locations on the *same* stream segment to lie on a smooth function. This natural extension has already been suggested by Cressie & O'Donnell (2010), who describe a need to build models that do not require the assumption of a stationary process as is the case with models based on covariance functions (such as Ver Hoef *et al.* (2006) and Cressie *et al.* (2006)). A functional approach based on continuous B-spline basis representation could allow relatively fast computation if the matrices remain sparse, and could therefore be computationally attractive. In moving towards this smooth model, there are three main challenges that must be overcome:

1) constructing an appropriate basis along stream paths

2) respecting the mixing behaviour at confluence points

3) setting up an appropriate roughness penalty on the basis coefficients.

While attempting to address each of these three points, the following section outlines a more advanced extension to the river network models described in Chapter 4 that could form a part of future research.

6.3.1 Basis functions on stream paths

One of the simplest ways to set up a basis could involve B-splines uniformly spaced over all branches of the river network using a basis expansion at each location defined as

$$(B_1(r), B_2(r), \ldots, B_p(r)) \tag{6.3}$$

where r is the distance upstream of a given network location from the river mouth. The advantage of this specification is that the spacing of the functions depends only on p, the chosen number of functions for the basis and can be adjusted so that each segment is represented by a minimum number of basis functions. In addition, the basis set is defined on the distance upstream r, and is therefore not dependent on the locations or features of particular stream segments in which the location resides. Ordinarily, the fitted value at a location resulting from this basis would be given by the linear combination $(\beta_1 B_1(r) + \beta_2 B_2(r) + \ldots + \beta_p B_p(r))$, however, this specification implies the same fitted value at different locations that share the same upstream distance r. This issue can be avoided most easily by simply allowing the vector of parameters $(\beta_1, \ldots, \beta_p)$ to vary across the network. In the following section, an attempt to address this issue while maintaining a relatively simple basis specification is sketched.

6.3.2 Representing mixing behaviour at confluences

Section 6.3.1 describes a simple way to define B-spline basis functions on the connected stream segments that make up the river network. The next step is to introduce an expanded set of parameters, and make some adjustments to the basis in Equation 6.3 in order to incorporate the influence of confluence points, particularly in the case where two upstream segments contribute very different flow volumes at the point of confluence. In order to develop this property, a similar argument to that used to justify Equation 4.3 can be employed.

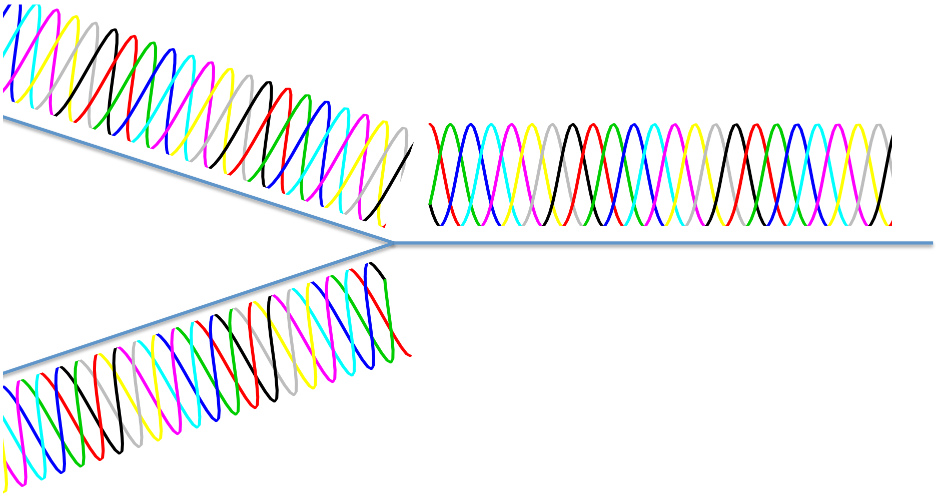


FIGURE 6.2: Illustration of B-spline basis across a confluence, where the basis is 'split' at the confluence point.

Figure 6.2 provides a visual representation for one of the basis defined in Equation 6.3, for three stream segments joined by a confluence point. The functions on the two upstream components are symmetrical, in the sense that at locations r_i and r_j, (distance r

upstream on segments i and j), the basis expansions are equal as a result of Equation 6.3 that $(B_1(r_i), B_2(r_i), \ldots, B_p(r_i)) = (B_1(r_j), B_2(r_j), \ldots, B_p(r_j))$. However, Figure 6.2 differs from Equation 6.3 because the basis defined over the downstream segment (on the right hand side of Figure 6.2) have been *split* from the parts of the basis functions that are defined on the two upstream segments. In other words, any basis function that is non-zero over both the downstream and upstream segments, is separated into 3 separate functions. Splitting the basis in this manner allows some notion of basis functions 'belonging' to each segment, which makes it easier to define separate penalties for small scale *within* segment smoothness, and for larger scale *between* segment variability that is due to mixing of water at confluences.

Following this adjusted basis specification, each stream segment s is then associated with a set of parameters $\beta(s,1), \ldots, \beta(s,p_s)$ and each of these in turn corresponds to a set of p_s basis functions, where p_s is the number of basis functions that occur on segment s as defined by the 'splitting' procedure just described. The expected value of a response variable on the network $y(s,r)$, is then given by $E(y(s,r)) = \sum_{i=1}^{p_s} \beta(s,i)B_i^*(r,s)$ where the $B_i^*(r,s)$ denotes the basis functions defined on segment s after the splitting operation. This specification automatically induces smoothness within each individual stream segment s, and with the addition of a roughness penalty for each segment smoothness could also be controlled. However, this description does not allow for borrowing of strength *between* stream segments, a mechanism which is now described.

For the idealised confluence shown in Figure 6.2, let the upstream segments be labelled a and b with the downstream segment called c. Associate with each of these a flow level f_a, f_b and f_c and a set of basis functions $(B_1(r,a), \ldots, B_{p_a}(r,a))$, $(B_1(r,b), \ldots, B_{p_b}(r,b))$ and $(B_1(r,c), \ldots, B_{p_c}(r,c))$, respectively. Then define the first functions in each set as $B_1(r,a)$, $B_1(r,b)$ and $B_1(r,c)$, those that lie farthest upstream on their respective stream segments. In order to enforce smoothness on the fitted function that occurs *within* each stream segment, define the roughness measure $R_1(\beta)$, by

$$\begin{aligned} R_1(\boldsymbol{\beta}) &= \sum_{i=2}^{p_a} \left(\beta(a,i) - \beta(a,i-1)\right)^2 \\ &\quad + \sum_{j=2}^{p_b} \left(\beta(b,j) - \beta(b,j-1)\right)^2 \\ &\quad + \sum_{k=2}^{p_c} \left(\beta(c,k) - \beta(c,k-1)\right)^2 . \end{aligned} \tag{6.4}$$

In addition to the above an additional penalty that represents roughness *across* adjacent stream segments that are connected by a confluence, and in proportion to the flow contribution of the upstream elements can be constructed.

$$R_2(\boldsymbol{\beta}) = \frac{f_a^2}{f_c^2}\left(\beta(a,p_a) - \beta(c,1)\right)^2 + \frac{f_b^2}{f_c^2}\left(\beta(b,p_b) - \beta(c,1)\right)^2 , \tag{6.5}$$

using the same idea of mass balance that underpins the simpler penalty shown in Equation 4.2. If a vector of observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ is observed across segments a, b and c, then incorporating the roughness measures described in Equations 6.4 and 6.5 the parameters associated with all of the basis functions could be obtained by minimising the objective function

$$\min_{\beta,\lambda_1,\lambda_2} \left[(\boldsymbol{y} - \boldsymbol{X\beta})^\top(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda_1 R_1(\boldsymbol{\beta}) + \lambda_2 R_2(\boldsymbol{\beta})\right].$$

where

$$\boldsymbol{X} = \begin{bmatrix} B_1(r_{y_1},a) & \ldots & B_{p_a}(r_{y_1},a) & B_1(r_{y_1},b) & \ldots & B_{p_b}(r_{y_1},b) & B_1(r_{y_1},c) & \ldots & B_{p_c}(r_{y_1},c) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_1(r_{y_n},a) & \ldots & B_{p_a}(r_{y_n},a) & B_1(r_{y_n},b) & \ldots & B_{p_b}(r_{y_n},b) & B_1(r_{y_n},c) & \ldots & B_{p_c}(r_{y_n},c) \\ & & & & & & & & \end{bmatrix}$$

6.3.3 Issues and considerations

In the previous section, a possible method for constructing a smooth functional representation for responses on river networks based on penalised B-spline basis functions was given. The ideas are similar to those presented in Chapter 4, except that here, smoothness within stream segments is possible.

The most important problem that would need to be overcome in implementing this extended model is the result of choosing the number of B-spline basis functions required to represent the network. As an illustrative example, assuming a moderate network size of 300 segments of equal length, a conservative choice might be to allow 5 basis functions per segment. This would result in a total of 1500 parameters to be estimated and could hamper any possibility of building space-time interaction models if not handled carefully. On the other hand, the model matrix that would be constructed to fit this model is sparse by construction, as the fitted value associated with an observation only depends on the basis functions within the stream segment on which it was measured, and it is highly likely that very fast bespoke routines could avoid much of the computational cost that would otherwise be incurred.

6.4 General remarks on future research

The work of each chapter of Chapters 3, 4 and 5 admit some desirable extensions that have been discussed in detail in the preceding two sections. However, the real potential for the methodologies described here lies in the possibility of integrating the components of rainfall, flow generation and measures on the network into a single modelling framework. To see the importance of this, consider that the methods developed by Ver Hoef *et al.* (2006) and also as described in Chapter 4 and in O'Donnell *et al.* (2013), the set of weights used to represent the proportional influences of upstream segments on downstream segments across confluences are assumed fixed. This assumption simplifies the analysis, as it means that the spatial penalty matrix of O'Donnell *et al.* (2013) and the spatial covariance matrix is a function of only one or two parameters. In addition, the set of weights are often intended to correspond to the relative influence of different flow rates, but because river flow is rarely measured across the whole set of stream segments proxy variables are often used instead, such as Shreve order sub-catchment drainage area, which might be appropriate when the measured variables are of a low temporal resolution, but not on shorter time scales when flow levels can change dramatically and differentially in space.

A relatively simple improvement might be obtained by treating the set of flow weights as random variables, assigning each an appropriate distribution that is centered on the weight resulting from a proxy measure or other estimate. This adjustment would be most easily implemented in a Bayesian hierarchical framework, and allows some measure of uncertainty about the set of weights. If the data are measured at high temporal resolution, and interest lies in capturing spatio-temporal structure, then it is important that the set of flow weights is not only treated as a random variable, but one with a time varying mean component. In this regard, it would be a very interesting possibility to build a distributed lag model for the purpose of building this time varying component, that should also vary across the network allowing a direct connection to be made between rainfall, flow levels and subsequent flow dependent attributes such as temperature and dissolved pollution.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory, 2nd, Tsahkadsor, Armenian SSR*, 267–281.

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica: Journal of the Econometric Society*, 178–196.

Baggaley, N., Langan, S., Futter, M., Potts, J. & Dunn, S. (2009). Long-term trends in hydro-climatology of a major Scottish mountain river. *Science of the Total Environment*, **407**, 4633–4641.

Bates, D. & Maechler, M. (2013). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.0-12.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2013). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.0-5.

Beven, K. (1985). Distributed models. *Hydrological Forecasting*.

Beven, K.J. (2001). *Rainfall-runoff modelling: the primer*, vol. 15. Wiley Chichester.

Bowman, A. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, Oxford.

Bowman, A., Crawford, E., Alexander, G. & Bowman, R.W. (2007). rpanel: Simple interactive controls for R functions using the tcltk package. *Journal of Statistical Software*, **17**.

Bowman, A.W., Giannitrapani, M. & Marian Scott, E. (2009). Spatiotemporal smoothing and sulphur dioxide trends over Europe. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **58**, 737–752.

Brezger, A., Kneib, T. & Lang, S. (2005). BayesX: Analysing Bayesian structured additive regression models. *Journal of statistical software*, **14**, 1–22.

Chambers, J.M. (2008). *Software for data analysis: programming with R*. Springer.

Clement, L. & Thas, O. (2007). Spatio-temporal statistical models for river monitoring networks. *Journal of Agricultural, Biological, and Environmental Statistics*, **12**, 161–176.

Clement, L., Thas, O., Vanrolleghem, P., Ottoy, J.P. *et al.* (2006). Spatio-temporal statistical models for river monitoring networks. *Water Science & Technology*, **53**, 9–15.

Craven, P. & Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377–403.

Cressie, N. & Majure, J.J. (1997). Spatio-temporal statistical modeling of livestock waste in streams. *Journal of Agricultural, Biological, and Environmental Statistics*, **2**, 24–47.

Cressie, N. & O'Donnell, D. (2010). Comment: Statistical dependence in stream networks. *Journal of the American Statistical Association*, **105**.

Cressie, N. & Wikle, C.K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, New York.

Cressie, N., Frey, J., Harch, B. & Smith, M. (2006). Spatial prediction on a river network. *Journal of agricultural, biological, and environmental statistics*, **11**, 127–150.

Currie, I., Durban, M. & Eilers, P. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 259–280.

Currie, I.D. & Durban, M. (2002). Flexible smoothing with p-splines: a unified approach. *Statistical Modelling*, **2**, 333–349.

Davis, T.A. (2006). *Direct methods for sparse linear systems*, vol. 2. Society for Industrial and Applied Mathematics.

De Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag.

Diblasi, A. & Bowman, A.W. (2001). On the use of the variogram in checking for independence in spatial data. *Biometrics*, **57**, 211–218.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive theory of functions of several variables*, 85–100, Springer.

Eilers, P. & Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, **11**, 89–102.

Eilers, P. & Marx, B. (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, **11**, 758–783.

Eilers, P.H. & Marx, B.D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 637–653.

Eilers, P.H., Currie, I.D. & Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational statistics & data analysis*, **50**, 61–76.

European Parliament (1991). Council Directive 91/676/EEC of 12 December 1991 concerning the protection of waters against pollution caused by nitrates from agricultural sources. *Official Journal of the European Communities*, **375**, 12.

European Parliament (2000). Directive 2000/60/EC. of the European Parliament, establishing a framework for community action in the field of water policy. *Official Journal of the European Communities*, **327**, 1–72.

Fahrmeir, L. & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society Series C-Applied Statistics*, **50**, 201–220.

Fahrmeir, L., Kneib, T., Lang, S. & Marx, B. (2013). Regression models. In *Regression*, 21–72, Springer.

Furrer, R. & Sain, S.R. (2010). spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. *Journal of Statistical Software*, **36**, 1–25.

Gardner, B., Sullivan, P.J. & Lembo Jr, A.J. (2003). Predicting stream temperatures: geostatistical model comparison using alternative distance metrics. *Canadian Journal of Fisheries and Aquatic Sciences*, **60**, 344–351.

Gardner, K.K. & McGlynn, B.L. (2009). Seasonality in spatial variability and influence of land use/land cover and watershed characteristics on stream water nitrate concentrations in a developing watershed in the Rocky Mountain West. *Water Resources Research*, **45**, W08411.

Garreta, V., Monestiez, P. & Ver Hoef, J.M. (2010). Spatial modelling and prediction on river networks: up model, down model or hybrid? *Environmetrics*, **21**, 439–456.

Gasparrini, A., Armstrong, B. & Kenward, M. (2010). Distributed lag non-linear models. *Statistics in medicine*, **29**, 2224–2234.

Gelfand, A., Ecker, M., Knight, J. & Sirmans, C. (2004). The dynamics of location in home price. *The Journal of Real Estate Finance and Economics*, **29**, 149–166.

Giannitrapani, M., Bowman, A. & Scott, E. (2011). Additive models for correlated data with applications to air pollution monitoring. In R. Chandler & E. Scott, eds., *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences*, chap. 7, 267–282, Wiley, London.

GREEN, P.J., SILVERMAN, B.W., SILVERMAN, B.W. & SILVERMAN, B.W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman & Hall London.

HASTIE, T. & TIBSHIRANI, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

HASTIE, T. & TIBSHIRANI, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 757–796.

HEIM, S., FAHRMEIR, L., EILERS, P.H. & MARX, B.D. (2007). 3d space-varying coefficient models with application to diffusion tensor imaging. *Computational Statistics & Data Analysis*, **51**, 6212–6228.

HOEF, J.V., PETERSON, E., CLIFFORD, D. & SHAH, R. (2014). SSN: An R package for spatial statistical modeling on stream networks. *Journal of Statistical Software*, **56**, To appear.

HURVICH, C.M. & TSAI, C.L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

HURVICH, C.M., SIMONOFF, J.S. & TSAI, C.L. (2002). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **60**, 271–293.

HUTCHINSON, M. (1989). A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, **18**, 1059–1076.

JAKEMAN, A., LITTLEWOOD, I. & WHITEHEAD, P. (1990). Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. *Journal of hydrology*, **117**, 275–300.

Kammann, E. & Wand, M.P. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **52**, 1–18.

Koenker, R. & Ng, P. (2013). *SparseM: Sparse Linear Algebra*. R package version 1.03.

Koyck, L.M. (1954). *Distributed lags and investment analysis*. North-Holland Publishing Company Amsterdam.

Lambert, P. & Eilers, P.H.C. (2005). Bayesian proportional hazards model with time-varying regression coefficients: a penalized poisson regression approach. *Statistics in Medicine*, **24**, 3977–3989.

Lang, S. & Brezger, A. (2004). Bayesian P-splines. *Journal of computational and graphical statistics*, **13**, 183–212.

Lang, S., Fronk, E.M. & Fahrmeir, L. (2002). Function estimation with locally adaptive dynamic models. *Computational Statistics*, **17**, 479–500.

Langan, S.J., Wade, A., Smart, R., Edwards, A., Soulsby, C., Billett, M., Jarvie, H., Cresser, M., Owen, R. & Ferrier, R. (1997). The prediction and management of water quality in a relatively unpolluted major Scottish catchment: current issues and experimental approaches. *Science of the Total Environment*, **194**, 419–435.

Lee, D. & Mitchell, R. (2014). Controlling for localised spatio-temporal autocorrelation in long-term air pollution and health studies. *Statistical methods in medical research*, To appear.

Lee, D.J. (2010). *Smoothing mixed models for spatial and spatio-temporal data*. PhD Thesis.

Lee, D.J. & Durbán, M. (2011). P-spline anova-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, **11**, 49–69.

Loecher, M. (2014). *RgoogleMaps: Overlays on Google map tiles in R*. R package version 1.2.0.6.

MacNab, Y.C. & Dean, C.B. (2002). Spatio-temporal modelling of rates for the construction of disease maps. *Statistics in Medicine*, **21**, 347–358.

Marx, B. & Eilers, P. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, **28**, 193–209.

Militino, A.F., Ugarte, M.D. & Ibáñez, B. (2008). Longitudinal analysis of spatially correlated data. *Stochastic Environmental Research and Risk Assessment*, **22**, 49–57.

Money, E., Carter, G.P. & Serre, M.L. (2009). Using river distances in the space/-time estimation of dissolved oxygen along two impaired river networks in New Jersey. *Water Research*, **43**, 1948–1958.

Muggeo, V.M. (2008). Modeling temperature effects on mortality: multiple segmented relationships with common break points. *Biostatistics*, **9**, 613–620.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability & Its Applications*, **9**, 141–142.

Nash, J. (1957). The form of the instantaneous unit hydrograph. *IAHS Publ*, **45**, 114–121.

Ng, E.G. & Peyton, B.W. (1993). Block sparse Cholesky algorithms on advanced uniprocessor computers. *SIAM Journal on Scientific Computing*, **14**, 1034–1056.

O'Donnell, D., Rushworth, A., Bowman, A.W., Scott, E.M. & Hallard, M. (2013). Flexible regression models over river networks. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Peterson, E.E. & Ver Hoef, J.M. (2010). A mixed-model moving-average approach to geostatistical modeling in stream networks. *Ecology*, **91**, 644–651.

Peterson, E.E., Ver Hoef, J.M., Isaak, D.J., Falke, J.A., Fortin, M.J., Jordan, C.E., McNyset, K., Monestiez, P., Ruesch, A.S., Sengupta, A. *et al.* (2013). Modelling dendritic ecological networks in space: an integrated network perspective. *Ecology letters*.

PINHEIRO, J., BATES, D., DEBROY, S., SARKAR, D. & R CORE TEAM (2013). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-113.

R DEVELOPMENT CORE TEAM (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAMSAY, J. (1988). Monotone regression splines in action. *Statistical Science*, 425–441.

ROBERTS, A.M. (2008). Exploring relationships between phenological and weather data using smoothing. *International Journal of Biometeorology*, **52**, 463–470.

RUE, H. & HELD, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC Press.

RUPPERT, D., WAND, M.P. & CARROLL, R.J. (2003). *Semiparametric regression*, vol. 12. Cambridge University Press.

RUSHWORTH, A.M., BOWMAN, A.W., BREWER, M.J. & LANGAN, S.J. (2013). Distributed lag models for hydrological data. *Biometrics*, **69**, 537–544.

SANSÃŞ, B. & GUENNI, L. (1999). Venezuelan rainfall data analysed by using a bayesian spaceâĂŞtime model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **48**, 345–362.

SCHWARZ, G. (1978). Estimating the dimension of a model. *The annals of statistics*, **6**, 461–464.

SHADDICK, G. & WAKEFIELD, J. (2002). Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society Series C-Applied Statistics*, **51**, 351–372.

SHAW, E.M. (2010). *Hydrology in Practice*. CRC Press Llc.

SHREVE, R.L. (1966). Statistical law of stream numbers. *The Journal of Geology*, 17–37.

Tabrizi, M., Said, S., Badr, A., Mashor, Y. & Billings, S. (1998). Nonlinear modeling and prediction of a river flow system. *JAWRA Journal of the American Water Resources Association*, **34**, 1333–1339.

Tobler, W.R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, **46**, 234–240.

UK Meteorological Office (2003). Rain radar products (NIMROD). *NCAS British Atmospheric Data Centre*.

UK Meteorological Office (2013). National meteorological library and archive fact sheet 15 - weather RADAR.

Varadhan, R., University, J.H., Borchers, H.W. & Research., A.C. (2011). *dfoptim: Derivative-free Optimization*. R package version 2011.8-1.

Ver Hoef, J.M. & Peterson, E.E. (2010a). A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association*, **105**, 6–18.

Ver Hoef, J.M. & Peterson, E.E. (2010b). Statistical dependence in stream networks rejoinder. *Journal of the American Statistical Association*, **105**, 22–24.

Ver Hoef, J.M., Peterson, E. & Theobald, D. (2006). Spatial statistical models that use flow and stream distance. *Environmental and Ecological statistics*, **13**, 449–464.

Vörösmarty, C.J., McIntyre, P., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S.E., Sullivan, C.A., Liermann, C.R. *et al.* (2010). Global threats to human water security and river biodiversity. *Nature*, **467**, 555–561.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association*, **93**, pp. 341–348.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.

Welty, L., Peng, R., Zeger, S. & Dominici, F. (2009). Bayesian distributed lag models: estimating effects of particulate matter air pollution on daily mortality. *Biometrics*, **65**, 282–291.

Wong, H., Ip, W.c., Zhang, R. & Xia, J. (2007). Non-parametric time series models for hydrological forecasting. *Journal of Hydrology*, **332**, 337–347.

Wood, S. (2006). *Generalized Additive Models: an introduction with R*. Chapman and Hall/CRC, London.

Wood, S.N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**, 413–428.

Zanobetti, A., Wand, M., Schwartz, J. & Ryan, L. (2000). Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*, **1**, 279–292.