Gaussian Markov Random Fields and Structural Additive Regression

Applications in Freshwater Fisheries Management

Colin Pearson Millar

University of
St Andrews

This dissertation is submitted in partial fulfilment for the degree of PhD

at the

University of St Andrews

March 2017

For Oscar, Niall and Lyra -

I owe you all some time! This work has taken your whole life to finish!

And,

For Sarah -

The reason I could keep going.

Declaration

1. Candidate's declarations:

I, Colin Pearson Millar, hereby certify that this thesis, which is approximately 60,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in September, 2004 and as a candidate for the degree of PhD in September, 2007; the higher study for which this is a record was carried out in the University of St Andrews between 2007 and 2016.

Date: March 2017,
Signature of candidate:

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date: March 2017,
Signature of supervisor:

3. Permission for publication:

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible

for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

PRINTED COPY

Embargo on all or part of print copy for a period of 2 years (maximum five) on the following ground(s): Publication would preclude future publication

Supporting statement for printed embargo request:

The research presented in this thesis is in the process of being written up for submission as peer reviewed articles. The publication of the thesis would preclude publication.

ELECTRONIC COPY

Embargo on all or part of electronic copy for a period of 2 years (maximum five) on the following ground(s): Publication would preclude future publication

Supporting statement for electronic embargo request:

The research presented in this thesis is in the process of being written up for submission as peer reviewed articles. The publication of the thesis would preclude publication.

Date: March 2017,
signature of candidate:

signature of supervisor:

Acknowledgements

I would like to express my appreciation and thanks to all the people who have supported me over the years. I greatly thank and acklowledge Rob Fryer, my supervisor and at one time line manager - he has been unfailingly supportive of all my ventures and has always been there to talk things through, be it statistics, career or just about cricket (though I can't really contribute on the cricket front). I would also like to thank David Borchers who has been a long standing motivation and support for me throughout my academic carreer. I have always greatly appreciated and admired the clarity that both Rob and David can bring to a problem. I would also like to thank Ruth King for her role as a supervisor in the early part of my PhD, and for the energy and enthusuasm she has instilled in me for statistics.

A special thanks to Sarah who kept me and the kids alive and healthy while I was of no help whatsoever. I owe you so much! Its hard to think that when I started this PhD I was a single loner, and now I am married with 3 children! That we have moved house six times going from Aberdeen, to Italy, to Pitlochry and now to Copenhagen! What next?

Abstract

In this thesis structural additive regression (STAR) models are constructed for two applications in freshwater fisheries management 1) large scale modelling of fish abundance using electrofishing removal data and 2) assessing the effect on stream temperature of tree felling. Both approaches take advantage of the central role Gaussian Markov random fields (GMRFs) play in the construction of structured additive regression components.

The R package mgcv can fit, in principle, any STAR model. In practice, however, several extensions are required to allow a non-specialised user to access this functionality, and a large part of this thesis is the developement of software to allow a general user ready access to a wide range of GMRF models within the familiar mgcv framework. All models are fitted making use of this extension where possible (and practical).

The thesis is divided into three main chapters. Chapter 2 serves to provide background and insight into penalised regression and STAR models and the role that GMRFs play in smoothing. Also presented are the extensions required to fit GMRF models in mgcv.

Chapter 3 presents a two stage model for fish density using electrofishing removal data. The first stage of this model estimates fish capture probability and is not a STAR model, but can utilise aspects of GMRFs through low rank approximations; software to make this available is developed and presented. The second stage is a Poisson STAR model and can therefore be fitted in the extended mgcv framework.

Finally, Chapter 4 presents a model for the impact of a clear felling event on stream temperature. This model utilises cyclic smoothers applied to the functional principal components of daily temperature curves. This allows for a detailed assessment of the effects of felling on stream temperature that is not possible when modelling daily summaries alone.

Table of Contents

List of Figures

List of Tables

1

Introduction

> *Despite [recent] advances ..., there is still a place for computationally simpler approaches which can be used routinely, especially when a range of candidate models are under consideration.*
>
> – Peter Diggle and Paolo Ribeiro, *Model Based Geostatistics*

1.1 Overview

The aim of this thesis is to promote the use of structured additive regression (STAR) modelling in freshwater fisheries management. STAR models (Fahrmeir et al., 2004, 2013) are simply an extension of generalised additive models (Wood, 2006) in which the options for modelling covariates extends beyond smoothers to encompass a more general class of *structural* forms, for example spatial models for rivers (Cressie et al., 2006). What is meant by structure in this context will be made explicit in the following chapters. The attractiveness of STAR models is their simple form and their

potential for wide practical use through additions to already existing and widely used software for fitting GAMs (Wood, 2006, 2011).

The complex dynamical nature of the natural systems which are of relevance to freshwater fisheries management require complex statistical models. Several authors (Cressie et al., 2006; Hilborn and Mangel, 1997; Thorson and Minto, 2014; Wikle, 2003) argue that Bayesian hierarchical models or mixed effect models are suitable, even necessary tools for the analysis of such ecological systems. The arguments are based on two key features of ecological data: the data is often sampled in a complex way, and the underlying systems generating the data are dynamic in nature. Cressie et al. (2009) argue that hierarchical statistical modelling is a powerful way to deal with *inevitable but quantifiable uncertainties*, including model uncertainty. However, in my experience, the implementation of such approaches tends to require the development of tailor-made computer code, using tools such as WinBugs (Lunn et al., 2000), ADMB (Fournier et al., 2012) and STAN (Stan Development Team, 2014b), and as such are not available to the non-technical ecological modeller. Diggle and Ribeiro (2007) acknowledge the need for computationally simpler approaches for routine use and when a range of models are being considered. They give two alternatives to Monte Carlo techniques: hierarchical generalised linear models (Lee and Nelder, 1996, 2001) and generalised estimating equations (Liang and Zeger, 1986). Recently, there has been a surge in the application of non-Markov chain Monte Carlo (MCMC) Bayesian modelling tools for hierarchical and spatio-temporal models (Blangiardo et al., 2013; Rue et al., 2009) using a technique called integrated nested Laplace approximation (INLA).

The common theme running through the above modelling approaches is that the form of the models are broadly defined by a linear model in which the model parameters are considered to be constrianed in some way, ether by following a particular distribution, or by being penalised for deviating from a mathematical form. This can impose

a wide variety of structure on the model and there are a variety of ways of doing this. This particular imposition of structure is often referred to as a penalty. There is a duality between penalties from a GAM perspective and prior structure through multivariate priors, from a Bayesian heirarchical sense, and this is a core element of this thesis. I will follow Rue and Held (2005) and only consider prior distributions that are multivariate normal. I will also restrict my attention to data distributions that belong to the exponential family. The INLA package is designed to fit this type of model, but I will use the frequentist alternative, mgcv (Wood, 2006) which I consider to have a simpler interface and more likely to be used by non-technical fisheries biologists.

Penalised likelihood (Green, 1987) and the related penalised least squares is at the core of the R package mgcv (Wood, 2006), and is a very popular statistical tool, used widely by fisheries biologists for fitting generalised additive models (GAMs). Fahrmeir et al. (2013) cover estimation of the wider class of STAR models pointing out that mgcv is capable in theory of fitting anything of the form

$$E(\boldsymbol{y}) = \eta\Big(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}\Big) \tag{1.1}$$

where $\boldsymbol{X}$ is an $n \times p$ matrix of covariates and $\boldsymbol{\beta}$ is a column vector of p parameters to be estimated; similarly $\boldsymbol{Z}$ is an $n \times q$ matrix of covariates and $\boldsymbol{\gamma}$ is a column vector of q parameters to be estimated. The function η is the so-called link function, and links the data $\boldsymbol{y}$, a column vector of n observations from an exponential family distribution, such that η maps $\mathbb{R}^n$ to the support of the chosen exponenial family distribution. The parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ differ in that $\boldsymbol{\gamma}$ is to be constrained, where as $\boldsymbol{\beta}$ are to be unconstrained. This gives the design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ different requirements: $\boldsymbol{X}$ must always have less than the number of observations in order for $\boldsymbol{\beta}$ to be estimable, where as $\boldsymbol{Z}$ can be over paramerised due to the fact that $\boldsymbol{\gamma}$

are penalised. This description leads to both linear mixed models and to penalised regression. In linear mixed models the parameter vector $\boldsymbol{\gamma}$ can be thought of as having a multivariate normal distribution, but as GAMs are usually described as penalised spline models, this is not the typical way of describing a GAM. However, as many have previously stated (e.g. Fahrmeir et al., 2013; Verbyla et al., 1999; Wood, 2006) there is symmetry between penalised splines and linear mixed effects models, which will be discussed in the following chapter. The main point here is that this simple formulation is very flexible indeed and the connection between mixed effects and penalised spline models provides the basis for a rich set of model options in one simple framework. With this simplicity there comes the restriction that STAR models are less general than fully hierarchical models, and data are restricted to following distributions in the exponential family. However, the simplicity of the form of STAR models means that they fit easily into standard software for penalised and liner mixed effects modelling.

In this thesis I focus on developing models for two specific case studies: modelling juvenile salmon abundance from removal electrofishing sampling, and modelling the effect of clear felling on stream temperature. Both are of importance to freshwater fishery management, and both present modelling problems that can be addressed by the use of STAR models.

1.2 STARs for ecological modelling

Structural additive regression (STAR) is a unifying class of models that extends additive models to incorporate any quadratically penalised linear model. This allows the inclusion of splines, discrete spatial models, varying coefficient terms and random

effects. The most general form of a structured additive predictor is

$$E(y_i) = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \boldsymbol{z}_{i1}\boldsymbol{\gamma}_1 + \cdots + \boldsymbol{z}_{iq}\boldsymbol{\gamma}_q. \tag{1.2}$$

where y_i is the ith observation, $x_{i1}, \ldots, x_{ip}$ are p covariates assocated with the ith observation and $\beta_1, \ldots, \beta_p$ are parameters to be estimated; $\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iq}$ are q row vectors of covariates assocated with the ith observation and $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q$ are q column vectors of parameters to be estimated. Typically $\boldsymbol{z}_{ij}, j = 1, \ldots, q$ are specific to a particular structural form (i.e. the values of a spline at an appropriate point) and the parameter vectors $\boldsymbol{\gamma}_j,\ j = 1, \ldots, q$ are each subject to a quadratic penalty, i.e.

$$\boldsymbol{\gamma}_j'\boldsymbol{Q}_j\boldsymbol{\gamma}_j \tag{1.3}$$

For example: a) x_{i1} could be an indicator function for whether the ith observation was taken in 2005 and β_1 the 2005 year effect. b) x_{i2} could be the altitude of the location where the observation was made and β_2 the altitude effect. c) $\boldsymbol{z}_{il}$ could be a set of indicator functions for whether the ith observation was made in one of 10 mutually exclusive spatial regions and $\boldsymbol{\gamma}_1$ are the 10 region effects; these region effects could be penalised as a random effect, and hence be unstructured, or as a spatial effect in which neighbouring regions are constrained to be similar. These would be encoded through the matrix $\boldsymbol{Q}_1$. d) $\boldsymbol{z}_{i2}$ could be the values of seven increasingly complex functions defined over the range 0 to 24, evaluated at the time of day that the ith observation was made. In this case, $\boldsymbol{\gamma}_2$ are uninterpretable coefficients, but the combination $\boldsymbol{z}_{i2}\boldsymbol{\gamma}_2$ results in a smooth function over time evaluated at the time of day that the ith observation was made. This function could have smooth first, second or third derivatives, and in addition be forced to be cyclical - all of these features would be encoded in the penalty matrix $\boldsymbol{Q}_2$.

In general all the additive elements of a STAR model can be defined by a single design matrix $\boldsymbol{Z}$ and a single quadratic penalty $\boldsymbol{Q}$, so that, in matrix form, the general definition of the structured additive predictor is

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} \qquad \text{with penalty} \qquad \boldsymbol{\gamma' Q \gamma} \tag{1.4}$$

The purpose of the penalty is to restrict the parameter vector $\boldsymbol{\gamma}$ in some way. An example of this is a smoother for a time trend in which estimates are restricted to vary gradually over time. In effect we are borrowing information from neighbouring time points in order to produce a smooth curve, hence a STAR model can be considered as a linear mixed effects model. What distingiuishes a STAR model is really the inner structure defined by (1.2) and (1.3).

To make the connection in the opposite direction, that is to look at GAMs as linear mixed effects models, consider the random effects model in which $\boldsymbol{\gamma}$ denotes group means which are to be normally distributed with zero mean and variance σ^2. The log probability density of the group means is proportional to

$$\frac{1}{2\sigma^2}\boldsymbol{\gamma' I \gamma} \tag{1.5}$$

where, $\boldsymbol{I}$ is the identity matrix. So the equivalent STAR model is one where the quadratic penalty is proportional to $\boldsymbol{Q} = \frac{1}{\sigma^2}\boldsymbol{I}$. In the language of GAMS and penalised regression, $\frac{1}{\sigma^2}$ is related to a parameter controlling the degree of smoothing; the smaller the variance, the less variable the random effect is and hence the more *smooth* it is. This highlights another subtlety in the move from GAMS to the more general STAR model. The role of the smoothing parameter in a GAM is directly interpretable as altering the smoothness of a curve or surface. But in a STAR model, the same parameter can influence a wider range of properties, such as shrinkage in

a random effect context; the specific influence that a smoothing parameter has is governed by the structure of the penalty itself.

Additionally, smoothing parameters allow for a kind of automatic model selection (Scheipl et al., 2013a) and can be thought of as incorporating model uncertainty (Cadigan, 2013). The smoothing parameter can be thought of as providing a continuous link across models of varying complexity. The process of finding appropriate smoothing parameters is thus a process very much like model selection, where rather than using a step-wise procedure to find an optimum model configuration, the continuous smoothing parameter allows something akin to a weighted average of models.

So why not just fit these models as random effects? STAR models can be fitted in a variety of ways, including as linear mixed effects and Bayesian hierarchical models (Fahrmeir et al., 2004). In this thesis I use generalised cross validation (Wahba, 1990) because it provides a way of estimating the parameters by relatively simple means. Generalised cross validation (GCV) is a predictive measure of model fit which is countered by the size of the model. This has much in common with other predictive measures of fit, such as AIC (Akaike, 1998) and BIC (Schwarz, 1978); in fact AIC is used to estimate smoothing parameters in Chapter 2, for demonstratation purposes. GCV is computed conditional on a set of smoothing parameters and provides a means to select an appropriate degree of smoothing. Wood (2006) develops computationally efficient ways to compute GCV which is implemented in the mgcv package in R. The developments of Wahba (1990) and Wood (2006) mean that this sophisticated modelling potential in STARs can be considered as an extension to simple linear and generalised linear models.

1.3 Linear algebra

1.3.1 Functional data and splines

As demonstrated by the use of matrices in (1.1) and (1.4) linear algebra plays an important role in this thesis. An important notion in linear algebra is the column space of a matrix. The column space can be thought of as a new coordinate system, so that a design matrix $\boldsymbol{Z}$ based on splines defines a column space often called a function space (function because the axes of the coordinate system are functions). Another way of thinking about it is that all linear combinations of the columns - that is, the result of $\boldsymbol{Z}\boldsymbol{\gamma}$ for any possible parameter vector $\boldsymbol{\gamma}$ - is a smooth function. If $\boldsymbol{Z}$ is an $n \times n$ matrix the representation of an n-dimensional data point $\boldsymbol{y}$ as a function in the column space of $\boldsymbol{Z}$ is one to one, that is, there is no information lost about $\boldsymbol{y}$ when it is represented in the function space as $\boldsymbol{\gamma}$. This is the notion of functional data (Ramsay and Silverman, 2005): data that has been transformed into a function space and is now represented by the vector $\boldsymbol{\gamma}$.

The functional data representation works very well for data that can be naturally thought of as curves, and often it is possible, if *good* basis functions are used to define the function space, that the dimension of the function space can be much smaller than the dimension of the data space. In this case the data is not transformed but *projected onto* the function space. Mathematically this is equivalent to fitting a regression model, but conceptually the spline coefficients represent the data as curves. This will inevitably result in some loss of information, but this can be minimised by a clever choice of basis functions. A pertinent example is daily variations in temperature (Chapter 4) where high frequency observations of temperature produce a very smooth curve. This curve may consist of a large number of data points, but could potentially be described very well by a small number of basis functions. One

way to find a good set of basis functions is a technique called functional principal component analysis (Silverman, 1996) which will be elaborated on in Chapter 2.

1.3.2 GMRFs and splines

The connection between GMRF penalties and splines is not obvious, but becomes clearer when reduced rank approximations of full rank GMRFs are constructed in Chapter 2. GMRFs are what is known as *full rank*, which essentially means, as a model, they have the maximum number of parameters. A reduced rank approximation is an alternative model defined by fewer parameters than the full rank case; clearly it is desirable to retain as much of the properties of the full model in the reduced one. Full rank GMRFs essentially share the same basis functions as the data, while a good reduced rank approximation requires a transformation into a more efficient functional space. This is similar to the motivation behind functional representations of data, and it turns out the the specific form of a GMRF penalty is associated with a particular function space defined by a particular set of spline bases. The most common way to find a good set of new basis functions is closely related to a principal component analysis of the GMRF penalty (see Chapter 2). These low rank approximations are particularly useful for problems where the full rank GMRF becomes inhibitively large, or where it is not practical to use penalised regression.

1.4 Personal motivation

My motivation is from the point of view of an applied statistician. I am interested in statistical theory and new developments, but there is often a need for pragmatic compromises when working with applied problems. This is particularly true when

communicating results and working in a tutorial capacity, helping researchers to fit models themselves to their data.

The statistical computing environment R (R Core Team, 2015), and the core packages for fitting linear models, generalised linear models, and generalised additive models (mgcv) provide a very flexible tool set for the practising researcher. In addition, many non-statisticians find mgcv an easy-to-use package and it has become part of the ecologist's basic analytical toolkit (see for example Zuur et al., 2007).

As argued by many (Cressie et al., 2009; Rivot et al., 2008; Wikle et al., 2013) hierarchical models such as STAR models are invaluable for ecological analysis and while there are R packages for STAR models such as BayesX (Belitz et al., 2015) and INLA (Martino and Rue, 2009), they require some specialist statistical knowledge and investment in time. Courses run at universities are increasing the user base of these packages, but in terms of non-statistical users mgcv still has a distinct advantage.

Since STAR models can be incorporated into the mgcv framework, part of my motivation in this thesis is to develop an extension to the mgcv package which will allow non-statisticians familiar with mgcv to use the wide array of STAR models out there.

1.5 Structure of the thesis

The statistical content of this thesis is demonstrated by two applications of STAR models in freshwater fisheries management. Chapter 2 will present and discuss relevant theory of smoothing splines and penalised regression and the development of software to fit STAR models in mgcv. These ideas will then be applied to two

non-standard problems. Although they require some non-standard tools, they are essentially extended STAR models, achieving their flexibility through the use of structural model components.

Chapter 2 introduces STAR models, beginning with the simple linear model, and the notion of a structured additive model term through the use of polynomials and splines. I give brief examples of the wide variety of forms that structural components can take, using examples from the applications to come. In particular I focus on the use of GMRF models (Rue and Held, 2005) which have widespread use in Bayesian modelling (for example in Bayesian disease mapping) but are much less prevalent in applied fisheries management literature. I also introduce some software developed to ease the fitting of STAR models (particularly GMRF models) using the well known mgcv package in R (Wood, 2006).

The remainder of the thesis is dedicated to two applications, each in the field of freshwater fisheries management, and each posing different problems with respect to modelling.

Chapter 3 presents the problem of estimating the spatio-temporal density of salmon fry in Scottish rivers when the data are observations from electrofishing removal samples. Current methods for such a problem are Bayesian hierarchical models (HBMs) commonly implemented in WinBugs (Rivot et al., 2008; Wyatt, 2002). However, these methods typically require the help of a highly technical scientist and HBMs coded in WinBugs are often not well suited to model comparison, or making conditional predictions. In this chapter I develop a two stage approach, which, by splitting the estimation of capture probability from density estimation, simplifies the modelling of both components. The application that motivated this development is the estimation of salmonid densities from a large number of electrofishing samples taken across Scotland. Software is developed to give non-technical users the ability

to easily explore the relationships between both capture probability and density and explanatory covariates, while also giving the user access to a wide array of structural forms of relationship.

Chapter 4 presents the problem of estimating the effect of tree removal on stream temperature. The core data are stream temperature readings taken at least every hour and up to every 15 minutes. A single felling event is assessed using a so-called before after control impact (BACI) experimental design, in which temperature loggers have continuously recorded stream temperature at the impacted stream and a control stream for a number of years before and after the felling event. Due to the number of data points it is not feasible to model the raw data directly. Two approaches are demonstrated. The first uses daily summaries such as the daily mean, where the models allow for seasonal variation and a felling impact that could decay with time. Felling effects were estimated by the difference between pre-felling and post-felling time periods. The second approach uses a technique from functional data analysis (Ramsay and Silverman, 2005) to compute the main *functional* components of variation (Silverman, 1996). The (daily) principal component scores, which are a linear transformation of the raw data, can then be modelled as the daily summaries. The benefit of the functional approach is that model predictions for each principal component can be combined to allow inference about the effect of felling on the full thermal regime, as opposed to daily summaries.

Chapter 5 synthesises the key ideas of the thesis and reflects on the successes and difficulties faced in the applications. I also take the opportunity to discuss some ideas for furthering the models developed in this thesis.

2

STARs and GMRFs

A GMRF is a really simple construct: It is just a random vector following a multivariate normal distribution.

– Hårvard Rue and Leonard Held, *GMRFs: Theory and Applications*

2.1 Introduction

2.1.1 Overview

Structural additive regression (STAR) models (Fahrmeir et al., 2004; Kneib and Fahrmeir, 2006) were originally proposed as Bayesian hierarchical models which extend penalised generalised additive models. Considered from a penalised likelihood perspective these models are a computationally simpler alternative to Bayesian hierarchical models due to the availability of robust estimation techniques for penalised likelihood (Wood, 2006). Hierarchical Bayesian models (and their penalised likelihood counterparts) are a valuable tool in freshwater fisheries modelling because

they can be designed to deal with a wide range of complex and time varying relationships (Cressie et al., 2009).

STARs are a rich class of models which extend additive models (penalised spline models) to more general forms of smoothness, such as discrete spatial models, varying coefficient terms and random effects, all based on quadratic penalties (Fahrmeir et al., 2004). The normal (or Gaussian) distribution is one of the most iconic distributions in statistics, and it is also one which has a lot of nice properties. One such property is that the log of the normal density function is a quadratic, so there is a strong link between quadratic penalties and Gaussian distributions (Rue and Held, 2005). It is for this reason that that Gaussian Markov random fields (GMRFs) provide much insight into structural additive regression.

The purpose of this chapter is to explore the connection between GMRFs (Rue and Held, 2005) and penalised regression splines (Green, 1987), and to show how GMRF models and their low rank approximations easily fit into existing statistical modelling tools in R (R Core Team, 2015). Due to its popularity and ease of use I show how the mgcv package (Wood, 2006) can be extended to fit a large number of GMRF type models, and through the use of low rank approximations to GMRF models (Strang, 2009), the time taken to fit these models can be minimal. The chapter begins with a simple example of a GMRF model. Next some theory for linear regression and spline models will be described. At this point the comparison between penalised regression and GMRFs will be made and a range of GMRF models will be presented. Some simple GMRFs will be fitted to simulated data with some more complex models fitted to demonstrate the wide scope of GMRF models in penalised regression.

The appeal of GMRFs is their outward simplicity, flexibility and practicality. A GMRF is simply a random vector that follows a multivariate normal (Gaussian)

distribution that has conditional independence (Markovian) assumptions. More insightfully, a GMRF is a random vector with a certain covariance structure[1].

2.1.2 Scope

The main scope of this chapter is as follows:

- To present the practical theory and applications of STAR and demonstrate the great breadth of potential use in everyday statistical modelling.
- To compare and discuss the similarity between regression splines and GMRFs, and show that GMRFs can be considered as central to the construction of STAR models.
- To demonstrate, through the use of examples, the incorporation of complex GMRF models into standard modelling software, through minor alterations to the mgcv package.

The innovative contribution in this chapter is in emphasising the link between GMRFs and smoothing splines for use in freshwater fisheries and providing a tool for the practically feasible incorporation of GMRF models into everyday model fitting.

2.1.3 An introductory example

An example of a frequently used GMRF was first discovered by T. N. Thiele, a Danish astronomer and statistician, in 1880 (Thiele, 1880). Lauritzen (1981) provides a sketch of how Thiele arrived at his model: Thiele considered the measurements made

[1]throughout this chapter the word *structure* refers to a correlation or penalty matrix with a particular pattern

by an instrument where part of the error is due to fluctuations of the position of the instrument itself. If $x(t)$ is the position of the instrument at time t, the most likely position of the instrument at time $t+\Delta t$ should be the position immediately before, that is $x(t)$, and deviations from this should be governed by the normal distribution law. He then considered that any sequence of instrument positions $x(t_1), \ldots, x(t_n)$ where $t_1, \ldots, t_n$ are consecutive time points, should have the property that the *increments* $x(t_{i+1}) - x(t_i)$ are independent and normally distributed with zero mean and variance depending on the time difference between observations. In other words, Theile described what is now called a *first order random walk*. If we take the time steps to be equidistant $t_i = i$ and assume constant variance for the increments τ, say, then the model for the position of the instrument is for each $i = 1, \ldots, n-1$

$$x_{i+1} - x_i \sim \mathrm{N}(0, \tau^{-1}) \tag{2.1}$$

The *process* that is being described here is a simple case of a GMRF. The distribution of the vector $\boldsymbol{x}$ is multivariate normal (this will become clear later) and the process is Markovian (in fact it is first order Markovian). Associated with this multivariate distribution is a covariance matrix which defines the random walk process. Later we will see that it is more practical to think of GMRFs in terms of the inverse covariance matrix, commonly denoted $\boldsymbol{Q}$ because it defines a quadratic equation.

GMRFs are by definition random variables and so fit easily into the Bayesian hierarchical model setting, but they can also be used in penalised regression. Notionally the reason for this is because random effect distributions from a frequentist and Bayesian perspective provide some *prior* assumptions of structure for the model parameters. A simple way to see this is to consider a simple random effects model in which group means vary about a common mean. This assumes that the group means are symmetrically distributed about some value, and any unusually large deviations will

be pulled or *shrunk* towards the mean. Enforcing structure on parameters or on the other hand penalising behaviour that doesn't fit with some structural assumptions are two sides of the same coin. It is this symmetry between structural prior distributions and regression parameter penalties that can be used to transfer ideas between the modelling frameworks.

2.1.4 GMRFs in practice

GMRFs have been used in a wide range of applications. Another early reference for both GMRFs and smoothing is Whittaker (1922), whose model is still used in actuarial modelling today. Both Thiele (1880) and Whittaker (1922) did not consider the breadth of applicability of their ideas and instead focused on a single model. A recent synthesis of GMRF models and their application can be found in Rue and Held (2005). There are four main areas of application:

Time-series analysis Time series models such as random walks, autocorrelated time series and models for seasonal variation are all examples of GMRFs and are extensively used in time series models. Any discrete time model, with normally distributed random effects, and with structure placed on the dependencies, are potential GMRF models.

Semiparametric regression and splines The most common example of GMRFs in univariate smoothing are first and second order random walks. These models can be generalised to the case where random walk models are used as the coefficients of B-splines. Other uses are where spline coefficients are modelled through time as independent AR1 processes.

Image analysis GMRFs are widely used in image analysis. There is a very large literature with many applications in image restoration from space telescopes, MRI scans, ultrasound, texture modelling; other applications include image discrimination to identify changes and differences in aspects images.

Spatial statistics This is another very large field with many applications of GMRFs in epidemiology, ecology, econometrics, oceanography, climatology. Wherever spatial variation is important, applications using GMRFs are common. Recent developments include spatial modelling of binary data and approximating continuous covariance functions using GMRFs.

2.2 GMRFs and regression splines

2.2.1 Basic theory of nonparametric regression

2.2.1.1 Linear models

In ecology, a linear model is often an economical approximation of a natural process using a sum of explanatory variables

$$y_i = \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \epsilon_i \tag{2.2}$$

where $x_{ij};\ j = 1, \ldots, q;\ i = 1, \ldots, n$ are explanatory variables for each observation $y_i;\ i = 1, \ldots, n,\ \beta_j;\ j = 1, \ldots, q$ are parameters to be estimated, and the errors are assumed to be identically and independently distributed (iid) normal variables with zero mean and variance σ^2. Importantly $q << n$ so that the model is a simplification of the data; it is a description of n observations using q parameters. Linear models

have had great success in statistical modelling due to the great flexibility that can be achieved with such a seemingly simple form and because parameter estimation in linear models is computationally efficient. Both the flexibility and efficiency are due to the use of linear algebra (Strang, 2009). I will first deal briefly with computational aspects then turn to flexibility.

Writing (2.2) in a compact form using matrices and vectors, i.e.,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.3}$$

where

$$\boldsymbol{\epsilon} \sim \mathrm{N}\left(\boldsymbol{0},\, \sigma^2 \boldsymbol{I}\right) \tag{2.4}$$

allows a range of techniques from linear algebra to be used to solve the linear regression problem.

A very common tool, and one that will be used later, is the QR decomposition (Strang, 2009). In a linear model, the use of the QR decomposition allows the regression to be restated in a computationally simpler form by finding a new set of explanatory variables which are orthogonal (or independent) to each other. This in turn means that the model parameters influence the solution in an independent way, simplifying parameter estimation.

To gain a deeper understanding of the QR decomposition as applied to a linear model it is useful to look at the linear model from a geometrical perspective. First think of the data vector $\boldsymbol{y}$ as a single point in n-dimensional space. As with all imaginings of multidimensional space, it is a good first step to start with familiar three dimensional space (i.e. $n = 3$). The model ($\boldsymbol{X}\boldsymbol{\beta}$, where $\boldsymbol{X}$ is a $n \times q$ matrix of explanatory variables, and $\boldsymbol{\beta}$ is a q-vector of coefficients) can be thought of as a

q-dimensional solution plane (again, imagine a $q = 2$-dimensional surface) that cuts through the n-dimensional data space. The parameters that result in a model that best fits the data is the point on the solution plane that is closest to the data point. From simple geometry this is known to be the point on the solution plane that defines a perpendicular line between the solution plane and the data point. To find this point on a general plane is quite possible using a range of techniques, however the QR decomposition provides a shortcut. The QR decomposition finds a rotation so that the solution plane becomes horizontal, like a table top. In this new and equivalent problem, the closest point on the table to the data point is found by simply dropping a line straight down, like a plumb line, or equivalently, looking down at the table from above and projecting the 3 dimensional data point onto the 2-dimensional table. Furthermore, the QR decomposition does not just rotate the problem. Because the original regression problem was posed in terms of potentially correlated explanatory variables, the coordinate system of the solution plane could be like a diamond shaped garden trellis, with the coordinate systems, or axes, not at right angles to one another. The QR decomposition also stretches the solution plane so that the coordinate system is right angled.

Mathematically this plays out as follows

$$\boldsymbol{X\beta} = \boldsymbol{QR\beta} = \boldsymbol{Q\beta'} \tag{2.5}$$

so that

$$\boldsymbol{Q}^T\boldsymbol{y} = \hat{\boldsymbol{\beta}}' \tag{2.6}$$

where, $\boldsymbol{Q}$ is an $n \times n$ orthogonal matrix and $\boldsymbol{R}$ is a $n \times p$ upper triangular matrix (the upper triangle corresponds to the rotation). The columns of $\boldsymbol{Q}$ form an orthonormal

basis and $\boldsymbol{\beta}'$ is the rotated version of $\boldsymbol{\beta}$. The factorisation is unique if the diagonal elements of $\boldsymbol{R}$ are restricted to be positive. Because $\boldsymbol{Q}$ is orthogonal $\boldsymbol{Q}^T = \boldsymbol{Q}^{-1}$ so the solution to $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{Q}\boldsymbol{\beta}'$ is easily computed (2.6). Note that $\boldsymbol{Q}$ is used here in the context of QR decomposition. Elsewhere in the thesis $\boldsymbol{Q}$ is used to refer to a quadratic penalty matrix.

2.2.1.2 Non-linear responses

A simple way to extend a linear model to include a nonlinear response is to transform the covariate. For example, modelling log temperature instead of temperature. Single transformations can only get you so far and an alternative is to include several transformations of the same variable. This has most commonly been done by including linear, quadratic and higher polynomial terms of a variable. For example, a polynomial of degree 3 is the model

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \beta_2 \boldsymbol{x}^2 + \beta_3 \boldsymbol{x}^3 + \boldsymbol{\epsilon} \tag{2.7}$$

where

$$\boldsymbol{\epsilon} \sim \mathrm{N}\left(\boldsymbol{0},\, \sigma^2 \boldsymbol{I}\right) \tag{2.8}$$

These simple polynomials form what are (in essence) basis functions. In practice it is better to either centre the covariate x beforehand or to use orthogonal polynomials (Kennedy and Gentle, 1980). Orthogonal polynomials, like the columns of $\boldsymbol{Q}$ earlier, are an example of a family of orthogonal basis functions. Furthermore, because the design matrix is now orthogonal by design, the solution to (2.7) can be computed using (2.6) with $\boldsymbol{X} = \boldsymbol{Q}$. To illustrate, the model (2.7) was fitted to some simulated data and is presented in Figure 2.1, along with each polynomial component.

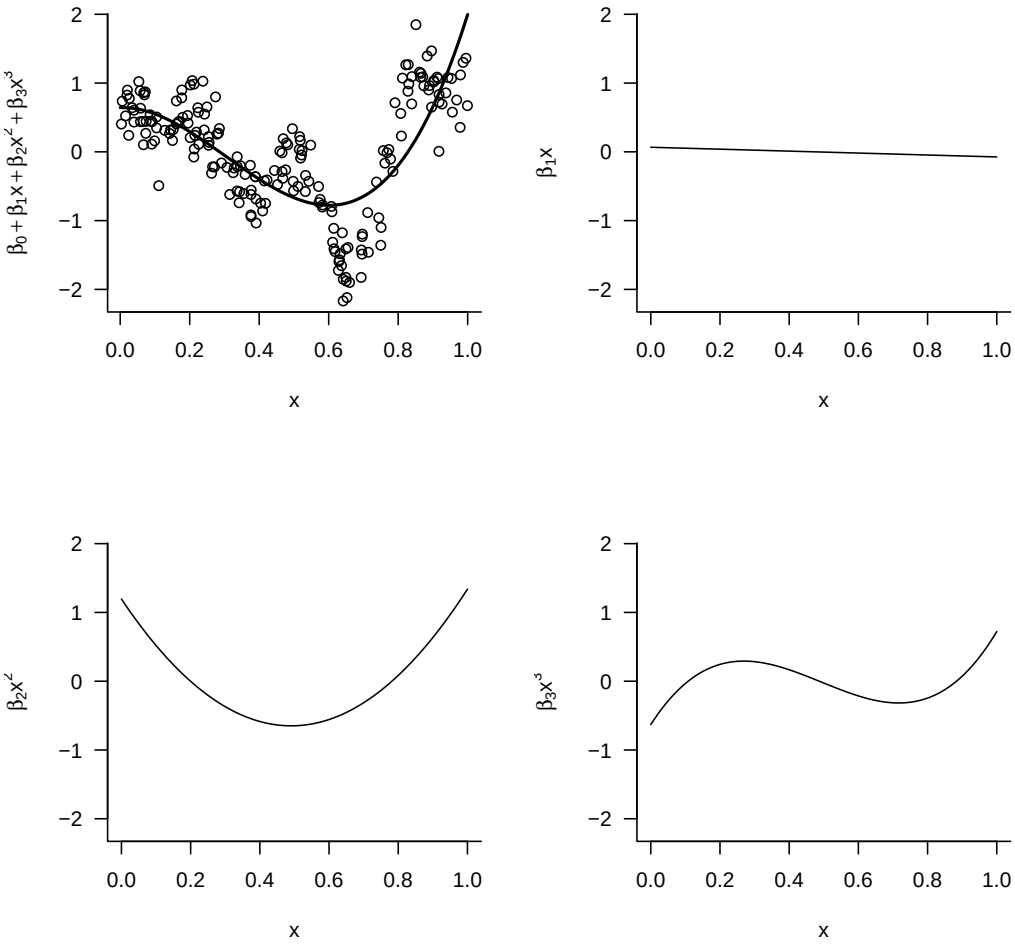


Fig. 2.1 Example fit to simulated data. The model fitted is a cubic polynomial. The top left panel shows the data and the model fit. The remaining panels (left to right, top to bottom) show the linear, quadratic and cubic components of the fit.

Clearly the fit in Figure 2.1 does not capture the trend very well and a higher degree polynomial would do better. It is possible to use AIC (Akaike, 1998) as a model selection criteria to find a suitable degree of polynomial. If this is done for the data in Figure 2.1, we get the fit in Figure 2.2 where a polynomial of degree 12 gave the fit with the lowest AIC.

Although polynomial functions are flexible, it can take a polynomial of high degree to adequately capture a nonlinear trend as seen in Figure 2.2 where a polynomial

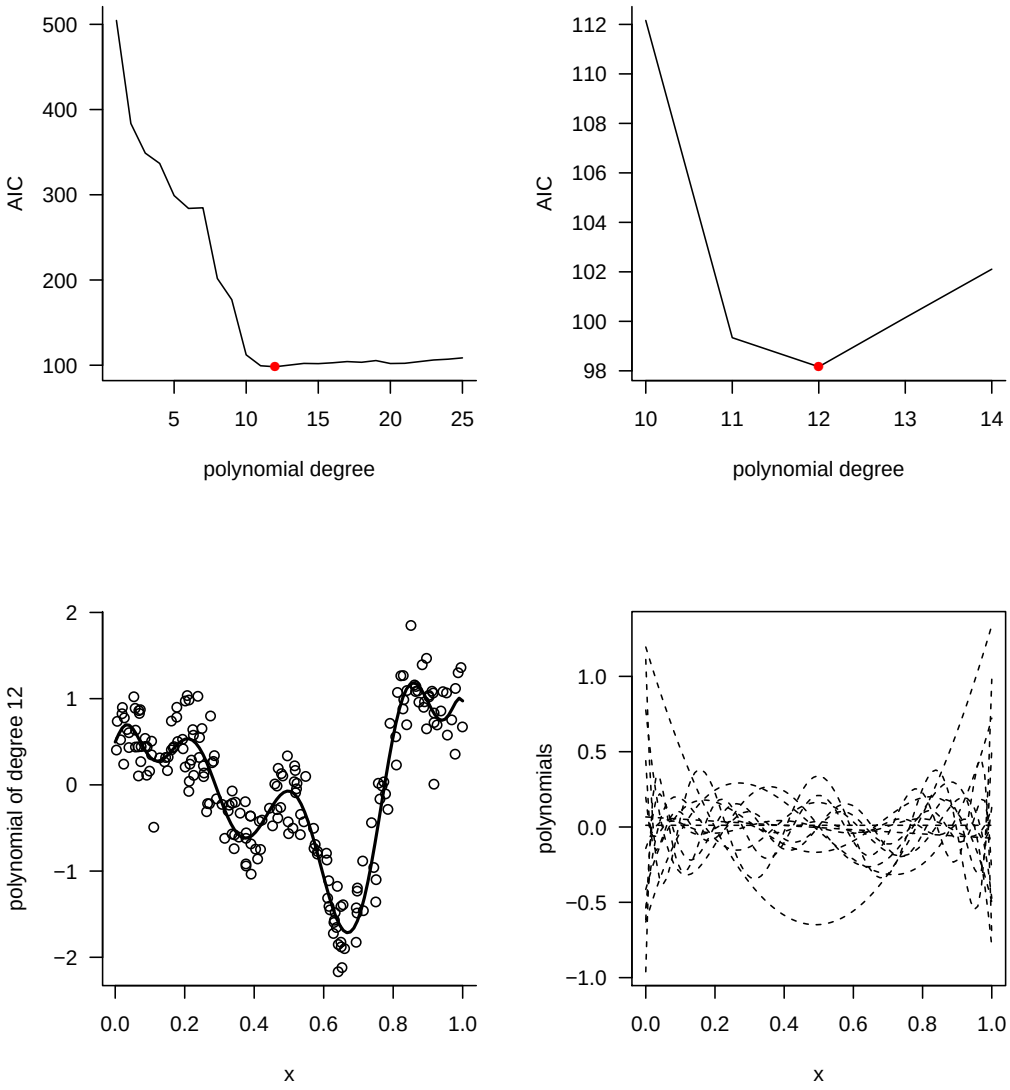


Fig. 2.2 Example fit to simulated data. The model fitted is a polynomial of degree 12. The top panels shows the AIC value for models with polynomial degree $1, 2, \ldots, 25$. The bottom left panel shows the data and the model fit. The bottom right panel shows all of the 12 polynomial components, the sum of which results in the fit shown on the left panel.

of degree 12 was selected via AIC. In the next section alternative families of basis functions called splines are presented.

2.2.1.3 Spline models

Splines are often written in shorthand as

$$f(\boldsymbol{x}) = \sum_{j=1}^{k} \beta_j B_j(\boldsymbol{x}) \tag{2.9}$$

where $B_j()$ is a function that produces the jth basis function of the spline. In the case of a raw polynomial in (2.7) we would have $B_j(\boldsymbol{x}) = \boldsymbol{x}^j$. There are a wide variety of options for basis function choice. Some common basis functions are Fourier basis functions and so called B-splines or basic splines (Eilers and Marx, 1996). I will discuss B-splines as they will be referred to in a later section.

There are several types of B-splines determined by their degree (Figure 2.3). Degree zero B-splines are essentially step functions, whereas degree 1 splines are spikes. It is not until degree 2 B-splines that things start to look smooth. The most common B-spline is the cubic (degree 3) B-spline. The reason for this is related to the degree of smoothness of the final function: cubic B-splines produce curves with continuous 2nd derivatives, where quadratic B-splines can only produce curves with continuous first derivatives. Figure 2.4 shows the best AIC fit of a cubic B-spline with evenly spaced basis functions: the location of the basis functions on the x axis is governed by a value known as a knot. The word knot is used because traditionally splines were made up of piece-wise polynomials *tied* together. Also shown are the individual basis functions scaled by the estimated coefficients to show how they contribute to the final fit. Notice the local influence of each basis function. In Section 2.2.2 we will see examples of spline basis functions that do not require knot position to be chosen.

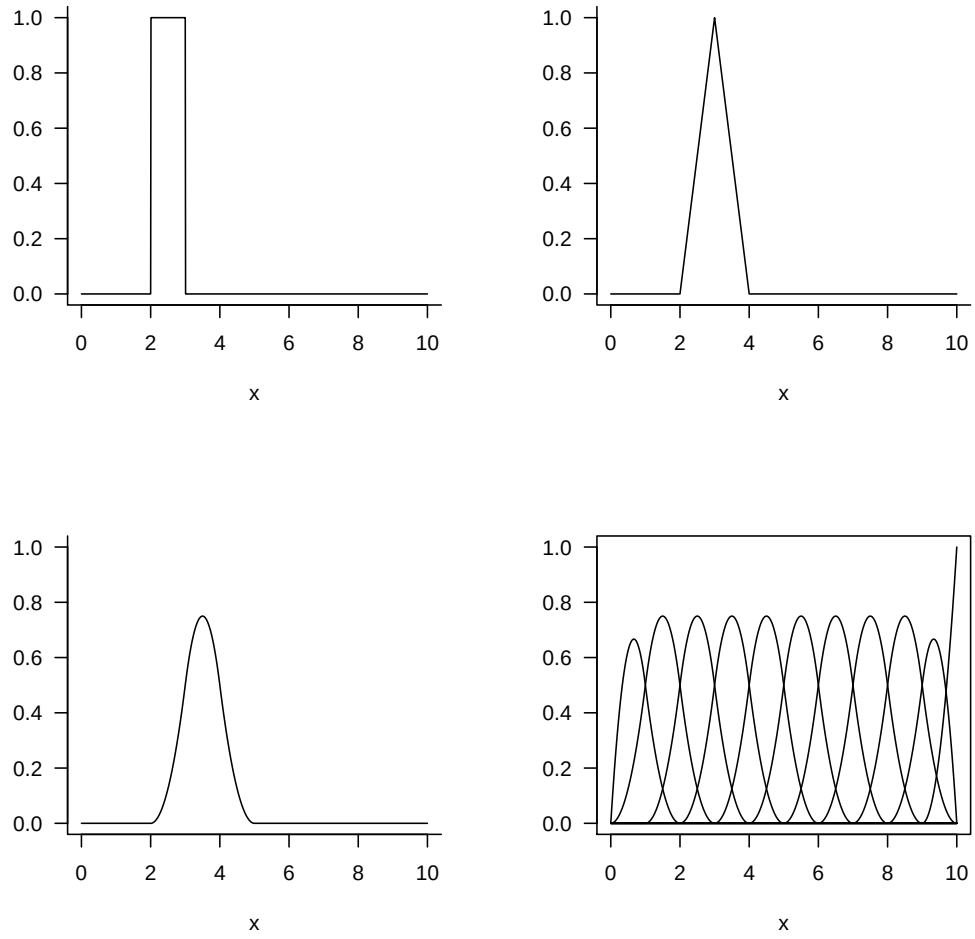


Fig. 2.3 Example of B-spline basis functions, from left to right, top to bottom: degree 0, degree 1, and degree 2 basis functions, and a complete set of 12 degree 2 basis functions for 9 evenly spaced internal knots.

Because splines are linear combinations of (basis) functions which are derived from the values of a covariate they can always be written in compact form

$$f(\boldsymbol{x}) = \sum_{j=1}^{k} \beta_j B_j(\boldsymbol{x}) = \boldsymbol{X}\boldsymbol{\beta} \tag{2.10}$$

and so, estimating the spline coefficients is exactly the same problem as estimating coefficients of explanatory variables for linear models.

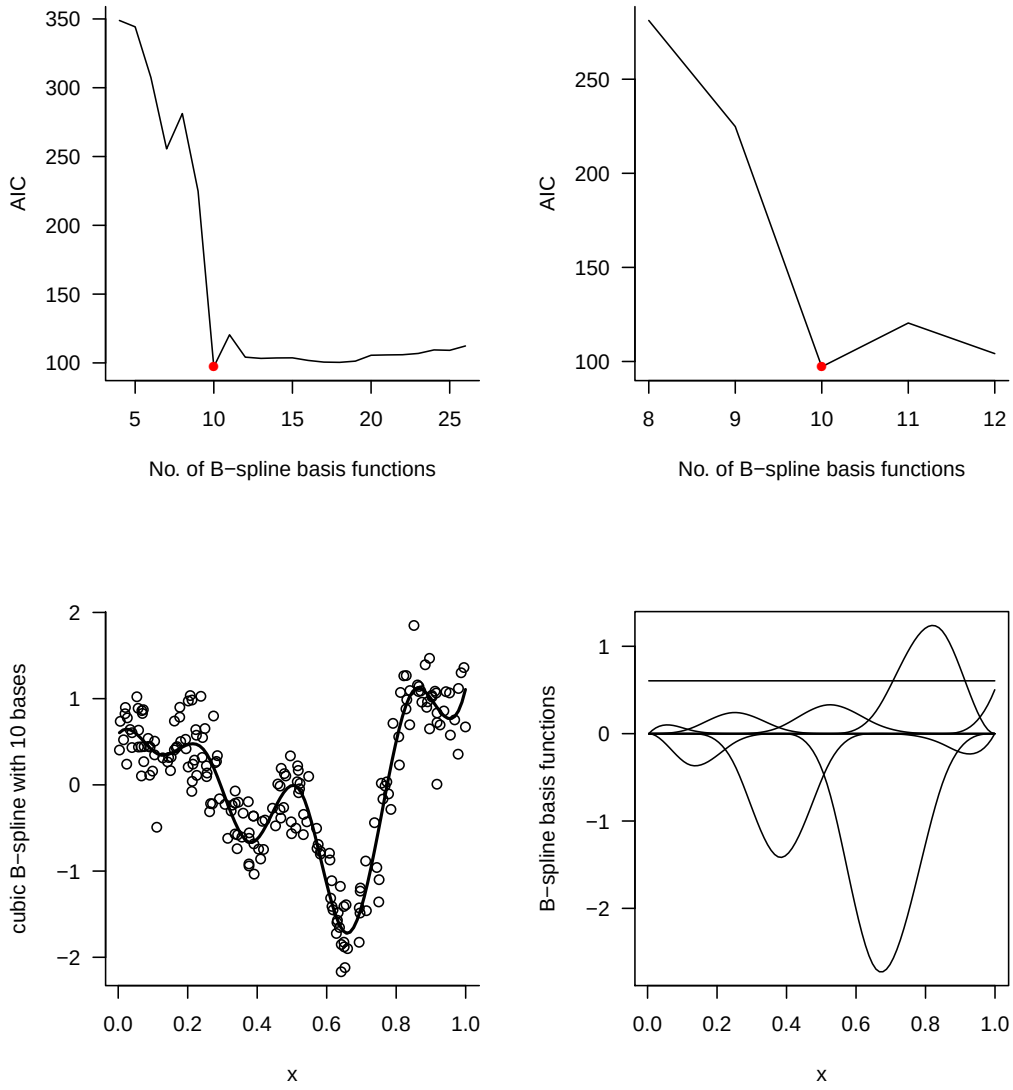


Fig. 2.4 Example fit to simulated data. The model fitted is a cubic B-spline composed of 10 basis functions. The left panel shows the data and the model fit. The right panel shows the contribution of each of the 10 basis functions, the sum of which gives the fit shown in the left panel.

2.2.1.4 Penalised splines

In the previous subsections we saw that to capture a nonlinear function a large number of knots were required. In addition, there was the need to perform explicit model selection in order to choose the best number of knots. This is not very satisfactory in

several ways. In the case where there is one term to be modelled as a spline, then any test of whether the term should be included or not is not one test but many. In the previous examples 25 models were compared to find the model with the lowest AIC, this is by no means an efficient strategy, but the point remains. If there are a number of spline terms in a model, finding the correct number of knots for each spline means an exponential increase in the number of tests. For example, if the number of basis functions under consideration is limited to 3 for each spline (1, 2 and 3 basis functions, say), then for n splines the number of model fits is 3^n, that is, it is exponential in the numer of splines being considered.

An interpretation of the number of knots is that more knots means more complexity. So the problem of estimating the appropriate number of knots is in a more general sense the problem of estimating the appropriate complexity of the spline. An alternative way to approach finding a nonlinear relationship is to begin by defining the nature of the relationship. For the types of situations where spline models were appropriate, a suitable property would be a function that is smooth, and a good definition of smooth in this context is to have a well behaved second derivative. In this loose description the problem of estimating the appropriate complexity is that of estimating appropriate variability in the second derivative: the smaller the variability the less the complexity. The variability of the second derivative is often quantified in terms of the total curvature of a function, where curvature is given by the square of the second derivative. Hence the total curvature (or total complexity) of a function is

$$\int f''(x)^2 dx \tag{2.11}$$

Penalised regression proceeds by considering the above integral like a Lagrange multiplier. The appropriate constraint is that the above integral is finite. If this is true, then the function $f(x)$ has the property of being smooth. Hence the Lagrangian

for penalised regression is set up by fitting the model

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon} \qquad \text{subject to} \qquad \int f''(x)^2 dx < \infty \tag{2.12}$$

where $\boldsymbol{\epsilon}$ are independently normally distributed errors. This leads to penalised log likelihood equations of the form

$$-2 \log L(\lambda, \boldsymbol{x}|\boldsymbol{y}) = \phi(\boldsymbol{y} - f(\boldsymbol{x})) + \lambda \int f''(x)^2 dx \tag{2.13}$$

where $\phi()$ is loosely defined as $-\frac{1}{2}$ of the normal log-likelihood function. The impact of the penalty is two-fold: it penalises functions that deviate from a straight line (since $\int f''(x)^2 dx = 0$ implies $f(x) = ax + b$); and it reduces the degrees of freedom of the function f. This allows the use of functions that have a large number of parameters, hence a large *space* of models to choose from, but is governed by a single *smoothing* parameter, λ. For example, although there may be 40 parameters in the function f, depending on the value of the smoothing parameter, there may only be 5 *effective* degrees of freedom. That is, the penalised function is only as complex as a function with 5 parameters.

The functions considered for f are always parameterised as a linear combination of its parameters such as the splines of the previous section, or, f is linear in terms of a basis function representation. This means that the penalty term can always be written as a quadratic form

$$\int f''(x)^2 dx = \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} \tag{2.14}$$

This gives rise to the notion of the penalty matrix $\boldsymbol{Q}$. For B-splines with equidistant knots the penalty matrices are simple expressions based on the differences of the basis coefficients (due to the local behaviour of the basis functions). To see this consider a

zero order B-spline (Figure 2.3) with knot locations at the points $x_1, \ldots, x_n$. The penalty matrix can be derived by writing the differences $x_{i+1} - x_i$ in terms of a $n-1 \times n$-matrix $\boldsymbol{D}_n$, for example the first derivatives are given by

$$\begin{pmatrix} x_1 - x_2 \\ x_2 - x_3 \\ \vdots \\ x_{n-1} - x_n \end{pmatrix} = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \boldsymbol{D}_n \boldsymbol{x} \tag{2.15}$$

and second derivatives

$$\begin{pmatrix} (x_1 - x_2) - (x_2 - x_3) \\ (x_2 - x_3) - (x_3 - x_4) \\ \vdots \\ (x_{n-2} - x_{n-1}) - (x_{n-1} - x_n) \end{pmatrix} = \begin{pmatrix} x_1 - 2x_2 + x_3 \\ x_2 - 2x_3 + x_4 \\ \vdots \\ x_{n-2} - 2x_{n-1} + x_n \end{pmatrix} = \begin{pmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \boldsymbol{D}_{n-1}\boldsymbol{D}_n\boldsymbol{x} \tag{2.16}$$

and hence the penalty is the quadratic

$$(\boldsymbol{D}_{n-1}\boldsymbol{D}_n\boldsymbol{x})^T\boldsymbol{D}_{n-1}\boldsymbol{D}_n\boldsymbol{x} = \boldsymbol{x}^T \underbrace{\boldsymbol{D}_n^T\boldsymbol{D}_{n-1}^T\boldsymbol{D}_{n-1}\boldsymbol{D}_n}_{\boldsymbol{Q}} \boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} \tag{2.17}$$

where

$$\boldsymbol{Q} = (\boldsymbol{D}_{n-1}\boldsymbol{D}_n)^T\boldsymbol{D}_{n-1}\boldsymbol{D}_n = \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix} \tag{2.18}$$

To see the effect of the smoothing parameter on the resulting fit (Figure 2.5), four values of λ were used to penalise a B-spline based on 50 knots (recall the best fit from unpenalised splines used 13 knots).

2.2.1.5 GMRF models

In one sense, GMRFs (Rue and Held, 2005) provide a practical way to generalise penalised splines and allow for the use in regression of functional forms with properties other than restricted derivatives. In the smoothing examples in the previous sections the goal was to find a function that had appropriate complexity, where complexity had to do with the variability of the function's second derivative. We saw in the previous section that this penalty took a quadratic form. There are two important points to note. Firstly, a $p \times p$ quadratic penalty can be constructed from any $q \times p$ matrix $\boldsymbol{L}$, say, where $\boldsymbol{L}$ defines a linear equation,

$$\boldsymbol{Q} = \boldsymbol{L}'\boldsymbol{L} \tag{2.19}$$

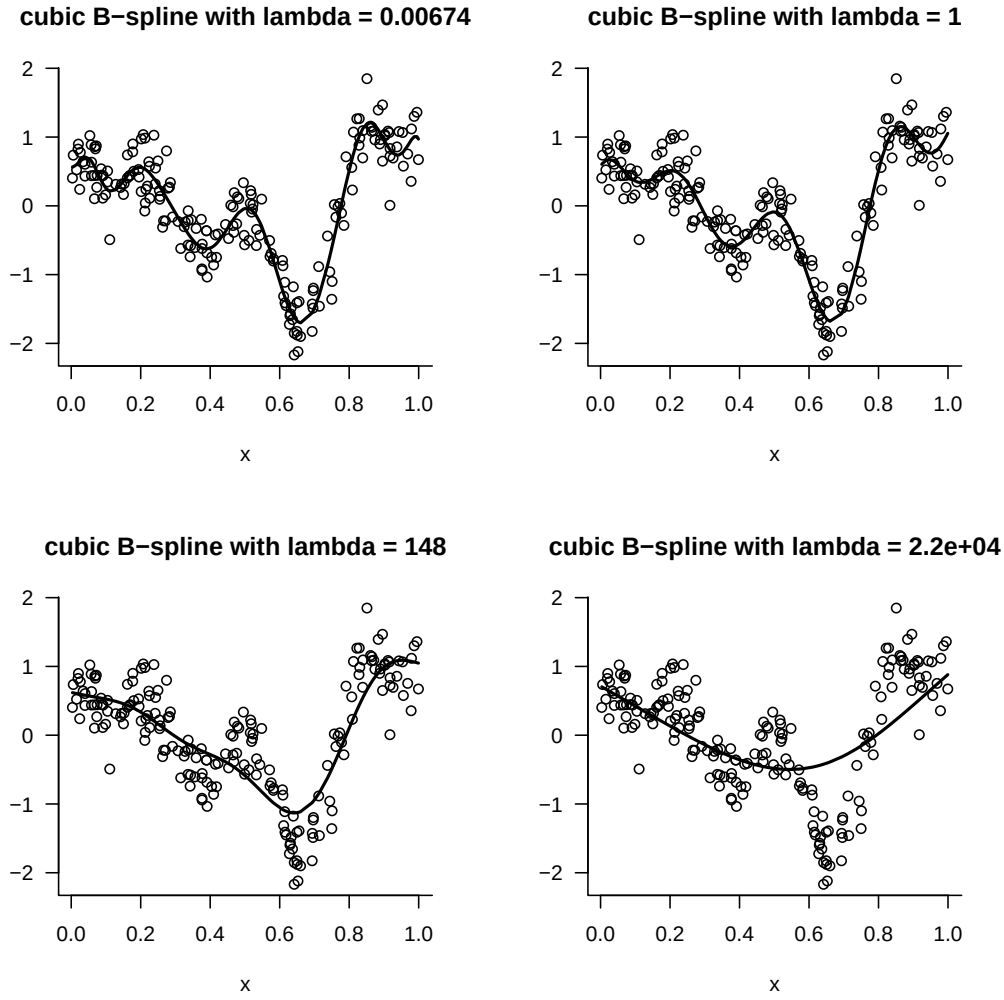


Fig. 2.5 Example fits to simulated data. The model fitted is a penalised cubic B-spline composed of 50 basis functions. The panels shows the data and the model fits for a range of smoothing parameter values; the lower the value the lower the penalty. The panels (left to right, top to bottom) show the effect of smoothing parameters with the values e^{-5}, e^{0}, e^{5} and e^{10}.

To see the generality of this, consider the previous example where $\boldsymbol{L}$ was the $n-2\times n$ matrix $\boldsymbol{D}_{n-1}\boldsymbol{D}_n$. Furthermore, recall the introductory example of a GMRF from Thiele (1880), this corresponds to $\boldsymbol{L}=\boldsymbol{D}_n$.

The second point is best seen by comparison. Consider a Bayesian regression model in which the likelihood is Gaussian,

$$\boldsymbol{y} \sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}\right) \tag{2.20}$$

and we assume that the model parameters $\boldsymbol{\beta}$ have a multivariate normal prior distribution,

$$\boldsymbol{\beta} \sim \mathrm{N}\left(0, \boldsymbol{\Sigma}\right) \tag{2.21}$$

Ignoring the hyperpriors of this for brevity, the log posterior is proportional to $\log f(\boldsymbol{y}|\boldsymbol{\beta}) + \log p(\boldsymbol{\beta})$, which can be written

$$\phi(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} \tag{2.22}$$

where ϕ is again $-\frac{1}{2}$ the normal log density function. This is now precisely in the form of the penalised log likelihood equations in (2.13), and hence, by comparison, a quadratic penalty matrix $\boldsymbol{Q}$ is equivalent to assuming that the model parameters are multivariate normal with variance matrix $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$.

Now, suppose we are interested in a seasonal pattern or a spatial pattern. GMRF precision matrices are a handy way to specify the desirable properties of a function. The remainder of this section will present some common and useful GMRF models fitted to a simulated dataset. Note that the applications of the GMRFs are as penalties applied to the smoothing coefficients. In the example given above the coefficients $\boldsymbol{x}$ follow what is known as a 2nd order random walk (in other areas it is known as a 2nd order Markov process because simple random walks have the memoryless Markov property). When the spline basis used is of zero order with equally spaced knots then the model is essentially a discrete process. GMRFs are always defined on discrete

lattices (Rue and Held, 2005), which correspond to a discrete process if zero order basis functions are used. It is possible to use a cubic B-spline basis with GMRF priors which allows a continuous process to be modelled by a discrete parameter vector which can then be modelled by a GMRF. The remainder of this section gives a few examples of GMRFs that can be used as penalties in a regression model

Seasonal models are models in which a pattern repeats, for example monthly mean temperatures. A desirable feature of a seasonal model is that the pattern can evolve. The following GMRF defines such a model for a seasonal time series x_i with a season length of m in which the variance σ^2 governs how much the seasonal pattern can change

$$\sum_{j=i}^{i+m-1} x_j \sim \mathrm{N}(0, \sigma^2) \tag{2.23}$$

defined for each $i = 1, \ldots, n - m + 1$. In matrix notation this is equivalent to saying that each element of the vector of linear combination ($\boldsymbol{Lx}$) is iid normal with variance

σ^2. With $m = 4$, the linear combination can is

$$\begin{pmatrix} 1 & 1 & 1 & 1 & & & & \\ & 1 & 1 & 1 & 1 & & & \\ & & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & 1 & 1 & 1 & \\ & & & & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ \vdots \\ x_{n-4} \\ x_{n-3} \\ x_{n-2} \\ x_{n-1} \\ x_n \end{pmatrix} = \boldsymbol{L}\boldsymbol{x} \tag{2.24}$$

where $\boldsymbol{L}$ is an $n - m + 1 \times n$ matrix. The penalty matrix for the GMRF with $m = 4$ is then

$$\boldsymbol{Q} = \boldsymbol{L}^T\boldsymbol{L} = \begin{pmatrix} 1 & 1 & 1 & 1 & & & & & \\ 1 & 2 & 2 & 2 & 1 & & & & \\ 1 & 2 & 3 & 3 & 2 & 1 & & & \\ 1 & 2 & 3 & 4 & 3 & 2 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & 2 & 3 & 4 & 3 & 2 & 1 \\ & & & 1 & 2 & 3 & 3 & 2 & 1 \\ & & & & 1 & 2 & 2 & 2 & 1 \\ & & & & & 1 & 1 & 1 & 1 \end{pmatrix} \tag{2.25}$$

Continuous 1st Order Random Walk models Rue and Held (2005) show that a 1st order random walk model can be defined by observing a continuous time process

known as a Wiener process at evenly spaced discrete time points. This observation allows for a model with unequally spaced discrete observations of a continuous time process to be defined. Because this link can be made, it means the models are equivalent regardless of the scale used. This is not the case for second order random walks which require more complex derivation (see Rue and Held, 2005).

Consider observations of this process x_i at locations s_i (assume that $s_1 < s_2 < \cdots < s_n$) and define the distance

$$\delta_i = s_{i+1} - s_i \tag{2.26}$$

The upper triangle of the precision matrix can be defined as

$$Q_{ij} = \kappa \begin{cases} \frac{1}{\delta_{i-1}} + \frac{1}{\delta_i} & j = i \\ -\frac{1}{\delta_i} & j = i+1 \\ 0 & \text{otherwise} \end{cases} \tag{2.27}$$

for $1 < i < n$, and the lower triangle through symmetry ($Q_{ij} = Q_{ji}$). Rue and Held define the boundaries Q_{11} and Q_{nn} via diffuse priors which results in $Q_{11} = \frac{\kappa}{\delta_1}$ and $Q_{11} = \frac{\kappa}{\delta_1}$.

1st Order Random Walks on irregular lattices The underlying principle behind the construction of random walk models is that the models can be written in terms of independent increments. For example, the 1st order random walk on a line can be written

$$x_i - x_j \sim \mathrm{N}(0, \kappa) \tag{2.28}$$

This same principle underlies the definition of a smoother often used for spatial variation; the regional smoother. This leads to a first order random walk model for spatial variation over regions with the following precision matrix (Rue and Held, 2005)

$$Q_{ij} = \kappa \begin{cases} n_i & j = i \\ -1 & i \sim j \\ 0 & \text{otherwise} \end{cases} \tag{2.29}$$

where n_i is the number of neighbouring regions to region i, and $i \sim j$ is true if region i neighbours region j. The regions need not be ordered in any particular way, however, the sparseness of $\boldsymbol{Q}$ can be maximised for particular orderings of regions depending on the connectedness of the regions. As an example, Figure 2.6 shows the regions of Scotland as defined by the Local Government (Scotland) act 1973. To build a regional smoother penalty matrix for Figure 2.6, the first row would have $Q_{11} = 6$ and $Q_{1j} = -1$ for $j = (2, 3, 4, 5, 7, 9)$ because region 1 neighbours regions 2, 3, 4, 5, 7 and 9. Likewise, the elements of $\boldsymbol{Q}$ can be built up with reference to the neighbourhood structure evident from Figure 2.6. A special case like region 10 could be treated as a neighbour of region 9 or alternatively be removed and considered separately.

2.2.2 Low rank approximations

Largely for computational reasons it is often necessary to work with low rank approximations to smoothers. A particular case in point is the implementation of the thin plate regression spline by Wood (2003). A thin plate spline is a full rank smoothing spline, that has as many parameters as there are data, and for general use the construction and fitting of such splines can be computationally restrictive.

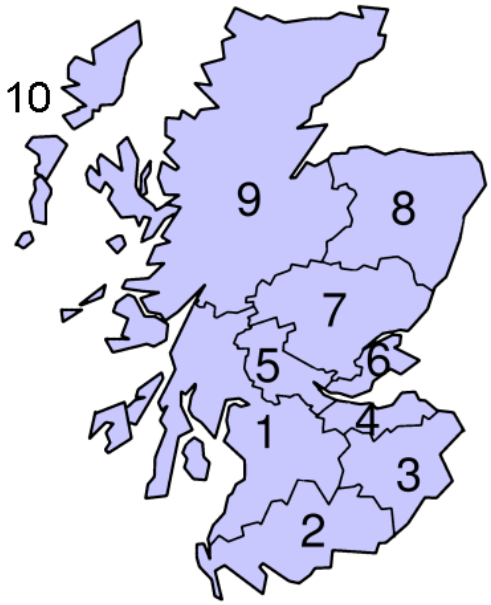


Fig. 2.6 The regions of Scotland as defined by the Local Government (Scotland) Act 1973. This map has a specific neighbourhood structure: region 1 neighbours regions 2, 3, 4, 5, 7 and 9, while region 8 neighbours regions 7 and 9.

Furthermore, the full model space spanned by a thin plate spline is, for the majority of applications, much larger than needed to get a suitable smooth fit. To these ends, Wood constructs low rank approximations of thin plate splines using low rank approximation techniques (Strang, 2009).

Low rank approximation is used in many statistical applications. A well known example is principle component analysis (PCA). In PCA, a low rank approximation of the empirical covariance matrix of a set of covariates is sought. To achieve this goal the data are rotated and scaled and possibly rotated again in order to find a reduced set of orthogonal bases (linear combinations of the original covariates) that explain a large part of the variation. Usually some sort of matrix decomposition is used, such as eigenvalue decomposition or the related singular value decomposition, and more recently CUR matrix decomposition (Mahoney and Drineas, 2009).

In smoothing applications, a very similar procedure leads to a reduced set of smoothing bases and a reduced set of parameters. Consider a regression spline model

where the spline bases are contained in the columns of a design matrix $\boldsymbol{X}$ and the coefficients β are constrained through the smoothing penalty matrix $\boldsymbol{S}$

$$y = \boldsymbol{X}\beta \qquad \text{with penalty} \qquad \beta^T \boldsymbol{S} \beta \tag{2.30}$$

The first step is to find a rotation of $\boldsymbol{X}$ that results in an orthogonal design matrix which is achieved through the so-called thin QR decomposition $\boldsymbol{X} = \boldsymbol{QR}$ where $\boldsymbol{Q}$ is $n \times p$ design matrix and $\boldsymbol{R}$ is a diagonal $p \times p$ matrix used to define a new parameter vector $\beta' = \boldsymbol{R}\beta$. This results in a transformed regression

$$y = \boldsymbol{QR}\beta \qquad \text{with penalty} \qquad \beta'^T \boldsymbol{R}^{-T} \boldsymbol{S} \boldsymbol{R}^{-1} \beta' \tag{2.31}$$

i.e.

$$y = \boldsymbol{Q}\beta' \qquad \text{with penalty} \qquad \beta'^T \boldsymbol{S}' \beta' \tag{2.32}$$

We now concentrate on the smoothing matrix. Following the procedure for PCA, an eigen decomposition of $\boldsymbol{S}'$ is $\boldsymbol{UDU}^T$ where $\boldsymbol{U}$ is orthogonal and $\boldsymbol{D}$ is diagonal matrix of eigenvalues arranged so that the eigenvalues decrease from top left to bottom right. If the rank of the smoother is less than p then the last eigenvalues will be zero. This final rotation results in a model in terms of $\beta'' = \boldsymbol{U}^T \beta' = \boldsymbol{U}^T \boldsymbol{R} \beta$

$$y = \boldsymbol{QU}\beta' \qquad \text{with penalty} \qquad \beta'^T \boldsymbol{UDU}^T \beta' \tag{2.33}$$

i.e.

$$y = \boldsymbol{X}''\beta'' \qquad \text{with penalty} \qquad \beta''^T \boldsymbol{D} \beta'' \tag{2.34}$$

So we now have a new design matrix of orthogonal basis functions and a diagonal penalty matrix. The nature of the penalty matrix implies that the first (left most) columns of X'' get penalised most heavily, and from the original definition this implies that these columns contain the high order bendy terms. So to create a low rank approximation of, say, rank $k < p$ would be to drop the first $p - k$ columns of X, the first $p - k$ elements/rows of the column vector β'' and penalise this with a diagonal penalty matrix which uses the $n - p + 1$th to the pth diagonal of D.

It is with this procedure that Wood (2003) constructs his thin plate regression splines, but there is no reason why these methods cannot be used to construct low rank approximations to other smoothing problems, for example for reducing the size of a GMRF smoothing problem.

2.2.3 Further examples of GMRF penalties

2.2.3.1 Spatio-temporal model

Although the GMRF models covered so far are usable in their own right, they can also be extended in a myriad of ways through the use of Kronecker products to give tensor products of GMRFs. A classic example of a tensor product is to allow a spatial effect to vary through time. Let $\boldsymbol{R}_s$ denote the structure matrix of a spatial smoother defined on a lattice containing n_s nodes, and let $\boldsymbol{R}_t$ denote a random walk model on a line with n_t nodes. These structure matrices can be combined to produce a GMRF that allows for a spatial model to evolve according to a (node-wise) random walk through time

$$\boldsymbol{Q}(\kappa_s, \kappa_t) = \kappa_s \boldsymbol{R}_s \otimes \boldsymbol{I}_{n_t} + \kappa_t \boldsymbol{I}_{n_s} \otimes \boldsymbol{R}_t \tag{2.35}$$

Essentially we have one component that provides a spatial model independently for each time point with an additional penalty in which each point in space should follow a random walk model. There is no restriction to the models through time. For example, it is possible to substitute the structure matrix $\boldsymbol{R}_t$ with a seasonal model or an AR1 model. It is not even necessary to restrict such constructions to GMRFs: low rank approximations could be used, as could conventional regression spline smoothers; but that is for a later section.

2.2.3.2 Vector time series

Consider four concurrent time series which are observations of different aspects of a temporal process. Suppose we wish to model each time series as an RW1 process. This is fairly straightforward to set up as a GMRF, and the precision matrix for this would have the form

$$\boldsymbol{R}_4(\boldsymbol{\kappa}) = \begin{pmatrix} \kappa_1\boldsymbol{R} & & & \\ & \kappa_2\boldsymbol{R} & & \\ & & \kappa_3\boldsymbol{R} & \\ & & & \kappa_4\boldsymbol{R} \end{pmatrix} \tag{2.36}$$

where $\boldsymbol{R}$ is a $n_t \times n_t$ structure matrix for a 1st order random walk. The overall model provided by this penalty is for independent random walks. However, as the observations are all of a different aspect of the *same* process, it is possible that they will be correlated. For the sake of example, let's assume that we require each 4-observation to have a uniform 4×4 correlation matrix $\boldsymbol{C}(\rho) = (1 - \rho)\boldsymbol{I} + \rho\boldsymbol{J}$, where $|\rho| < 1$ is a correlation parameter. In order to further constrain the 4-d random

walk model defined above, the combined penalty matrix is

$$\boldsymbol{Q}(\boldsymbol{\kappa}, \rho) = \boldsymbol{R}_4(\boldsymbol{\kappa}) + \boldsymbol{I}_{n_t} \otimes \boldsymbol{C}(\rho) \tag{2.37}$$

which is between a tensor product and independent GMRFs.

2.3 Fitting GMRF effects with mgcv

In the current implementation of mgcv, there are facilities to fit GMRF models using the mrf (Markov random fields) basis. This is done using

```
mrffit <- gam(y ~ s(id, bs = "mrf", xt = list(penalty = Q)), data = dat)
```

where Q is a suitable penalty matrix, and id is a factor whose levels must match the row names of Q. Quoting from the help file "*These [mrf smooths] are popular when space is split up into discrete contiguous geographic units (districts of a town, for example). In this case a simple smoothing penalty is constructed based on the neighbourhood structure of the geographic units.*". There are also help functions in mgcv to build a suitable penalty matrix from a list of polygons called poly2nb.

Obviously the focus is on using Markov random fields as regional smoothers, specifically 1st order random walks on irregular lattices (Section 2.2.1.5). However, in a practical application, the user runs into two problems: 1) the mrf basis constructor will not remove columns and rows in the penalty matrix if the associated nodes are not represented in the data, and 2) if you want to implement anything other than a regional smoother, you need to create the penalty matrix from scratch.

One of the objectives of this thesis is to demonstrate the ease with which GMRFs can be fitted using mgcv and to encourage the use of GMRF models by non-technical users. So far we have seen it is not immediately easy. However, by altering the

code associated with the mrf basis to not restrict the penalty to only having nodes represented in the data, and by writing small functions to generate penalty matrices for commonly used GMRF models, fitting GMRFs in mgcv becomes much more straightforward.

The good design of the mgcv package it was quite straightforward to amend the mrf code to create a new gmrf smoother type. In mgcv, the basis constructor functions all have the form: smooth.construct.xxx.smooth.spec, where xxx is substituted with the type of smooth specified in the gam formula. For example, the formula y ~ s(x, bs = "mrf") results in smooth.construct.mrf.smooth.spec being used to construct the smoother basis functions. The first lines of this function are

```
smooth.construct.mrf.smooth.spec <-
function (object, data, knots) {
    x <- as.factor(data[[object$term]])
    k <- knots[[object$term]]
    if (is.null(k)) {
        k <- as.factor(levels(x))
    }
    else k <- as.factor(k)
    ...
}
```

which in effect uses the unique values in the vector supplied to the function as factor levels. Later there is a check that the row names of the penalty matrix match the number of *knots* k. This is the single obstacle to opening up fitting with and predicting from GMRF models, and is easily solved by creating a new function which is identical to smooth.construct.mrf.smooth.spec, but the above lines are replaced with:

```
smooth.construct.gmrf.smooth.spec <-
function (object, data, knots) {
    k <- factor(rownames(object$xt$penalty),
                levels = rownames(object$xt$penalty))
    x <- data[[object$term]]
    x <- factor(x, levels = levels(k))
    ...
}
```

in this version of the function, which defines a new basis type `bs = 'gmrf'`, the knot levels are taken from the penalty, not the data, and nodes in the penalty are added as factor levels to the data. Essentially the penalty matrix drives the definition of the smoother. Fits using this basis will be referred to as extended mgcv fits. There is nothing especially complex happening here, it is just a work around for a sanity check incorporated into the mcgv code. In the case of GMRFs, this check is restrictive, and the `smooth.construct.gmrf.smooth.spec` function provides a less restrictive, but essentially equivalent constructor funtion to `smooth.construct.mrf.smooth.spec`. It is very much the same as fitting a smoother with knots at the data points, then predicting for an unobserved point in between two knots. The new point that is being predicted at did not exist explicitly in the original fit, but it it may as well have - its existence was guaranteed by the definition of the smoother. In a similar way, the GMRF penalty defines the relationship of the missing knots to the observed knots. The fact that the knots exist in the penalty but not the data is not an issue, and is accounted for in the calculation of the effective degrees of freedom. This brings out another sanity check performed by mgcv, which is to restrict the number of parameters to be less than the number of observations. This is also restrictive, because as long as the penalty is strict enough, there can be many more parameters than data - what is important is that the effective degrees of freedom is less than the number of data points. However, to remove this restriction requires making a whole new copy of mgcv, and this was considered too much of a deviation from the original.

For a simple demonstration of the increased flexibility afforded by the slightly altered basis constructor function, consider a model with random intercepts. The data for such a model can be simulated by

```
dat <- data.frame(id = rep(1:10, 20),
                  mu = rep(rnorm(10), 20))
dat $ y <- dat $ mu + rnorm(200, 0, 0.1)
```

The GMRF penalty in this case is the identity matrix. But, suppose we want to predict for a new group with confidence intervals? This can be done by specifying a bigger GMRF with a new node where we want a prediction,

```
Q <- diag(11)
rownames(Q) <- colnames(Q) <- 1:nrow(Q)
```

Note that the row names match up with the `id` vector. The model can now be fitted in the same way as the standard `mrf` model, but now using the `gmrf` basis constructor. The prediction for the new group can be made as usual, using the `predict` function

```
fit <- gam(y ~ s(id, bs = "gmrf", xt = list(penalty = Q)), data = dat)
pred <- predict(fit, newdata = data.frame(id = 1:11), se.fit = TRUE)
```

	1	2	3	4	9	10	11
fit	-1.2697	0.6079	-1.4223	-1.2352	-0.9216	-1.1404	-0.4620
se.fit	0.0229	0.0229	0.0229	0.0229	0.0229	0.0229	0.9151

In the above example the standard errors for the observed group means are 0.023 and the standard error for the unobserved group mean is 0.915.

The random effect GMRF is the simplest GMRF possible, and its construction was not taxing. However, the construction of more complex models such as the first order random walk (RW1) and the cyclic first order random walk (CRW1) can also be straightforward given a few tools. The most useful tool in defining GMRF penalties is a function that creates a difference matrix. The following function serves as a good building block

```
D <- function(n) {
  out <- diag(n)
  diag(out[,-1]) <- -1
  out[n,1] <- -1
  rownames(out) <- colnames(out) <- 1:n
  out
}
```

and produces a $n \times n$ matrix that computes the differences between successive nodes including the difference between the last node and the first node, that is it produces a cyclic difference matrix. For example,

```
D(5) # returns:
   1  2  3  4  5
1  1 -1  0  0  0
2  0  1 -1  0  0
3  0  0  1 -1  0
4  0  0  0  1 -1
5 -1  0  0  0  1
```

To compute the penalty for a cyclic 2nd order random walk (CRW2) model for 52 weeks of the year using this function, the function D can be applied recursively:

```
Dw <- D(52) # first order cyclic difference matrix
L <- Dw %*% Dw # applied twice gives a second order difference
Q <- t(L) %*% L # the quadratic penalty matrix
Q[1:10,1:10]
    1  2  3  4  5  6  7  8  9 10
1   6 -4  1  0  0  0  0  0  0  0
2  -4  6 -4  1  0  0  0  0  0  0
3   1 -4  6 -4  1  0  0  0  0  0
4   0  1 -4  6 -4  1  0  0  0  0
5   0  0  1 -4  6 -4  1  0  0  0
6   0  0  0  1 -4  6 -4  1  0  0
7   0  0  0  0  1 -4  6 -4  1  0
8   0  0  0  0  0  1 -4  6 -4  1
9   0  0  0  0  0  0  1 -4  6 -4
10  0  0  0  0  0  0  0  1 -4  6
```

To fit this model, it should be ensured that the data has a column with entries giving the week of the year as an integer to link with the penalty.

In addition to extending mgcv to make it easier to fit GMRF models, I have also developed functions to construct a variety of GMRF penalties:

`RW1`	1st order random walk
`CRW1`	Cyclic 1st order random walk
`RWn`	nth order random walk
`CRWn`	Cyclic nth order random walk
`Seasonal`	Seasonal model

The remainder of the section presents examples of models that can be fitted in mgcv as GMRFs.

2.3.1 Three cyclic smoother models

The first example is a cyclic RW2 model in which there are 100 nodes. The simulated data is presented in Figure 2.7 along with the mgcv fit (using low rank approximations to a cyclic cubic regression spline), a full rank cyclic 2nd order random walk GMRF and a low rank approximation to the same GMRF.

The data consists of 200 random observations of a process that occurs over 100 equally spaced points. The data is stored in an R data frame called `dat`, observations are stored in `y` and the location ids from where the observations were made are stored in `id`. Given this data, a fit using a cyclic cubic regression spline with 9 knots is

```
ccfit <- gam(y ~ s(id, bs = "cc", k = 10), data = dat)
```

The reason that $k = 10$ in the code is due to the way that the mgcv package treats the extra constraint for cyclicity.

The full GMRF fit requires one extra argument: the penalty matrix, $\boldsymbol{Q}$, which is supplied via an argument called `xt`. In addition, the number of 'knots' is set to -1 to indicate that the full penalty is to be applied, not an approximation. This would be similar to setting $k = 100$ for the spline fit. The code to fit this model is

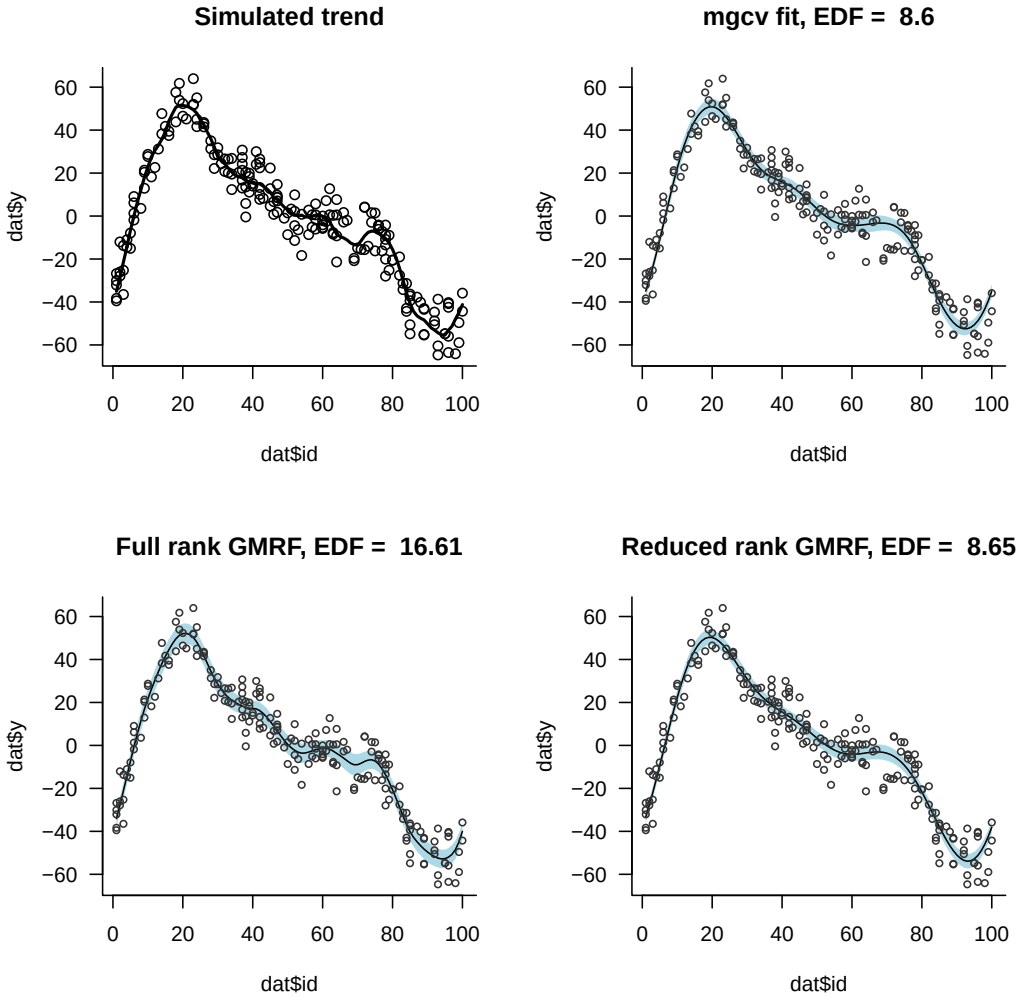


Fig. 2.7 Model fits to a simulated data set. The top left panel shows the simulated trend and simulated observations; the trend is a cyclic second order random walk. The top right panel shows a fit using a cyclic cubic regression spline with 9 knots, the bottom left shows a full rank cyclic second order random walk model, and the bottom right shows a low rank (rank = 9) approximation to a cyclic second order random walk. The estimated effective degrees of freedom (EDF) are shown above each fit.

```
gmrffit <- gam(y ~ s(id, bs = "gmrf", k = -1, xt = list(penalty = Q)),
               data = dat)
```

And finally, to fit a reduced rank GMRF, a non-zero value is supplied for the number of knots. Here we will use $k = 9$ for comparison.

```
gmrfapproxfit <- gam(y ~ s(id, bs = "gmrf", k = 9, xt = list(penalty = Q)),
                     data = dat)
```

The extended mgcv fits of all three models is shown in Figure (2.7). Notice that the full rank GMRF fit estimates 16 effective degrees of freedom, while the fits restricted to 9 degrees of freedom give an estimate close to the upper bound. This indicates that when an effective degree of freedom is near its upper bound it is likely that the data are being oversmoothed (the fits are too smooth) and that there is a higher signal to noise ratio (more signal in the data) than the current fit is capable of representing.

2.3.2 INLA comparison: drivers data

This example is originally given in Harvey and Durbin (1986) and appears in Rue and Held (2005) as an example of a hierarchical GMRF model. The model can also be fitted in mgcv as a penalised least squares problem. The data consist of 16 years of monthly statistics on road traffic injuries from 1969 to 1984, with the introduction of seatbelts occuring in February 1983. The data show a strong seasonal pattern and a trend over time. There is also interest in predicting forward one year into 1985. The model being fitted is

$$y_i \sim \mathrm{N}(s_i + t_i, \sigma^2) \tag{2.38}$$

where the observations are the square root counts, s_i is a seasonal trend and t_i is a second order random walk. The data is contained in the R package INLA

```
require(INLA)
data(Drivers)
```

The drivers data has the following structure:

```
str(Drivers)
'data.frame': 204 obs. of  4 variables:
 $ y       : int  1687 1508 1507 1385 1632 1511 1559 1630 1579 1653 ...
 $ belt    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ trend   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ seasonal: int  1 2 3 4 5 6 7 8 9 10 ...
```

There are 204 rows in the data, but the last 12 rows do not contain observations and so will be forecasts when prediction are made from the fitted models. The first task is to create the penalty matrices for the seasonal and trend models, this is done as follows

```
# seasonal model
m <- 12
n <- nrow(Drivers)
L <- t(sapply(1:(n-m+1), function(i) as.numeric(1:n %in% i:(i + m-1))))
Qs <- t(L)%*%L
rownames(Qs) <- colnames(Qs) <- Drivers $ seasonal

# RW2 model
D1 <- -1 * diag(n)[-1,]
diag(D1) <- 1
D2 <- D1[-1,-1] %*% D1
Qt <- t(D2) %*% D2
rownames(Qt) <- colnames(Qt) <- Drivers $ trend
```

The inla model is fitted as follows

```
# Define the formula
i1 <- inla(sqrt(y) ~ belt +
             f(trend, model="rw2") +
             f(seasonal, model="seasonal", season.length=12),
             family = "gaussian",
             data = Drivers,
```

while the model in mgcv is fitted

```
g1 <- gam(sqrt(y) ~ belt +
             s(trend, bs="gmrf", k=20, xt=list(penalty=Qt)) +
             s(seasonal, bs="gmrf", k=40, xt=list(penalty=Qs)),
           data = Drivers)
```

The extended mgcv fit was undertaken using low rank approximations to the seasonal and second order random walk GMRFs. This is necessary when using mgcv as there is a restriction that the upper bound on the model degrees of freedom has to be less than the number of data points; in INLA this is not necessary. If full rank GMRF models were fitted the number of GMRF model parameters would be $2(n + 12)$, because there are two GMRFs with a coefficient for each data point and we are predicting 12 points ahead for each GMRF, plus the intercept and the seat-belt term.

Following the example in Rue and Held (2005), the model was fitted with and without the additional term `belt` to model the introduction of seat-belts. The mgcv fit without the seat-belt covariate had an AIC of 715.472, while the model with the seat-belt covariate had an AIC of 699.912. The estimate of the seat-belt effect was -4.914 (SE 0.899) in mgcv compared to -4.88 (SE 0.895) in the INLA fit. In both models the introduction of a seat-belt effect reduced the effective degrees of freedom of the trend, and reduced the residual variability. A presentation of the fits is shown in Figure 2.8 along with a comparison between the fits (and predictions) from INLA (red) and mcgv (black). An apparent difference between the mgcv and INLA fits is the estimate of the seatbelt effect, however, it is likely this is a feature of how the trend effects were constrained for identifiability. Both the seasonal and trend components (plus the overall mean) are very similar in the INLA (black) and mgcv (red) fits, and both are strikingly different from the model without the seatbelt effect (blue) (Figure 2.9).

2.3.3 Correlated smoothers

The following example shows how two correlated smoothers can be fitted using mgcv.

A situation where the use of a complex GMRF model is beneficial is where two trends are correlated with each other. In this example the correlation is in the sense that

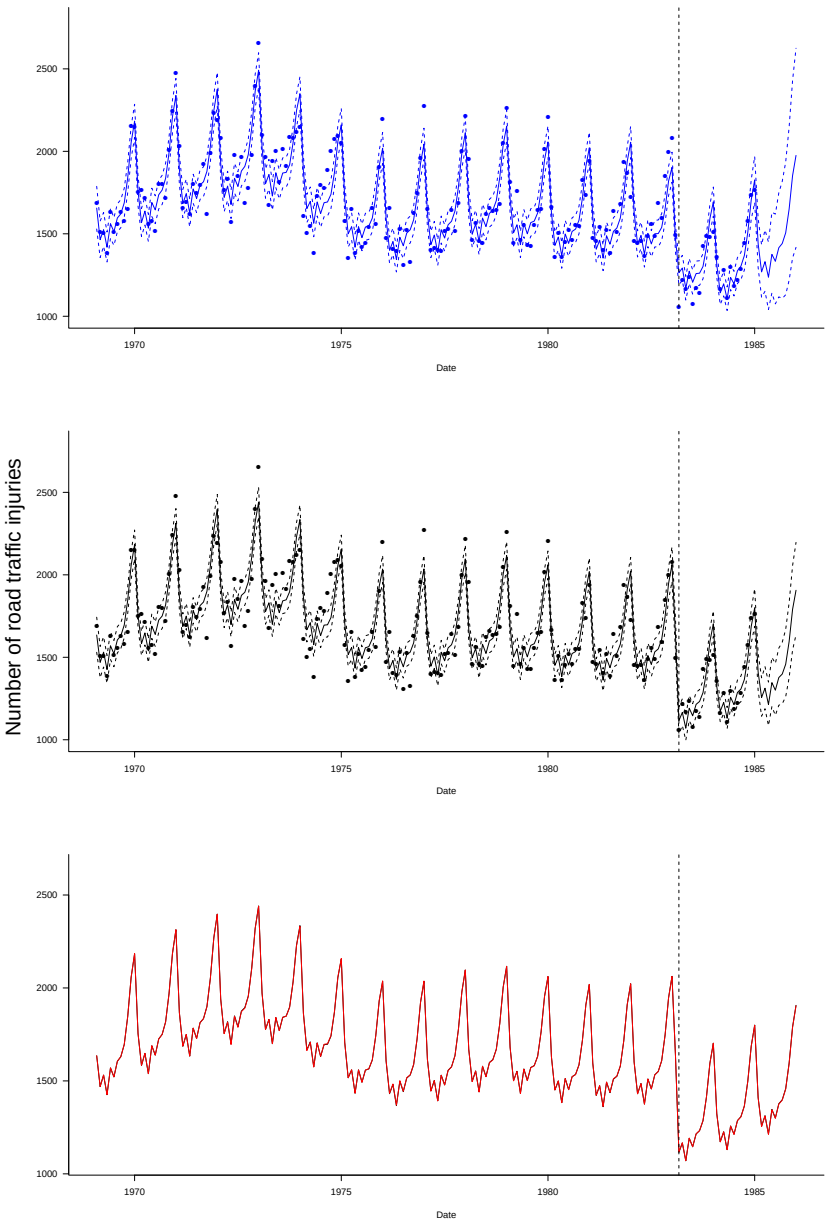


Fig. 2.8 Observed and predicted counts of road traffic injuries between 1969 and 1984 with a forecast into 1985. The top panel shows (in blue) a fit without an effect for the introduction of seat-belts, the middle panel shows (in black) the fit with the inclusion of the seat-belt covariate, and the bottom panel shows a comparison between the same model with seat-belt covariate fitted in INLA (red) to the model fit using low rank GMRFs in mgcv (black). Also shown as vertical dotted lines is the date of the introduction of seat-belts.

changes in gradient occur together and are to some degree in the same or opposite

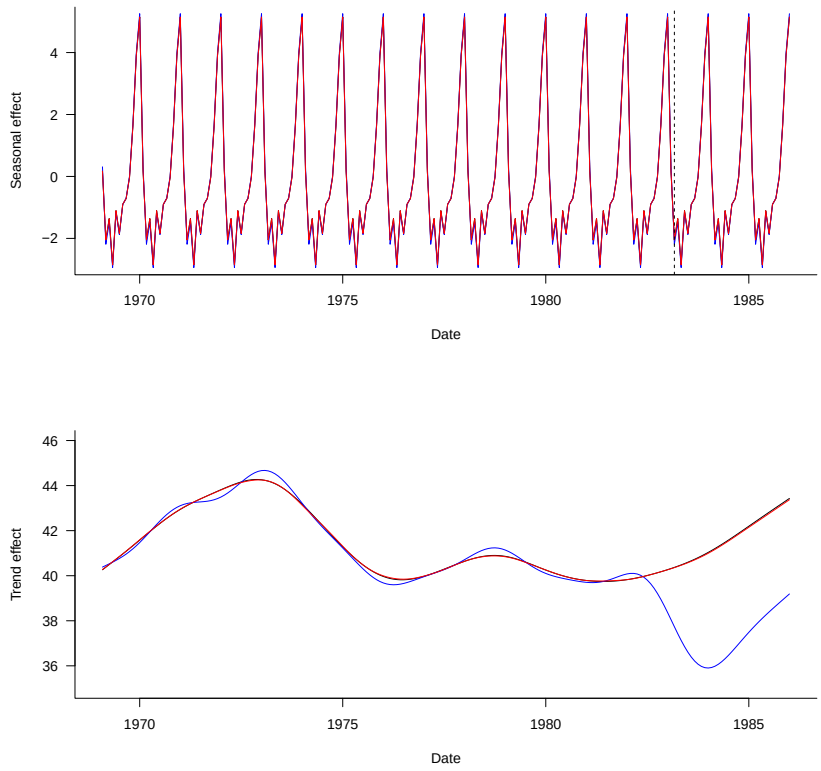


Fig. 2.9 Predicted seasonal and trend components of the reduced rank mgcv fit without (blue) and with (black) the seatbelt covariate and the INLA fit with the seatbelt covariate (red). Also shown as vertical dotted lines is the date of the introduction of seat-belts. Note that the trend component is plotted here as the trend plus overall mean.

direction. To model this behaviour a GMRF can be constructed as follows

$$\boldsymbol{Q}(\rho) = \kappa \boldsymbol{C}(\rho) \otimes \boldsymbol{Q} \tag{2.39}$$

where $\boldsymbol{C}$ is

$$\boldsymbol{C}(\rho) = (1-\rho)\boldsymbol{I} - \rho \boldsymbol{J} \tag{2.40}$$

where ρ is the unknown correlation parameter, $\boldsymbol{I}$ the identity matrix and $\boldsymbol{J}$ a matrix of ones. That is, $\boldsymbol{C}(\rho)$ is a correlation matrix with off diagonals equal to ρ. $\boldsymbol{Q}$ is a

second order random walk GMRF. The GMRF is defined in R with the following function:

```
Qfunc <- function(rho, n, d = 2) {
  # define the individual time series GMRF
  D <- Drw(n, order=2)
  Qt <- t(D) %*% D
  # define correlation structure
  Crho <- (1-rho)*diag(d) + rho*matrix(1, d, d)
  # combine
  Q <- solve(Crho) %x% Qt
  rownames(Q) <- colnames(Q) <- 1:nrow(Q)
  Q
}
```

This model corresponds to second differences that are penalised to be small (smooth) while also being penalised to be correlated. Thus, it is the gradients that are correlated here. The benefit of fitting such a model is to share information between the smooths. This can be useful in filling in missing data and making predictions, but also to reduce the effective degrees of freedom. For example, if a high degree of correlation is found between the curves then the effective degrees of freedom could be less than that of a model fitted with no correlation, however, the main benefit is in predictive performance.

Figure 2.10 shows some data simulated from two correlated 2nd order random walks in which the correlation parameter is high (0.9). In order to test the prediction the last 20 points are not observed in one of the processes. The code used to create the penalty matrix and simulate the data is as follows:

```
# simulate to time series, each with 100 points
n <- 100
rho <- 0.95
Q <- 0.1 * Qfunc(rho = rho, n = n)

set.seed(9384745)
x <- simQ(Q, rank = qr(Q)$rank)
id <- sample(1:length(x), 300, replace = TRUE)
```

```
y <- x[id] + rnorm(length(id), sd = 15) + ifelse(id > n, 30, 0)

dat <- data.frame(y = y, id = id)
dat $ group <- paste(as.numeric(dat $ id <= n))
dat <- subset(dat, id < 2*n-20)
```

and the code to fit the models and estimate the correlation parameter is:

```
# find the best value of rho by AIC
trans <- function(par) (exp(par)-1) / (exp(par)+1)
opt <- optimize(
  function(par) {
    rho <- trans(par)
    Q <- Qfunc(rho = rho, n = n)
    m0 <- gam(y ~ group +
                s(id, bs = "gmrf", k = 24, xt = list(penalty = Q)),
              data = dat)
    AIC(m0)
  }, c(-5, 5))
```

The correlation in the data is evident by noticing that changes in gradient of the two functions tend to occur together and in the same direction. Note that curves can have opposite signed gradients in this model; it is the change in gradient that is correlated, not the mean values of the curves. The two beneficial features of modelling correlation where it exists is shown well in this example (Figure 2.10): the effective degrees of freedom for the correlated model is 21.39, which is similar to the uncorrelated model which is 20.45, however the predictive power is better for the correlated model as shown by predicted trend and the width of the confidence intervals for the last 20 points.

2.4 Oscillating GMRFs

Seasonality is often modelled using oscillating functions, for example, Gomi et al. (2006) and Watson et al. (2001) modelled seasonal trends in river temperature using

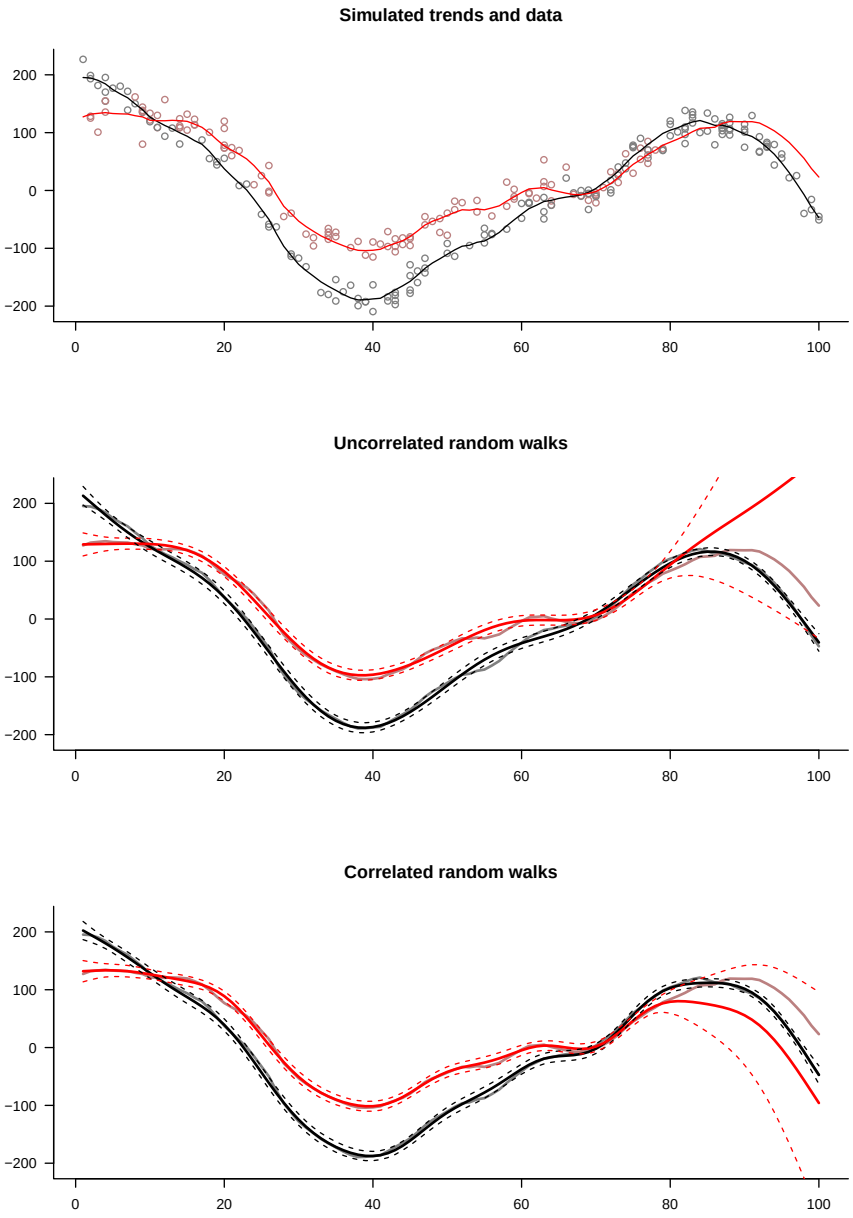


Fig. 2.10 Fits to simulated data. The data was simulated as observations from two trends (thin grey lines) which had correlated changes in gradient. For the trend plotted in red, no observations were made from the 80th time point. The middle panel shows a fit using independent smoothers. The bottom panel shows a model fit using a correlated second order random walk in which the correlation was estimated via outer iteration based on AIC. Both models were given the same total degrees of freedom (24). 95% point-wise confidence intervals are shown by dashed lines.

a sinusoidal function. I consider a number of more flexible alternatives based on GMRFs in this section.

Seasonality, used here to mean cyclicity, can be designed into a GMRF in more than one way. Consider a model for average daily temperature in which the cycle covers one year. A simple possibility is the 1st order cyclic random walk GMRF model (CRW1, Chapter 2 and Rue and Held, 2005). Here, the CRW1 model is based on penalising day to day changes in temperature where there is an explicit link between the last day in the year and the first day in the year. This is often referred to as a GMRF on a cylinder. In terms of derivatives, this model is based on penalising 1st differences, and can be written in terms of a differential operator $\boldsymbol{D}$ where, if $g(x)$ is a function defined for $x = 1, \ldots, 365$, then

$$\boldsymbol{D}g = g'(x) \tag{2.41}$$

In the discrete time case, the operator $\boldsymbol{D}$ is a matrix which computes the differences between each day, and because the 365th day is linked to the 1st day, the last row of $\boldsymbol{D}$ should computes this difference. The discrete (cyclic) differential operator is then the $n \times n$ matrix

$$\boldsymbol{D} = \begin{pmatrix} 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & \ddots & \ddots & & \\ & & & & 1 & -1 \\ -1 & & & & & 1 \end{pmatrix} \tag{2.42}$$

This is useful to describe several cyclic smoothers. The cyclic 1st order random walk is given by the quadratic penalty ($\boldsymbol{D}'\boldsymbol{D} = \boldsymbol{Q}_{crw1}$), which is invariant to the addition

of a constant term to the function g

$$\boldsymbol{Q}_{crw1} = \begin{pmatrix} 2 & -1 & & & -1 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ -1 & & & -1 & 2 \end{pmatrix} \tag{2.43}$$

The equivalent matrix for penalising 2nd differences is simply $\boldsymbol{D'D'DD} = \boldsymbol{Q}_{crw2}$.

A very useful GMRF for modelling periodic time series is known as the harmonic oscillator. The model penalises the 1st and 3rd derivatives together forming a penalty that is invariant to the addition of a phase shifted sine wave $a\sin(\omega x + \phi)$. The appropriate penalty is based on the famous wave equation that underlies string theory and was first discovered by the French scientist Jean-Baptiste le Rond d'Alembert. It is defined through a combination of differences to form a more general linear differencing operator L

$$\boldsymbol{L}g = \boldsymbol{D}^3 g + \omega^2 \boldsymbol{D}g \tag{2.44}$$

where ω is the desired period. The quadratic penalty for this is simply ($\boldsymbol{L}'\boldsymbol{L} = \boldsymbol{Q}_{harm}$). The structre of $\boldsymbol{Q}_{harm}$ is given below, where ω was set to equal $\frac{\pi}{365}$ to give a period of 365 days.

$$\boldsymbol{Q}_{harm} = \begin{pmatrix} 20 & -15 & 7 & -1 & & & -1 & 7 & -15 \\ -15 & 20 & -15 & 7 & -1 & & & -1 & 7 \\ 7 & -15 & 20 & -15 & 7 & -1 & & & -1 \\ -1 & 7 & -15 & 20 & -15 & 7 & -1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & -1 & 7 & -15 & 20 & -15 & 7 & -1 \\ -1 & & & -1 & 7 & -15 & 20 & -15 & 7 \\ 7 & -1 & & & -1 & 7 & -15 & 20 & -15 \\ -15 & 7 & -1 & & & -1 & 7 & -15 & 20 \end{pmatrix} \tag{2.45}$$

Both these models require a parameter vector $\boldsymbol{\gamma}$ with 365 elements - a parameter for each day - which would be penalised by a 365x365 matrix. For this moderate sized problem, the dimension of the smoothing matrix makes for slow computation. In order to speed up the fitting process a reduced rank smoother can be used.

Reduced rank smoothers provide a means to reduce the number of parameters in a model by reducing the maximum model space. However, the process involves computing the eigen decomposition of the penalty matrix, which again comes with computational cost. There is a short cut by noting the form of the new bases induced by diagonalising the penalty matrix. Recall that the *best* parameterisation of a smoother could be achieved by transforming the parameter space so that the smoothing matrix was diagonal. Wood (2006) called this the *natural parameterisation* for a smoother, and it is appealing because the new basis functions contribute independently to

the complexity of the fit. With a diagonal penalty matrix, the smoothing problem becomes almost a ridge regression.

Consider the GMRF smoother for a discrete time process composed of n time points. The basis functions for this model are coded by the identity matrix, each point in the process has a single basis vector, much like the basis vectors (i, j, k) for points in euclidean 3 dimensional space. Following the procedure for reduced rank smoothers (Wood, 2006), the penalty $\boldsymbol{\gamma}'\boldsymbol{Q}\boldsymbol{\gamma}$ can be written $\boldsymbol{\gamma}'\boldsymbol{U}'\boldsymbol{V}\boldsymbol{U}\boldsymbol{\gamma}$, by replacing $\boldsymbol{Q}$ with its eigen decomposition, so a transformed model with diagonal penalty matrix $\boldsymbol{V}$ is

$$\boldsymbol{y} = \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{e} \tag{2.46}$$

where $\boldsymbol{U}$ defines the new basis functions. The form of these basis functions is very informative and the first four are given in Figure 2.11, for two cyclic GMRF penalties: 1)the 1st order cyclic random walk and 2) the harmonic oscillator random walk. Also shown are the first four Fourier basis functions $(\cos(0), \sin(\frac{2\pi}{365}), \cos(\frac{2\pi}{365}), \sin(\frac{4\pi}{365}))$ for comparison.

Notice the similarity between the basis functions of the CRW1 and the harmonic oscillator in their natural parametrisation and the sine and cosine functions: reduced rank models are intimately related to penalised splines based on Fourier basis functions. In fact, the first four basis functions of the CRW1 penalty results in the first four discrete Fourier basis functions. The harmonic basis functions are very similar to the CRW1 apart from the 1st basis function, which has a sinusoid form.

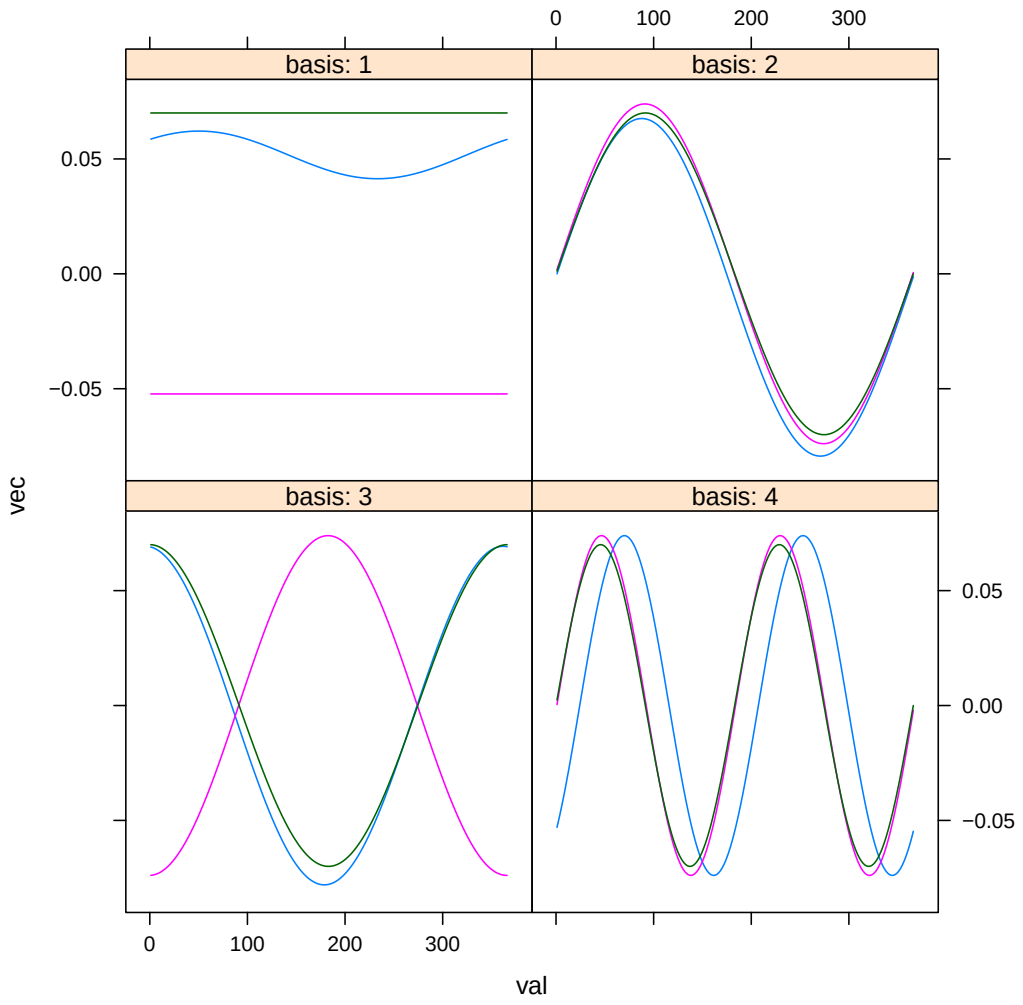


Fig. 2.11 The first four basis functions of two GMRF models transformed to their *natural parameterisation*, with a comparison to the Fourier bases. Pink: CRW1, Blue: harmonic oscillator, and green: Fourier basis functions.

3

A large scale removal method model for salmon fry

Preamble

In this chapter I develop a model for estimating fish abundance from electrofishing data. A two stage approach is used in which the first stage is the development of a model for capture probability, and the second stage is the development of a model for fish density. An R package named `ef` after **e**lectro**f**ishing, was written to aid the capture probability modelling stage, and allows the incorporation of standard linear models and splines; there is also the possibility of fitting penalised effects with fixed penalties. This package is not restricted to electrofishing data and can be used with any removal data. The approach was applied to a dataset covering the whole of Scotland and ranging over 17 years. This large scale dataset was a primary reason for adopting a two stage approach, since joint models of capture probability and density are cumbersome when working applied to large datasets.

The package has been developed on Github and can be found at http://www.github.com/faskally/ef. The latest version of the package is installed in R using the function `install_github` from the package `devtools`:

```
devtools::install_github("faskally/ef")
```

and to install the version used in this thesis use

```
devtools::install_github("faskally/ef@v1.0")
```

3.1 Introduction

3.1.1 Motivation

Large scale models of juvenile fish abundance are required to understand and predict spatial variability in fish productivity. Fish productivity varies at a range of spatial scales depending on water quality, food availability, hydraulic and sedimentary characteristics (Armstrong et al., 2003; Fausch et al., 1988). In fisheries management, such models can inform the development of juvenile assessment tools (Godfrey, 2005; SNIFFER, 2011) and provide an intermediate step in scaling stock-recruitment relationships between data rich and data poor catchments for the development of conservation limits (Wyatt and Barnard, 1997).

Models of juvenile fish abundance are typically based on electrofishing data collected at a large number of sites, often annually. Electrofishing is a fishing method in which an alternating electric current is introduced to a stream via the use of an electrode and an earth. The current stimulates muscle contractions in fish in the locality of the electrode which in turn causes them to swim towards the electrode. The fish are then netted and removed from the sample site. A predefined area of stream is traversed and all fish caught are removed. This constitutes a single electrofishing

pass. Often several passes are conducted in a consistent manner so that the sampling effort is constant from pass to pass. With successive passes the density of fish in the sample site decreases so the number of fish removed decreases. The combination of electrofishing passes made at a site on a single sampling occasion is referred to here as an *electrofishing event*. The number of fish caught on an electrofishing event depends not only on the fish density, but also on the capture probability (or catchability) of the fish. Capture probability could depend on factors such as habitat, sampling equipment, procedures and personnel. Any large scale model of fish density based on electrofishing data should therefore be underpinned by a suitable large scale model of capture probability.

Recent attempts to model spatial variability in juvenile fish abundance at large spatial scales have attempted to combine models of capture probability and spatial prediction using hierarchical Bayesian models (HBM) (Rivot et al., 2008; Wyatt, 2002, 2003). HBMs can readily accommodate complicated model structures and are well suited for modelling electrofishing data. They provide joint distributions of the estimates of capture probability and abundance, integrating uncertainty across multiple sites. However, complex HBMs often require considerable time (days) to fit, and can be difficult to explore, for example simply investigating conditional posterior expectations may not be practically feasible. In addition, because such models are often tailored for specific analyses, model exploration and comparison is time consuming and thus limiting. Maximum likelihood based alternatives could offer significant benefits in terms of rapid fitting. Furthermore where generalised additive models can be used, this provides model flexibility (Wyatt et al., 2007) and ease of implementation using standard software such as mgcv (Wood, 2006).

Unfortunately, spatial models of density and capture probability cannot be estimated simultaneously using GAMs because the resulting likelihood is not an exponential family. However, where capture probability is estimated separately, GAMs can be

used to model density. Consequently, many previous approaches have estimated capture probability independently for individual site visits (Huggins and Yip, 1997; Lanka et al., 1987; Otis et al., 1978) or, at the other extreme, assumed a constant capture probability across site visits (Bohlin et al., 2001). Approaches that focus on individual site visits can result in poorly defined capture probabilities (and thus density estimates) or provide no information on density in the case where no fish are caught. Conversely, an assumption of constant capture probability across all site visits is an over simplification and could lead to an under-estimate of uncertainty or to model bias in density estimates. Improvements in density estimates could be made if capture probability were to be modelled in a flexible framework where it was allowed to vary with covariates.

There have been many statistical developments in recent years that could improve large scale models of fish densities. For example, geostatistical approaches (Cressie et al., 2006; Peterson et al., 2013) or the use of Gaussian Markov random fields (Rue and Held, 2005; Wyatt, 2003). However, at present, there is no single software package that allows ready specification of models for the analysis of electrofishing data that could include variation at different spatial scales, over time and due to explanatory covariates. Such models would be desirable for fish abundance studies and are provided by the `ef` package developed as part of this chapter (see section 3.4) and the `gmrf` package developed for the thesis in general (see section 2.1).

3.1.2 Some historical context

Removal sampling has long established roots in the statistical literature. The original likelihood was reported by Moran (1951), and solutions to this likelihood developed by Zippin (1956). At the same time alternative ways to analyse such data had arisen, such as that of DeLury (1947) and the sampling based estimates of Horvitz

and Thompson (1952). From early on, it was recognised that the estimation of capture probability was crucial and the monograph by Otis et al. (1978) set out the the terminology still used today to refer to different forms of variation in capture probability.

Alternative models were also being proposed, seemingly driven from Bayesian considerations of the problem, but nonetheless giving rise to maximum likelihood estimates. The first such example is that of Carle and Strub (1978) who derived a compound mixture of the Moran multinomial with a beta distribution for capture probability. This model allows for overdispersion in the counts, in much the same way as a beta-binomial or dirichlet-mutinomial distribution. The original Moran (1951) likelihood was prone to give unstable estimates, for example, if no depletion was observed, the estimate of density tended to infinity due to the estimate of capture probability tending to zero. Allowing counts to be overdispersed helped this problem but did not solve it.

3.1.3 Chapter structure

This chapter describes the development of models that can be used to characterise, understand and predict spatio-temporal variability in fish abundance at the Scottish scale. This was achieved by developing two separate models: (1) a model to predict capture probability from electrofishing data and (2) a generalised additive model to predict spatio-temporal variability in fish abundance using explanatory covariates and the capture probabilities estimated from the catch probability model. The methods described were incorporated into an R package (ef) which is presented towards the end of the chapter.

Section 3.2 describes the Scotland-wide electrofishing dataset.

Section 3.3 develops a method to allow efficient modelling of capture probability and fish density from a large dataset such as that described in Section 3.2.

Section 3.4 presents the R package developed to fit the models described in the chapter.

Section 3.5 develops a model for capture probability using the methods described in Section 3.3 and the R package described in Section 3.4.

Section 3.6 develops a model for fish density conditional on the model estimated for capture probability in Section 3.5.

3.2 Data

3.2.1 Data availability

Electrofishing data was obtained from the Scottish Fisheries Coordination Centre (SFCC) database, Marine Scotland Science (MSS) FishObs database, and in spreadsheet format from the Scottish Environmental Protection Agency (SEPA) and Caithness District Salmon Fishery Board. Sites that were known to be stocked (as identified by the SFCC stocking code or knowledge of MSS staff) were excluded to simplify modelling requirements as were sites above impassable barriers. Because many data sources do not reliably obtain ages from scale reading, electrofishing data was resolved to life-stage (fry or parr) rather than individual age class. Only those data collected between 1997 and 2013 and between the 1st June and 21st November were used in subsequent analyses because few of the collated data lay outside of these time periods and thus there was insufficient information to inform large scale spatio-temporal models. The compiled data was also visually assessed in relation to

covariates and any extreme outliers removed. The resulting dataset consisted of 2353 sites and 4648 electrofishing events.

The spatial and temporal coverage of multi-pass electrofishing data in the final dataset is shown in Figure 3.1. In general, spatial coverage increased between 1997 and 2013, the exceptions being North West Scotland, where data coverage was better in earlier years and a recent reduction in sampling effort in the South East. There was a general paucity of data from the islands, with no data available from Orkney or Shetland. A schematic showing the temporal variability in the number of site visits provided by each Organisation is shown in Figure 3.2.

3.2.2 Covariates

Covariates were obtained for each sampling location. The selection of covariates was informed by previous habitat modelling studies (Fausch et al., 1988; Niemelä et al., 2000; SNIFFER, 2011; Wyatt, 2005) or relationships between covariates and processes that influence abundance, or capture probability. For example Altitude influences river temperature and in turn potentially affects fish productivity, whereas Organisation may be expected to affect capture probability through differences in equipment or personnel.

The selected covariates can be broadly assigned to four groups (1) spatial covariates (Hydrometric Area and Catchment), (2) habitat covariates (Altitude, Upstream Catchment Area, Distance to Sea, Gradient, Land-use and Channel Width (3) sampling covariates (Organisation) and (4) temporal, covariates (Year and day of the year).

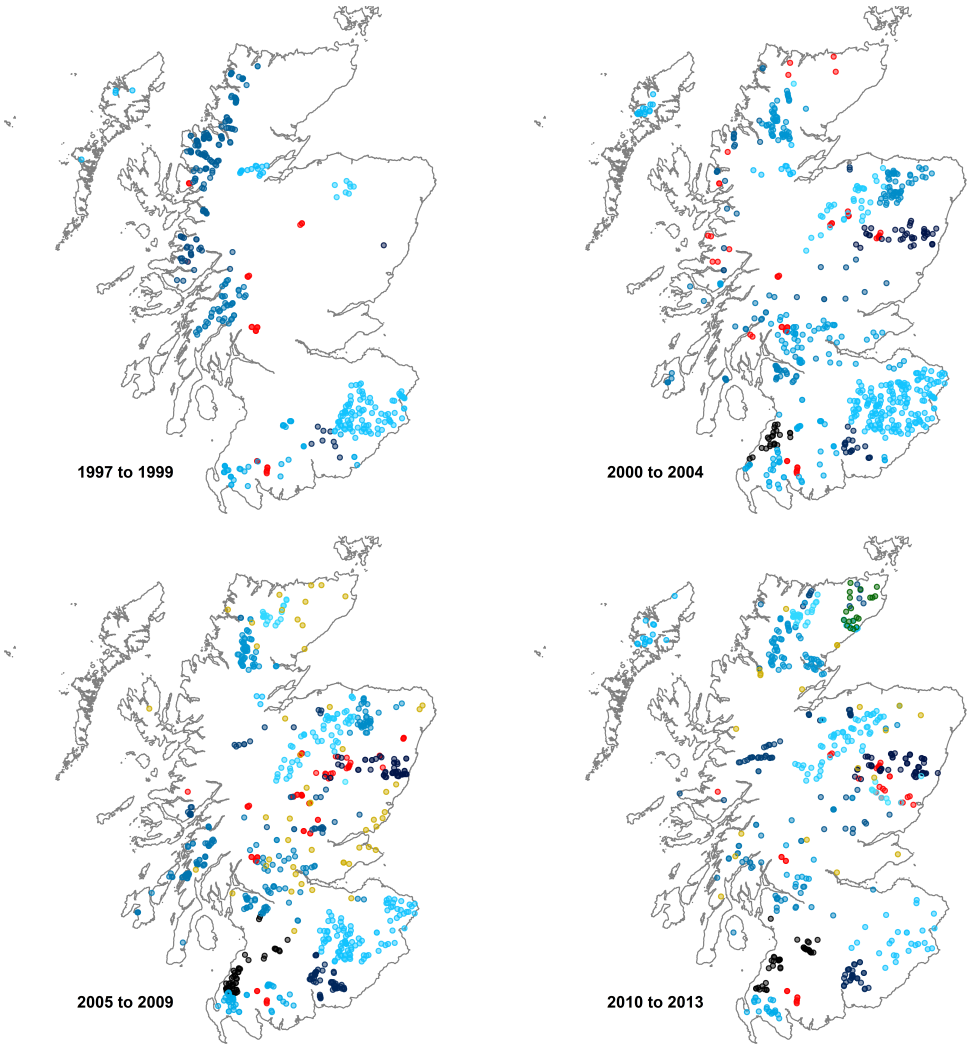


Fig. 3.1 Spatial and temporal data coverage (un-stocked sites with multi-pass electrofishing, below impassable barriers) between 1997 and 2013. Prior to 1997 there are too few data from too constrained an area for useful large scale model fitting. Data are colour coded by source MSS (red), SEPA (yellow), Other (green), SFCC (shades of blue).

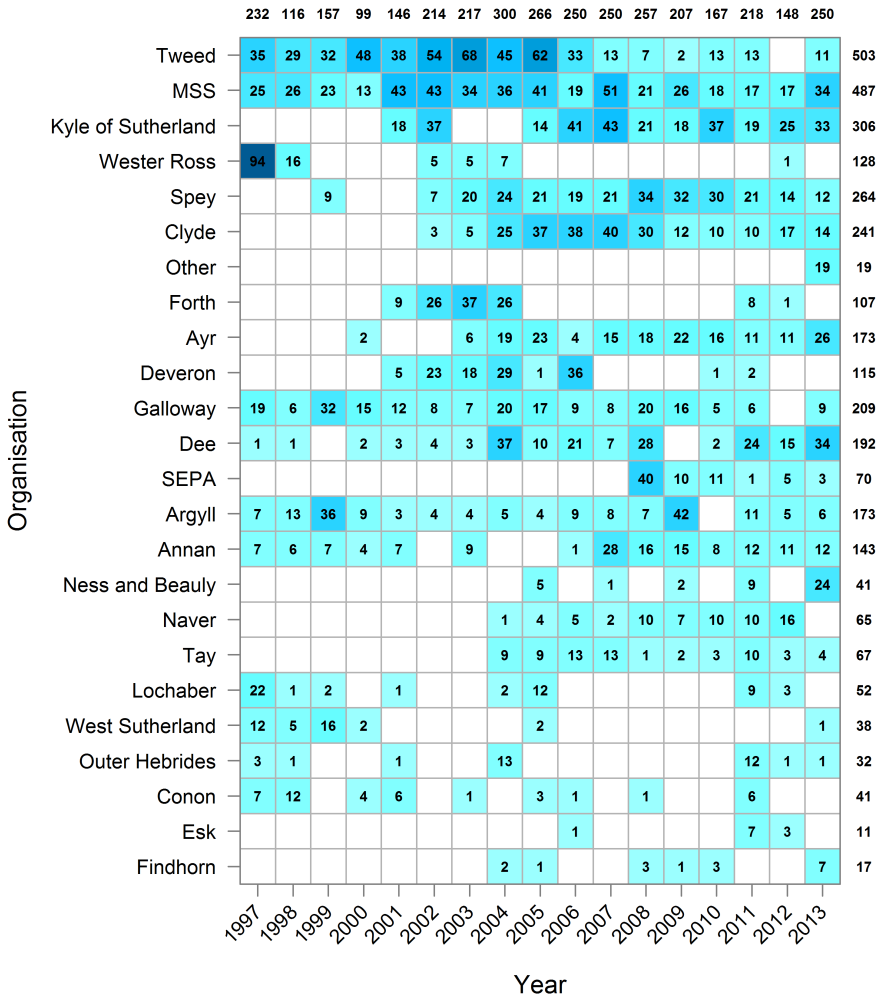


Fig. 3.2 Schematic showing the temporal coverage of data by provider. Organisations are ordered by the mean annual number of site visits. Total site visits by Year and by Organisation are given in the margins.

Spatial covariates

Hydrometric Area are administrative areas defined by SEPA that encompass a number of catchments. They represent an intermediate spatial scale useful for modelling correlated regional variability in fish abundance.

Catchment All sampling points were allocated to a SEPA Baseline or Coastal river catchment. SEPA baseline rivers are only those catchments with an area of >10km2. Catchments were too numerous for describing large scale regional variability in fish abundance, but were useful for describing finer scale deviations from regional trends.

Habitat covariates

Altitude was derived for each sample location from a national digital terrain model (DTM) provided by the Centre for Ecology and Hydrology (CEH).

Upstream Catchment Area was obtained using the *flow accumulation grid* dataset provided by CEH.

Distance to Sea (along river network) was derived using the SEPA rivers dataset.

Gradient was derived for each sample location from a national digital terrain model (DTM) provided by the Centre for Ecology and Hydrology (CEH).

Land-use metrics were obtained from the Ordinance Survey MasterMap dataset. A 50m circular buffer around each sampling location was used to derive percentage land-use characteristics.

Channel Width was derived from the area of water within the circular buffer (see above) used to derive percentage land use.

Sampling covariates

Organisation was included because of likely effects on catch probability such as sampling protocol or the equipment used. All data were assigned an organisation based on who collected the data. There was some spatial structuring in the data from fisheries trusts (see Figure 3.1). However, data provided by MSS and SEPA was geographically spread, and some trusts generated data in other trust areas thereby reducing the potential for confounding the effects of organisational and spatial variability.

Temporal covariates

Year and **Day of Year** (DoY) were included to allow for temporal variability.

3.2.3 Data coverage

Pairwise density plots of the data coverage in relation to covariates are presented in Figures 3.3 and 3.4. Latitude and Longitude have been included to identify the spatial coverage of available data, although they were not included as covariates during model fitting. The plots highlight where combinations of covariate values are well represented, for example low slopes and short distance to sea, but also where combinations are absent or rare, for example low elevation, high distance to sea. In addition, some combinations of land use variables are not possible because the landuse combinations need to sum to 100%.

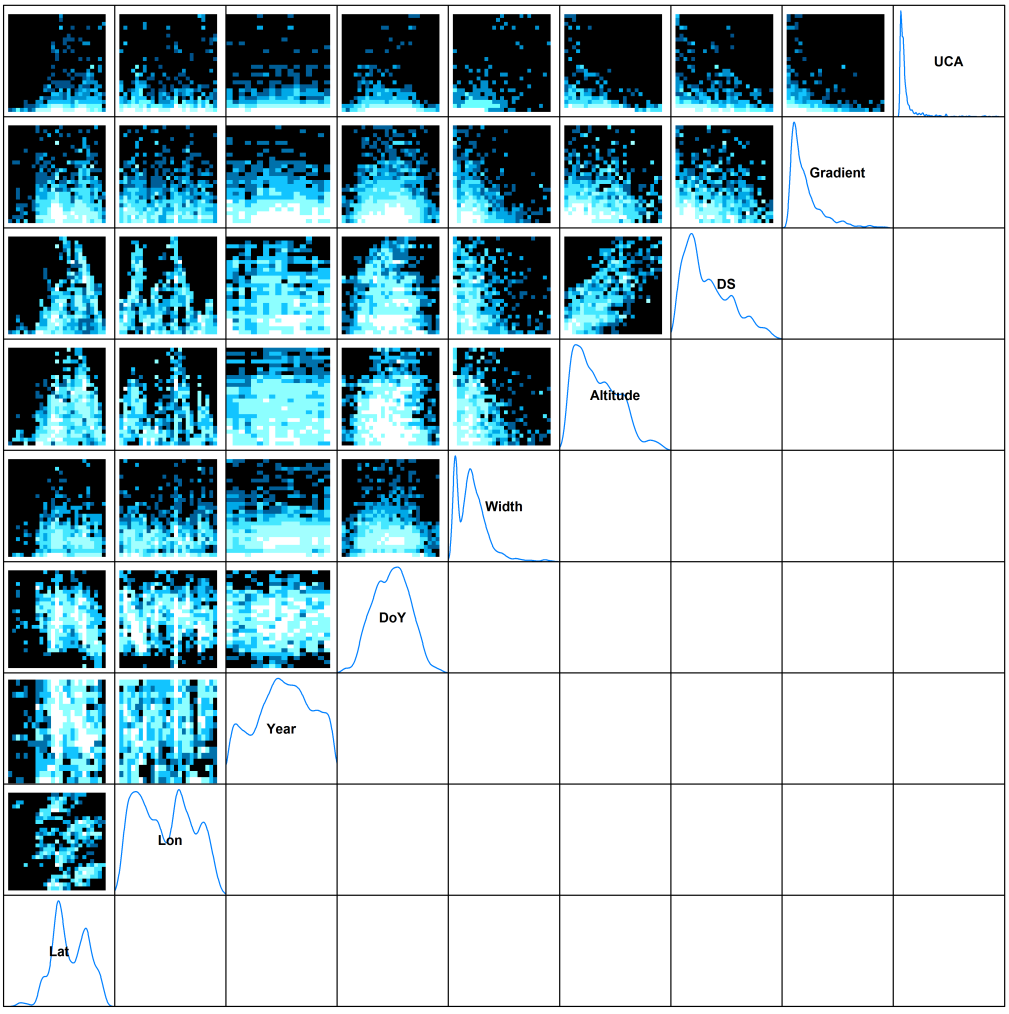


Fig. 3.3 Density plots showing the distribution of available data in relation to combinations of environmental covariates (white: lots of data, blue: few data, black no data).

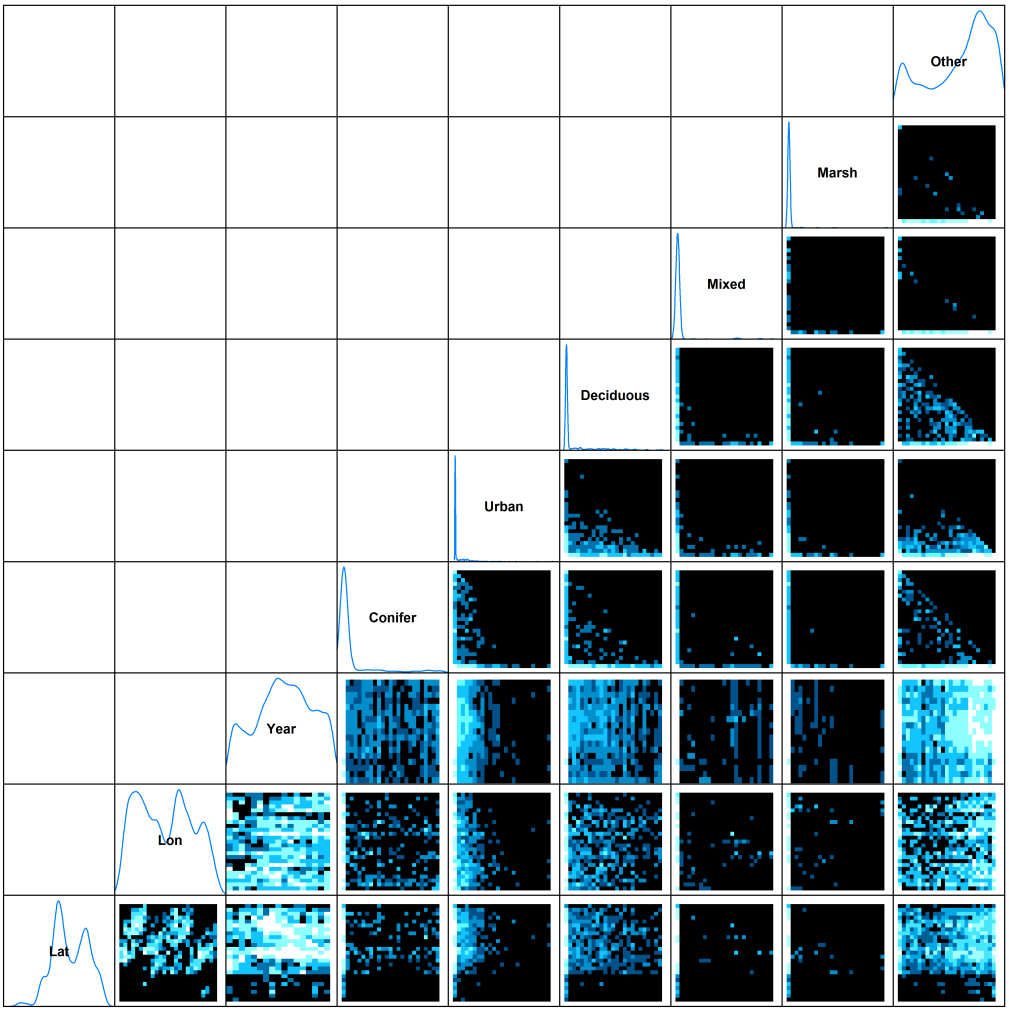


Fig. 3.4 Density plots showing the distribution of available data in relation to combinations of environmental covariates (white: lots of data, blue: few data, black no data).

3.3 A two-stage removal model

3.3.1 The likelihood for a single electrofishing event

An electrofishing event is defined as a complete electrofishing sample, composed of multiple fishing passes, taking place at a predetermined site. Sites are well defined areas of water, typically covering the entire breadth of the river, beginning and ending at defined locations. In some cases, nets spanning the river (stop nets) completely enclose the sampling site. In others, the boundaries are open and net immigration and emigration to the site are assumed to cancel each other out. On any one electrofishing event, there is assumed to be a well defined number of fish, denoted by N.

Throughout this section, it will be assumed that (for each electrofishing event) the capture probability of each of the N fish is the same value p, that the capture of any one fish does not affect the capture of another, and that the capture probability is the same in all electrofishing passes (that is, it does not change over time). In the following section I will extend this to allow capture probability to vary by pass and investigate the extent to which capture probability varies within pass. Constant capture probability implies that the number of fish caught on the first electrofishing pass, n_1, will be binomial

$$n_1|N, p \sim \text{Bin}(N, p) \tag{3.1}$$

It follows that the numbers caught on subsequent passes will also be binomial and will depend on the previous catches, for example, for the second pass

$$n_2|n_1, N, p \sim \text{Bin}(N - n_1, p) \tag{3.2}$$

The joint distribution of the data from multiple passes is the product of the individual binomial densities. For example, for three passes the joint distribution is

$$\pi(n_1, n_2, n_3|N, p) = \pi(n_3|n_2, n_1, N, p)\pi(n_2|n_1, N, p)\pi(n_1|N, p) \tag{3.3}$$

This extends to an arbitrary number of fishing passes s. The original likelihood was reported by Moran (1951) and solutions to this likelihood developed by Zippin (1956).

The joint distribution can be written in a compact form using a pair of statistics that are jointly sufficient for p and N: $T = \sum_{i=1}^{s} n_i$ and $X = \sum_{i=1}^{s}(i-1)n_i$

$$\pi(\boldsymbol{n}|N, p) = \frac{N!}{(N-T)!\prod_{i=1}^{s} n_i!} p^T (1-p)^{X+s(N-T)} \tag{3.4}$$

where $\boldsymbol{n} = (n_1, n_2, \ldots, n_s)$ (Otis et al., 1978). T is the total number of captures and X can be interpreted as the total number of times the captured fish evaded capture on previous passes. Conditional maximum likelihood estimates of N and q can be easily derived

$$\begin{aligned} \hat{N}|p, T &= \frac{T}{1-(1-p)^s} \\ \hat{p}|N, X, T &= \frac{T}{X + s(N-T)} \end{aligned} \tag{3.5}$$

The estimates above conveniently ignore the fact that N should be integer valued. This has led to some debate about the effects of bias (see the discussion of Zippin (1956) by Carle and Strub (1978)). An alternative approach which avoids this problem is to view N as a nuisance parameter, assumed to be an independent random variable with density $f(N; \theta)$, and then integrate (3.4) over this prior distribution on abundance. One sensible choice of f is the Poisson density,

$$f(N, \lambda) = \frac{1}{N!} e^{-\lambda A} (\lambda A)^N \tag{3.6}$$

where λ is the density of fish and A is the area electrofished. The Poisson assumption on N is appealing because it arises under the assumption that the distribution of individuals within some region is a homogeneous Poisson point process, which is often sensible for organisms which have a random distribution in space with an underlying constant density.

Estimation can now focus on the integrated likelihood

$$L(p, \theta; \boldsymbol{n}) = \sum_{N=T}^{\infty} L(N, p; \boldsymbol{n})\, f(N; \theta) \tag{3.7}$$

Note that the support of N is from T to infinity, as it is physically impossible for $N < T$. For the Poisson case, the integrated likelihood reduces to

$$\begin{aligned} L(p, \lambda; \boldsymbol{n}) \propto\ & p^T (1-p)^X \\ & \times \exp\left[-\lambda A \left\{1 - (1-p)^s\right\}\right] (\lambda A)^T \end{aligned} \tag{3.8}$$

It is a simpler matter to maximise (3.8) to give

$$\hat{\lambda}|p, T = \frac{T}{A(1-(1-p)^s)} \tag{3.9}$$

$$\hat{p}|N, X, T = \frac{T}{T + X + s(N-T)} \tag{3.10}$$

which are essentially identical to (3.5), see for example Seber (1982).

3.3.2 A conditional likelihood for p

With capture probability the focus, it is possible to regard the abundance N or density λ as a nuisance parameter and find a likelihood in terms of p alone. A simple and direct way of achieving this is to plug the conditional MLE estimate $\hat{\lambda}|p = \frac{T}{A(1-(1-p)^s)}$

into the joint likelihood (3.8) resulting in a conditional likelihood for p

$$L(p; \hat{\lambda}, \boldsymbol{n}) \propto \left(\frac{p(1-p)^Z}{1-(1-p)^s} \right)^T \tag{3.11}$$

where $Z = \frac{X}{T}$. Since X is the number of failed captures, Z is the average number of failed captures. The MLE of p can be derived using Z alone, however, the variance of the maximum likelihood estimate $\hat{p}$ depends also on T.

There are a number of ways to arrive at this likelihood. For example, Huggins and Yip (1997) derive it for a single site by considering the likelihood of capturing all fish and then conditioning on only the captured fish. An alternative is to assume a uniform distribution for $\log \lambda$ and integrate it out. All of these lead to the same conditional probability derived here and by Huggins and Yip (1997).

3.3.3 Extending the conditional model

In this section I extend equation (3.11) to include variability between passes, essentially following the framework of Huggins and Yip (1997). Let n_{ijk} be the number of fish of life-stage j caught on pass k in sample i, $i = 1, \ldots, N$, $j = 1, 2$ (fry, parr respectively), $k = 1, 2, 3$. Further, let p_{ijk} be the corresponding capture probability, assumed by necessity to be common across individuals within samples, life stages and passes because length information is not available. The log-likelihood of the data is then (subject to an additive constant):

$$L = \sum_{ij} (n_{ij1} \log(p_{ij1}) + n_{ij2} \log(q_{ij1} p_{ij2}) + n_{ij3} \log(q_{ij1} q_{ij2} p_{ij3}) - n_{ij.} \log(p_{ij1} + q_{ij1} p_{ij2} + q_{ij1} q_{ij2} p_{ij3})) \tag{3.12}$$

where $q_{ijk} = 1 - p_{ijk}$ and $n_{ij.} = n_{ij1} + n_{ij2} + n_{ij3}$. Changes in capture probability between samples are then incorporated by assuming that the capture probabilities are a linear logistic function of explanatory variables $\boldsymbol{x}_l$, $l = 1, \ldots, M$, i.e.:

$$\log\left(\frac{p_{ijk}}{1 - p_{ijk}}\right) = \beta_0 + \sum_l \beta_l x_{ijkl} \tag{3.13}$$

where x_{ijkl} is the value of $\boldsymbol{x}_l$ for sample i, life-stage j and pass k, and β_0 and β_l, $l = 1, \ldots, M$, are parameters to be estimated. Although the use of linear logistic models may appear limiting, the formulation above can incorporate categorical variables, interactions and non-linear responses through smoothing splines and spatial models with fixed degrees of freedom (Yee and Hastie, 2003).

3.3.4 Investigating overdispersion

In order to investigate the degree of within and between pass overdispersion the following procedure was used. More details of the application to the dataset are described in section 3.5. First a 'large' model is fitted to capture most of the systematic variation in the data and the log-likelihood of the large model is compared with the log-likelihood of a saturated model, in which there is a separate first- and second-pass capture probability for each sample and life-stage. This will show if the data are over-dispersed or not. That is, is there more variation in the data than expected given the multinomial likelihood. The overdispersion can then be partitioned into within- and between-sample overdispersion by also computing the log-likelihood of a 'sample-wise' model in which there is an extra parameter for each sample (i.e. logit $p \sim$ sample + terms in the large model. Due to the potentially large number of parameters, this model can be fitted by conditioning on the fitted values from the large model and estimating the sample effect for each sample in turn. The within-

and between-sample overdispersions can then estimated to be:

$$\Psi_{\text{within}} = \frac{2(L_{\text{saturated}} - L_{\text{sample}})}{2N - M_{\text{large}} - 1} \tag{3.14}$$

$$\Psi_{\text{between}} = \frac{2(L_{\text{sample}} - L_{\text{large}})}{N} \tag{3.15}$$

where $L_{\text{saturated}}$, L_{sample}, L_{large} are the log-likelihoods of the saturated, sample-wise, and large model respectively, and M_{large} is the number of linear terms in the large model. If Ψ_{between} is the dominant term, and since our focus is to produce a model that can be used to predict capture probabilities at the national level (i.e. across samples), Ψ_{between} can be used to adjust the BIC.

$$\text{BIC}_{\text{adj}} = -\frac{2L}{\Psi_{\text{between}}} + \log(N)(M + 1) \tag{3.16}$$

This should prevent any over-fitting of the variables that change between samples (such as the habitat variables), albeit at the loss of some power for detecting within-sample effects (those involving pass or life-stage).

3.3.5 The two stage approach

National scale models of juvenile abundance (or density) are based on data from multiple electrofishing events Wyatt (2005). Existing models are typically hierarchical Bayesian models (Rivot et al., 2008; Wyatt, 2003), which jointly model capture probability and density, but can take days to run (SNIFFER, 2011). This in itself is not a problem, but due to the size of the dataset and the diversity of data sources, it is common to have to rerun models with certain samples removed. Unfortunately, knowledge about which samples to remove is largely informed by exploratory model

runs. If models take days to run, then the whole process including exploratory modelling becomes prohibitive. Here, I develop an alternative two-stage approach in which capture probability is first modelled assuming different densities for each electrofishing event and then density is modelled conditional on the fitted capture probabilities. The two main advantages of this modelling process are that model selection is simplified, as only one aspect of the model (capture probability or density) is being tested at a time and that conditional modelling of density based on previously estimated capture probabilities allows the use of generalised additive models (GAMs) and structural additive regression (STAR) models (Fahrmeir et al., 2013) due to the standard form of the likelihood that results from this approach. This allows the incorporation of a wide range of penalised smoothers and spatial and random effects and the use of standard software (Wood, 2006).

Because the capture probability model (3.12) is not an exponential family model, it is not possible to fit this model using standard software. So, as part of this thesis I have developed an R package called `ef`, hosted on line at https://www.github.com/faskally/ef. The R package and its usage are described in the following section (3.4) on software.

Estimation of fish density is more straightforward. If we take the joint likelihood for capture probability and density for s single electrofishing pass given in (3.8) and condition on p we find that

$$T \sim \text{Pois}\left(A\lambda(1-(1-\hat{p})^s)\right) \tag{3.17}$$

which extends to multiple samples, lifestages and passes to

$$T_{ij} \sim \text{Pois}\left(\lambda_i A_i(\hat{p}_{ij1} + \hat{q}_{ij1}\hat{p}_{ij2} + \hat{q}_{ij1}\hat{q}_{ij2}\hat{p}_{ij3})\right) \tag{3.18}$$

where $T_{ij} = n_{ij1} + n_{ij2} + n_{ij3}$, with n_{ijk} being the number of fish of life-stage j caught on pass k in sample i, $i = 1, \ldots, N$, $j = 1, 2$ (fry, parr respectively), $k = 1, 2, 3$, $\hat{p}_{ijk}$ is the corresponding estimate of capture probability, assumed by necessity to be common across individuals within samples and $\hat{q}_{ijk} = 1 - \hat{p}_{ijk}$. And recall that A_i is the area fished for the ith sample

Therefore, conditional on $\hat{p}_{ijk}$, the total catch for lifestage and sample T_{ij} can be modelled as a Poisson GAM or STAR model. Density is then modelled using a log link for $E(T_{ij})$ where the estimate of capture probability $\hat{p}_{ijk}$ enters via an offset given by $\log(A_i(\hat{p}_{ij1} + \hat{q}_{ij1}\hat{p}_{ij2} + \hat{q}_{ij1}\hat{q}_{ij2}\hat{p}_{ij3})$,

$$\begin{aligned} \log\Big(E(T_{ij})\Big) &= \log\Big(\lambda_{ij}A_i(\hat{p}_{ij1} + \hat{q}_{ij1}\hat{p}_{ij2} + \hat{q}_{ij1}\hat{q}_{ij2}\hat{p}_{ij3})\Big) \\ &= \log\lambda_{ij} + \log\Big(A_i(\hat{p}_{ij1} + \hat{q}_{ij1}\hat{p}_{ij2} + \hat{q}_{ij1}\hat{q}_{ij2}\hat{p}_{ij3})\Big) \\ &= \log\lambda_{ij} + \text{offset} \end{aligned} \tag{3.19}$$

In practice, fish are often patchily distributed even after taking into consideration habitat characteristics and therefore there is likely to be overdispersion in the densities. Placing a between-event gamma distribution on density allows for over-dispersion, and results in a negative binomial likelihood in place of the Poisson likelihood. The negative binomial, like the Poisson, is an exponential family distribution and therefore fits within the scope of generalised linear and additive modelling.

3.3.5.1 Propagating error between stages

A problem with modelling density conditional on capture probability is that the error in the estimation of $\hat{p}_{ijk}$ is not carried forward into the density model. In hierarchical models the error is automatically transferred in the fitting process, but as discussed previously these can be cumbersome to specify and fit. The proposal here is to use a

parametric boostrap procedure to transfer the error in the estimation of $\hat{p}_{ijk}$ through to the model for density λ_{ij}. This is achieved as follows:

1. a model selection procedure is undertaken to estimate a suitable capture probability model
2. conditional on the estimates of capture probability, a model selection procedure is undertaken to estimate a suitable density model
3. a large number (1000, say) simulations of possible capture probability values are made by taking draws from a multivariate normal distribution with mean given by the parameter estimates of the capture probability model, and variance matrix by the inverse of the negative Hessian matrix of the solution surface at the parameter estimates
4. the final density model is refitted using each simulated set of capture probabilities in turn and the model parameter estimates retained
5. appropriate confidence intervals can then be constructed for effects and predictions from the density model that account for the uncertainty in the estimates of capture probability.

3.4 Implementation in R

This section gives an overview of the modelling software developed, such as the `ef` package for the estimation of capture probability from removal data. I show how the functions can be used and give some simple examples.

An R package was developed to combine and extend existing functions in R, particularly from the packages mgcv to allow capture probability and density to be

modelled using the GMRF models described in Chapter 2. The package will fit the model described in this chapter to any removal method data, it is not limited to electrofishing.

In the case of capture probability, the model building functionality of mgcv (Wood, 2011) was used to generate a design matrix to include covariates in the likelihood (3.12) to allow rapid specification of complicated models. However, this method does not currently allow the use of generalised cross-validation to determine the degree of smoothing and as such it is necessary to use fixed degrees of freedom splines.

The following is to serve as an introduction to the practical use of the `ef` package. Data should be prepared in a data-frame with covariates stored in columns and lines representing site visits. The most basic sample is one in which there are 3 passes and 2 lifestages; this constitutes a single sample.

```
# create a single electrofishing site visit with 3 passes and 2 lifestages
ef_data <- data.frame(n      = c(100, 53, 24, 50, 26, 12),
                      pass   = c(  1,  2,  3,  1,  2,  3),
                      stage  = c(  1,  1,  1,  2,  2,  2),
                      sample = c(  1,  1,  1,  2,  2,  2))
ef_data
#     n pass stage sample
# 1 100    1     1      1
# 2  53    2     1      1
# 3  24    3     1      1
# 4  50    1     2      2
# 5  26    2     2      2
# 6  12    3     2      2
```

A sensible model to fit to this is one where the capture probability is constant across passes but varies by life-stage. This can be achieved by

```
# Fit a simple model
m1 <- efp(n ~ 1 + factor(stage), data = ef_data, pass = pass)
m1
#
# Call:  efp(formula = n ~ 1 + factor(stage), data = ef_data, pass = pass)
```

```
#
# Coefficients:
#    (Intercept)  factor(stage)2
#        0.00306         0.00920
#
# Degrees of Freedom: 6 Total (i.e. Null);  4 Residual
# Null Deviance:     NA
# Residual Deviance: NA  AIC: 510
```

Notice that it is necessary to supply the information on which electrofishing pass via the argument 'pass'.

Currently the object returned by the function `efp` is a class in its own right, to distinguish it, but it inherits from a glm

```
class(m1)
# [1] "efp" "glm" "lm"
```

So as long as I create the appropriate elements within the the efp object, such as `loglik` and `coeff`, then the printing methods, `logLik`, `AIC`, `fitted`, `coef`, etc. that are defined for a `glm` object, will all work on the `efp` object.

```
AIC(m1)
# [1] 510
```

Prediction is also possible because the parameters are fitted on the logistic scale, and by ensuring that the `family` element of the fitted model is set to `binomial(logit)` i.e. the model coefficients are on the logistic scale, then predictions with new data are straightforward.

3.4.1 Efficient and flexible optimisation

The `ef` package uses an efficient optimiser which employs the BFGS algorithm (an approximation to Newton's method that doesn't require the explicit calculation of the matrix of second derivatives, named after its discoverers Broyden, Fletcher, Goldfarb

and Shanno, but affectionately known as Big Friendly Giant Steps) with automatic differentiation carried out by tools in the (RSTAN, Stan Development Team, 2014a) package. The model code in the STAN language is

```
stanmod <- rstan::stan_model(model_code = "
    data {
      int<lower=0> N; // number of observations
      int<lower=0> K; // number of parameters
      row_vector[N] y[3]; // data - 3 passes
      vector[N] offset[3]; // offset - 3 passes
      matrix[N,K] A[3]; // the design matrices - 3 passes
    }
    parameters {
      vector[K] alpha;
    }
    model {
      vector[N] p[3]; // calculate all the probs required
      for (s in 1:3) {
        p[s] = 1.0 ./ (1.0 + exp(-1.0 * A[s] * alpha - offset[s]));
      }
      target += y[1] * log(p[1]);
      target += y[2] * log((1-p[1]) .* p[2]);
      target += y[3] * log((1-p[1]) .* (1-p[2]) .* p[3]);
      target += -1.0 * (y[1] + y[2] + y[3]) *
                       log(p[1] +
                            (1-p[1]) .* p[2] +
                              (1-p[1]) .* (1-p[2]) .* p[3]
                           );
    }")
```

which is translated and compiled against automatic differentiation libraries to produce a dynamic library that performs the optimisation and returns the maximum likelihood estimates and an estimate of the matrix of second derivatives. The optimiser is coded to take as input a design matrix for each pass `A[3]` and data `y[3]`, along with some data summaries to define the sizes of objects. This design allows a large amount of flexibility as the design matrix is specified in R before being passed to the optimiser. This allows the use of mgcv and extensions developed in this thesis to be used in model specification, combining flexibility with fast, accurate, and stable optimisation.

See the file https://github.com/Faskally/ef/blob/master/R/zzz.R for more advanced versions of the above code which can fit to data with variable numbers of passes and also incorporate a fixed quadratic penalty.

3.4.2 Simple fits to real data

The following sections will fit simple models to the data used later in this chapter. The data is stored in an object called `ef` and has the following structure

```
str(ef)
# Classes 'tbl_df' and 'data.frame': 17022 obs. of  58 variables:
#  $ totlanduse   : num  1964 1792 1758 1758 1758 ...
#  $ keep         : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
#  $ sampleID     : num  2512 2673 2877 2942 3017 ...
#  $ pass         : num  1 1 1 1 1 1 1 1 1 1 ...
#  $ n            : num  0 0 38 38 28 23 16 16 0 0 ...
#  $ pass23       : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
#  $ Site_OBJECTID: int  2079 2918 2022 2022 2022 2022 2022 2022 2022 2022 ...
#  $ Site.Name    : chr  "ACHRIDIGILL" "Adielinn Burn 1" "ALLT A MHARCAIDH LOWER" "ALLT A MHARC
#  $ Dataset      : chr  "fobs" "fobs" "fobs" "fobs" ...
#  $ Width        : num  NA NA NA NA NA NA NA NA NA NA ...
#  $ Date         : chr  "20/08/2002" "18/09/2013" "03/09/1997" "02/09/1998" ...
#  $ Runs         : num  3 3 3 3 3 3 3 3 3 3 ...
#  $ Area         : num  73.5 40.7 180.2 180.2 180.2 ...
#  $ Trust        : chr  "MSS" "MSS" "MSS" "MSS" ...
#  $ Species      : chr  "Salmon" "Salmon" "Salmon" "Salmon" ...
#  $ LifeStage    : chr  "Fry" "Fry" "Fry" "Fry" ...
#  $ coords.x1    : num  288767 335398 285517 285517 285517 ...
#  $ coords.x2    : num  962796 771511 803628 803628 803628 ...
#  $ NEAR_X       : num  288767 335398 285517 285517 285517 ...
#  $ NEAR_Y       : num  962796 771511 803628 803628 803628 ...
#  $ Elevation_   : num  25 234 251 251 251 251 251 251 251 251 ...
#  $ Slope_deg    : num  9.07 2.84 2.43 2.43 2.43 ...
#  $ Upcatch_km   : num  2.18 5.07 20.93 20.93 20.93 ...
#  $ Water_A      : num  0 172 205 205 205 ...
#  $ Water_W      : num  0 3.12 4.31 4.31 4.31 ...
#  $ Distance_s   : num  1.08 53.99 110.8 110.8 110.8 ...
#  $ Distance_1   : num  1.08 53.99 110.8 110.8 110.8 ...
#  $ Urban        : num  0 0 0.0159 0.0159 0.0159 ...
```

```
#  $ CTrees       : num  0 0.994 0 0 0 ...
#  $ NCTrees      : num  0 0 0.113 0.113 0.113 ...
#  $ Mixed        : num  0 0 0 0 0 0 0 0 0 0 ...
#  $ Marsh        : num  0 0 0 0 0 0 0 0 0 0 ...
#  $ Other        : num  1 0.0057 0.8707 0.8707 0.8707 ...
#  $ CATCH_       : num  135 41 28 28 28 28 28 28 28 28 ...
#  $ RIVCODE      : chr  "135/0.0000" "41/0.0000" "28/70.0000" "28/70.0000" ...
#  $ OBJECTID     : int  2079 2918 2022 2022 2022 2022 2022 2022 2022 2022 ...
#  $ rivs_OBJECTID: int  250711 46478 55524 55524 55524 55524 55524 55524 55524 55524 ...
#  $ CATCH_ID     : chr  "592" "441" "492" "492" ...
#  $ DESCRIPTIO   : chr  "Halladale River at mouth" "River South Esk at mouth" "River Spey at m
#  $ barrier      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
#  $ OBJECTID.1   : int  39 12 7 7 7 7 7 7 7 7 ...
#  $ HACode       : int  96 13 8 8 8 8 8 8 8 8 ...
#  $ HAName       : chr  "Naver group" "Esk Group" "Spey" "Spey" ...
#  $ SHAPE_AREA   : num  0 0 0 0 0 0 0 0 0 0 ...
#  $ SHAPE_LEN    : num  0 0 0 0 0 0 0 0 0 0 ...
#  $ hmidx        : int  15 1 5 5 5 5 5 5 5 5 ...
#  $ optional     : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
#  $ s            : num  3 3 3 3 3 3 3 3 3 3 ...
#  $ T            : num  0 0 57 48 45 34 26 27 0 0 ...
#  $ X            : num  0 0 90 80 70 51 39 39 0 0 ...
#  $ Z            : num  0 0 1.58 1.67 1.56 ...
#  $ phi          : num  0 0 0.211 0.167 0.222 ...
#  $ year         : num  2002 2013 1997 1998 1999 ...
#  $ doy          : int  231 260 245 244 248 251 255 251 247 244 ...
#  $ fyear        : chr  "2002" "2013" "1997" "1998" ...
#  $ sinSlope     : num  0.1577 0.0496 0.0424 0.0424 0.0424 ...
#  $ logGradient  : num  -1.83 -3 -3.16 -3.16 -3.16 ...
#  $ woodland     : num  0 0.994 0.113 0.113 0.113 ...
```

Fitting factors, linear effects and smoothers

Factors (e.g., organisation or hydrometric area), linear effects (e.g., width) and smoothers (e.g., day of year) can be specified for the capture probability model as follows

```
  m1 <- efp(n ~ s(doy, k = 4) + Trust,
            data = ef, pass = pass)
  m1
#
```

```
# Call:  efp(formula = n ~ s(doy, k = 4) + Trust, data = ef, pass = pass)
#
# Coefficients:
#             (Intercept)              TrustArgyll                 TrustAyr
#                  0.3723                  -0.1860                  -0.0705
#              TrustClyde               TrustConon                 TrustDee
#                  -0.0612                  -0.2498                  -0.1653
#            TrustDeveron                TrustEsk            TrustFindhorn
#                  -0.6668                  -0.3000                  -0.0714
#              TrustForth            TrustGalloway  TrustKyle of Sutherland
#                  -0.3105                   0.2733                  -0.6980
#           TrustLochaber                TrustMSS               TrustNaver
#                  -0.3987                   0.2224                  -0.7416
#    TrustNess and Beauly               TrustOther      TrustOuter Hebrides
#                  -0.2460                   0.1096                  -0.4083
#               TrustSEPA               TrustSpey                 TrustTay
#                  -0.3737                  -0.1879                  -0.0418
#              TrustTweed     TrustWest Sutherland         TrustWester Ross
#                  -0.3621                  -0.4342                  -0.4379
#
#                  -0.1070                   0.4987                   0.0892
#
# Degrees of Freedom: 17022 Total (i.e. Null);  16995 Residual
# Null Deviance:     NA
# Residual Deviance: NA  AIC: 519000
```

where DoY is restricted to a smoother with 3 degrees of freedom (`k = 4`).

If we wanted to go forward and model fish density we would first have to compute the total fish captured and the probability $\hat{p}_{ij1} + \hat{q}_{ij1}\hat{p}_{ij2} + \hat{q}_{ij1}\hat{q}_{ij2}\hat{p}_{ij3}$, which is equivalent to $1 - \hat{q}_{ij1}\hat{q}_{ij2}\hat{q}_{ij3}$. This is done by the following functions

```
getp <- function(x, data) {
  data $ pijk <- x
  tmp <- data %>%
         select(sampleID, LifeStage, pass, pijk) %>%
         as.data.frame(.) %>%
         spread(pass, pijk)

  1 - (1-tmp[["1"]]) * (1-tmp[["2"]]) * (1-tmp[["3"]])
}
getn <- function(data) {
```

```
  c(tapply(data $ n, list(data $ LifeStage, data $ sampleID), sum))
}
getSummary <- function(model, data) {
  tmp <- data %>%
         select(sampleID, LifeStage, pass, n) %>%
         as.data.frame(.) %>%
         mutate(pass = paste0("n_", pass)) %>%
         spread(pass, n)
  within(tmp, {
      y = getn(data)
      p = getp(model$fitted, data)
    })
}
```

and the application of these functions results in a new data frame with a row for each lifestage and sample, to which we add on selected covariates from the original data set.

```
efdens <- getSummary(m1, ef)
efdens <- left_join(efdens,
                    unique(
                      ef[c("sampleID", "LifeStage", "Area",
                           "Trust", "fyear", "HACode", "Water_W", "doy")
                         ]))
efdens$offset <- log(efdens$Area * efdens$p)
str(efdens)
# 'data.frame': 5674 obs. of  14 variables:
#  $ sampleID : num  1 1 2 2 3 3 4 4 5 5 ...
#  $ LifeStage: chr  "Fry" "Parr" "Fry" "Parr" ...
#  $ n_1      : num  0 2 12 8 33 6 10 12 22 3 ...
#  $ n_2      : num  0 0 13 7 28 2 3 1 4 2 ...
#  $ n_3      : num  0 1 5 1 4 2 1 1 0 1 ...
#  $ p        : num  0.941 0.941 0.941 0.941 0.941 ...
#  $ y        : num  0 3 30 16 65 10 14 14 26 6 ...
#  $ Area     : num  168 168 118 118 217 ...
#  $ Trust    : chr  "Annan" "Annan" "Annan" "Annan" ...
#  $ fyear    : chr  "2000" "2000" "2000" "2000" ...
#  $ HACode   : int  78 78 78 78 78 78 78 78 78 78 ...
#  $ Water_W  : num  10.37 10.37 7.81 7.81 7.59 ...
#  $ doy      : int  228 228 228 228 229 229 230 230 208 208 ...
#  $ offset   : num  5.06 5.06 4.71 4.71 5.32 ...
```

Now we are in a position to model fish density in terms of some covariates using mgcv. In these examples I assume a negative binomial distribution (e.g., see `?mgcv::nb`).

```
gam(y ~ factor(HACode) + Water_W + s(doy, k=6),
    offset = offset, data = efdens, family = nb)
#
# Family: Negative Binomial(0.585)
# Link function: log
#
# Formula:
# y ~ factor(HACode) + Water_W + s(doy, k = 6)
#
# Estimated degrees of freedom:
# 3.8  total = 48.8
#
# REML score: 26490
```

where `HACode` is modelled as a factor, `Water_W` as a linear term and `doy` as a thin plate spline. In this case, because the flexibility of the smoother is estimated from the data, a larger (maximum) degrees of freedom can be safely specified. Other standard GAM functionality remains. A range of 'smooth types' are possible e.g., thin plate splines with shrinkage (`bs='ts'`), random effects (`bs='re'`) or GMRF smoothers (`bs='gmrf'`), see for example,

```
gam(y ~ s(HACode, bs='re')  + Water_W + s(doy, k=6),
    offset = offset, data = efdens, family = nb)
#
# Family: Negative Binomial(0.532)
# Link function: log
#
# Formula:
# y ~ s(HACode, bs = "re") + Water_W + s(doy, k = 6)
#
# Estimated degrees of freedom:
# 0.968 3.569  total = 6.54
#
# REML score: 26809
```

where this time `HACode` is modelled as a random effect. Note the reduction in effective degrees of freedom compared to the previous fit.

Fitting regional smoothers

Regional covariates such as `HACode` could be estimated independently, that is, as a factor, but given likely spatial correlation it is more appropriate to fit a model where neighbouring regions have similar effects. This requires specification of the spatial (neighbourhood) structure using `poly2nb` and `nb2mat` functions in the spdep package. The shape file should be read into R using `readOGR` from the rgdal package and stored in an object (in this case '`hma`') where the neighbourhood connections are calculated and converted to an adjacency matrix before conversion to GMRF. The shapefile describing the outlines of the hydrometric areas is contained in a data package called `CLdata` and can be installed using

```
devtools::install_github("colinpmillar/CLdata")
```

The neighbourhood structure is then identified by the following (fairly involved) code.

```
hma <- hma[!hma $ HAName %in% c("Orkneys","Shetlands"),]
hmaadj <- spdep::poly2nb(hma, queen = FALSE)
hmaadj <- spdep::nb2mat(hmaadj, style = "B", zero.policy = TRUE)
# add connections for inner and outer hebs
hmaadj[hma $ HAName == "Inner Hebrides", hma $ HAName == "Outer Hebrides"] <- 1
hmaadj[hma $ HAName == "Outer Hebrides", hma $ HAName == "Inner Hebrides"] <- 1
hmaadj[hma $ HAName == "Inner Hebrides", c(21, 42, 43)] <- 1
hmaadj[c(21, 42, 43), hma $ HAName == "Inner Hebrides"] <- 1
Qhma <- methods::as(hmaadj, "dgTMatrix")

Qhma @ x[] <- -1/Qhma @ x
diag(Qhma) <- rowSums(hmaadj)
Qhma <- as.matrix(Qhma)
colnames(Qhma) <- rownames(Qhma) <- hma $ HACode
```

First the polygons that neighbour each other are identified, then these are converted to a matrix. For modelling purposes, the Shetland and Orkney are first removed, and the Inner and Outer Hebrides are artificially joined to the mainland. Finally a penalty

matrix is created, the first 7 columns and rows of which are

$$\boldsymbol{Q}_{hma} = \begin{pmatrix} 3 & & & & & -1 & \\ & 2 & & & & & \\ & & 4 & & & -1 & -1 \\ & & & \ddots & & \ddots & \\ & & & & 7 & -1 & \\ -1 & & -1 & -1 & -1 & 10 & \\ & & -1 & & & & 6 \end{pmatrix} \tag{3.20}$$

The regional effect for `HACode` can be specified using the GMRF basis, `bs='gmrf'` to estimate spatial variability in capture probability, note that a reduced rank version is being used here (`k=12`). The construction of the reduced rank GMRF is being done behind the scenes by the mgcv smooth construct functions, which are then fed into the capture probability optimiser.

```
efp(n ~ s(HACode, k = 12, bs = 'gmrf', xt = list(penalty = Qhma)),
          data = ef, pass = pass)
#
# Call:  efp(formula = n ~ s(HACode, k = 12, bs = "gmrf", xt = list(penalty = Qhma)),
#     data = ef, pass = pass)
#
# Coefficients:
# (Intercept)
#      0.1326       0.0175      -0.0218      -0.0477       0.0746
#
#     -0.0595      -0.0286      -0.0514       0.1108       0.1047
#
#     -0.1436      -0.0478
#
# Degrees of Freedom: 17022 Total (i.e. Null);  17010 Residual
# Null Deviance:     NA
# Residual Deviance: NA  AIC: 521000
```

Although it is possible to incorporate spatial smoothers using existing functionality in mgcv, the 'gmrf' extension provides a more flexible interface that can be used to fit

regional effects. It can also automatically accommodate regions without observations but where there is information on spatial structure. By default the maximum degrees of freedom for regional smoothers would be equal to the number of regions. However, this can be restricted by specifying a lower value, such as, `k=12`, which results in a reduced rank GMRF being fitted.

3.5 Modelling capture probability

3.5.1 Modelling options

The conditional likelihood for capture probability is not an exponential family distribution and therefore does not allow the use of penalised regression components through the use of standard packages. However, nonlinear responses can be modelled through the use of fixed degrees of freedom splines as presented in the previous section. A nice way to specify a design matrix for a wide range of splines is through the mgcv interface allowing models to be specified using widely recognised syntax, with the added advantage that identifiability checks are carried out by mcgv.

Continuous variables such as day of year and gradient can be specified as reduced rank forms of the continuous random walk models described in Section 2.2.1.5, however, since there are so many possibilities for constructing smoothing splines in mgcv, a simple and approximately equivalent choice is one of the standard smoothers (for example, reduced rank thin plate regression splines, Wood (2003)).

It is possible to model Hydrometric area (HA) as a spatial effect using a reduced rank GMRF. This is done by setting up a penalty matrix seen in the previous section, based on a difference matrix that defines the connections between the hydrometric areas. In this example, the Outer Hebrides were linked to the Inner Hebrides, which

were in turn linked to adjacent mainland areas. The reduced rank version of the full rank regional smoother was then constructed resulting in a set of basis functions which encode the spatial dependence between the regions. These basis functions specified a spatial spline regional model for HA.

When working with fixed degrees of freedom smoothers there are several options for specifying how many degrees of freedom to use. One way would be to attempt to fill all possible combinations of degrees of freedom and select the best using a criteria such as AIC. However, even with a modest collection of covariates this approach is impractical due to the huge number of models that would be required to fit. A pragmatic alternative is to select the degrees of freedom to allow for particular nonlinear forms. For example, with continuous covariates, 2 degrees of freedom would allow a simple modal response, where 3 degrees of freedom may allow for modal response with asymmetry. The number of degrees of freedom for a regional smoother is less obvious and preliminary analysis can help with this choice.

3.5.2 Model fitting

Using the likelihood defined in Section 3.3.3 and the tools is Section 3.4 I first fitted a 'large' model to capture most of the systematic variation in the data. The reason for this is to investigate the potential for overdispersion in the data. The large model can

be written symbolically as:

$$\begin{aligned}\text{logit } p \sim \quad & \text{life-stage + pass + life-stage:pass + organisation + organisation:pass} \\ & \text{+year + organisation:year + hydrometric area + s(altitude)} \\ & \text{+s(upstream catchment area) + s(distance to sea)} \\ & \text{+s(gradient) + s(channel width) + s(urban) + s(woodland)} \\ & \text{+s(marsh) + s(other) + s(day of year)} \end{aligned} \tag{3.21}$$

and is fitted in R, and the fitted probabilities added to the ef dataset using

```
big <- c("LifeStage",
         "Trust",
         "fyear",
         "pass23",
         "Trust:pass23",
         "LifeStage:pass23",
         "Trust:fyear",
         "s(HACode, k = 12, bs = 'gmrf', xt = list(penalty = Qhma))",
         "s(Water_W, k = 3)",
         "s(Elevation_, k = 3)",
         "s(Distance_s, k = 3)",
         "s(sinSlope, k = 3)",
         "s(Upcatch_km, k = 3)",
         "s(Urban, k = 3)",
         "s(woodland, k = 3)",
         "s(Marsh, k = 3)",
         "s(Other, k = 3)",
         "s(doy, k = 3)"
         )
bigf <- formula(paste("n ~", paste(big, collapse = " + ")))
bigmod <- efp(bigf, data = ef, pass = pass, verbose = TRUE, hessian = FALSE)
ef $ bigfit <- bigmod $ fitted
```

In this model, pass was considered as a categorical variable with two levels, for the first two passes, and the constraint that the capture probabilities for the second and third pass are equal (i.e. $p_{ij2} = p_{ij3}$). This avoided problems with model identifiability which arose because there was little information in the data to differentiate the

capture probabilities in the second and third passes. Interactions, denoted by ':', were included to allow for different changes in capture probability between passes for each life-stage and organisation, and for changes in staff or protocol over time within organisation. Hydrometric area was fitted as a categorical variable with 44 levels, but with spatial structure imposed so that neighbouring regions were correlated. This was achieved using a reduced rank Gaussian Markov random field spatial smoother (Rue and Held, 2005; Wood, 2003) with 12 basis functions (11 degrees of freedom) providing a reasonable compromise between complexity and model fit. Continuous variables were fitted as reduced rank thin plate regression splines (Wood, 2003), denoted by s(.), with three basis functions (two degrees of freedom) allowing, at most, a modal form.

A forwards and backwards stepwise selection procedure was then used to refine the model. At each stage I considered dropping interaction terms involving the categorical variables, replacing the smooth functions of the continuous variables with linear effects, and dropping the main effects of the categorical and continuous variables (provided they weren't involved in any interactions or expressed as smooth functions). I also considered introducing interactions between the continuous variables and life-stage and pass, to allow for different relationships for fry and parr and for the first- and second-pass capture probabilities. Model selection was based on an adjusted version of the Bayesian Information Criterion:

$$\mathrm{BIC}_{\mathrm{adj}} = -\frac{2L}{\Psi_{\mathrm{between}}} + \log(N)(M+1) \tag{3.22}$$

where L is the log-likelihood, N is the number of samples (2,837), M is the number of linear terms in the model (see equation 3.13), and Ψ_{between} is an estimate of overdispersion (see equation 3.24) based on the fit of the large model (3.21). The large number of samples makes $\mathrm{BIC}_{\mathrm{adj}}$ a strict selection criterion that is likely to

be successful at both including relevant and excluding irrelevant terms (Murtaugh, 2009; Raffalovich et al., 2008).

Comparing the log-likelihood of the large model (3.21) with the log-likelihood of a saturated model, in which there is a separate first- and second-pass capture probability for each sample and life-stage, calculated by

```
# fit samplewise saturated model
n <- nrow(ef) # data points
N <- n / 6    # site visits (samples)
llsat <- rep(NA, N)
sIDs <- sort(unique(ef $ sampleID))
for (i in 1:N) {
  samp <- subset(ef, sampleID == sIDs[i])
  mod <- efp(n ~ factor(pass) * LifeStage, data = samp, pass = pass)
  llsat[i] <- logLik(mod)
}
```

showed that the data were over-dispersed. That is, there was more variation in the data than expected given the multinomial likelihood. Residual plots indicated that this was not due to any systematic lack of fit. The overdispersion was partitioned into within- and between-sample overdispersion by also computing the log-likelihood of a 'sample-wise' model in which there was an extra parameter for each sample (i.e. logit $p \sim$ sample + terms in model 3.21). Due to the number of parameters, this model was fitted by conditioning on the fitted values from the large model and estimating the sample effect for each sample in turn as follows:

```
# now estimate a sample effect for each sample
n <- nrow(ef) # data points
N <- n / 6    # site visits (samples)
llsample <- rep(NA, N)
sIDs <- sort(unique(ef $ sampleID))
for (i in 1:N) {
  samp <- subset(ef, sampleID == sIDs[i])
  offset <- logit(samp $ bigfit)
  mod <- efp(n ~ 1, offset = offset, data = samp, pass = pass)
  llsample[i] <- logLik(mod)
}
```

The within- and between-sample overdispersions were then estimated to be:

```
# compare deviance
N <- nrow(ef) / 6    # 3 passes, 2 life-stages
M_large <- bigmod $ df.null - bigmod $ df.residual
## within sample overdispersion
(psi_within <- 2 * (sum(llsat) - sum(llsample)) / (3*N - M_large))
# [1] 1.39
## between sample overdispersion
(psi_between <- as.numeric(2 * (sum(llsample) - logLik(bigmod)) / N))
# [1] 2.98
```

that is,

$$\Psi_{\text{within}} = \frac{2(L_{\text{saturated}} - L_{\text{sample}})}{2N - M_{\text{large}} - 1} = 1.389 \tag{3.23}$$

$$\Psi_{\text{between}} = \frac{2(L_{\text{sample}} - L_{\text{large}})}{N} = 2.982 \tag{3.24}$$

where $L_{\text{saturated}}$, L_{sample}, L_{large} are the log-likelihoods of the saturated, sample-wise, and large model respectively, and M_{large} is the number of linear terms in the large model. Since Ψ_{between} was the dominant term, and since our focus is to produce a model that can be used to predict capture probabilities at the national level (i.e. across samples), I used Ψ_{between} to adjust the BIC. This should prevent any over-fitting of the variables that change between samples (such as the habitat variables), albeit at the loss of some power for detecting within-sample effects (those involving pass or life-stage).

The importance of the explanatory variables in the final model was assessed by calculating the change in BIC_{adj} and by an F-test based on the change in log-likelihood, when each variable was removed in turn. The F-test used an updated estimate of Ψ_{between} from the final model.

The adequacy of parametric inference based on the final model (such as the F-tests above) was assessed by comparing parametric estimates of the parameter standard errors (from the inverse of the Hessian matrix scaled by Ψ_{between} estimated from the final model) with bootstrap standard errors (Efron and Tibshirani, 1994) obtained by re-sampling the electrofishing samples with replacement and re-estimating the model parameters 2000 times. This showed that the parametric standard errors tended to be too small for the between-sample effects and too large for the within-sample effects (by factors of about 0.85 and 1.18 respectively). Hence, the corresponding F-tests were likely to be 'too significant' and 'not significant enough' respectively. The F-tests for the between-sample effects were therefore supplemented by permutation tests in which the values of the explanatory variable were permuted across electrofishing samples and the F-statistics recalculated, thus generating a reference distribution under the null hypothesis of no effect.

3.5.3 Results

The final capture probability model was:

$$\begin{aligned}\text{logit } p \sim \quad & \text{life-stage + pass + life-stage:pass + organisation} \\ & \text{+year + altitude + altitude:life-stage + distance to sea} \\ & \text{+channel width + s(day of year) + s(day of year):life-stage}\end{aligned} \tag{3.25}$$

and was fitted as follows, by first centering the continuous covariates to reduce correlation in the solution surface.

```
ef <- within(ef, {
              cDistance_s = c(scale(Distance_s))
              cWater_W = c(scale(Water_W))
              cElevation_ = c(scale(Elevation_))
```

```
              fyear = factor(fyear)
              Trust = factor(Trust)
              LifeStage = factor(LifeStage)
              })
contrasts(ef $ fyear) <- "contr.sum"
contrasts(ef $ Trust) <- "contr.treatment"
```

The the model formula was the specified and fitted

```
finalf <-  n ~ LifeStage + Trust + fyear + pass23 + cWater_W +
                  cElevation_ + cDistance_s +
                  LifeStage:pass23 + s(doy, k = 3, by = LifeStage) +
                  cElevation_:LifeStage

finalm <- efp(finalf, data = ef, pass = pass)
```

The fit of all the models considered in the model selection process is given in the supplementary material. Organisation (a composite measure of people, equipment and protocols) had the greatest effect on capture probability, both in terms of the change in BIC_{adj} (Table 3.1) and effect size (Fig. 3.5a). Life-stage was the next most important (Table 3.1), with parr being more catchable than fry, and the magnitude of the effect greater at higher altitudes and in the earlier or later parts of the year (Fig. 3.5b). Pass was also significant, as was the interaction between life-stage and pass (Table 3.1). The odds of being caught on the second or third pass was 0.91 or 0.81 times the odds of being caught on the first pass for fry and parr, respectively. How this translates into capture probabilities depends on the values of the other explanatory variables. But, to illustrate, for organisation MSS, year 2006 and assuming all the other variables take their median value, the capture probabilities in the first and subsequent passes are estimated respectively to be 0.67 and 0.65 for fry and 0.72 and 0.68 for parr. Capture probability showed a modal response to day of year for both fry and parr, with the response more pronounced for fry (Fig. 3.5c). Capture probability also varied with year, but showed no systematic change over time (Fig. 3.5d). The remaining continuous variables were all characterised by linear effects: capture probability declined with increasing distance to sea, increased with altitude

(with a steeper gradient for parr) and decreased with channel width (Figs. 3.5e-g). Hydrometric area and the four landuse variables were absent from the final model.

Figure 3.6 shows the spatial distribution of modelled capture probabilities (Fig. 3.6a) and the partial effects of organisation (Fig. 3.6b) and the habitat variables distance to sea (Fig. 3.6c), altitude (Fig. 3.6c) and channel width (Fig. 3.6f). Large scale spatial variability was dominated by the effect of organisation (Fig. 3.6b). Organisations in the south-west tend to have higher capture probabilities than those in the north-west. However, there is no pattern evident across organisations on the east coast. The partial effects of distance to sea and altitude are difficult to interpret from Figures 3.5e and f, because the two variables are moderately correlated (0.73). However, their combined effect allows for increased capture probability in higher altitude sites near the coast and decreased capture probability in lower altitude catchments distant from the coast (Fig. 3.6e). The effect of channel width was relatively small and shows no broad spatial trend (Fig. 3.6f).

3.6 Modelling fish density

3.6.1 Modelling considerations

Conditional on an estimate of capture probability for each electrofishing event, the total fish caught on each event is modelled as negative binomial (Section 3.3). This is a pragmatic choice to make allowances for overdispersion in the counts. Because the likelihood for this model can be fitted in the generalised additive modelling framework, it can also incorporate structural additive components and hence is a STAR model. This allows a wide range of options for modelling covariates.

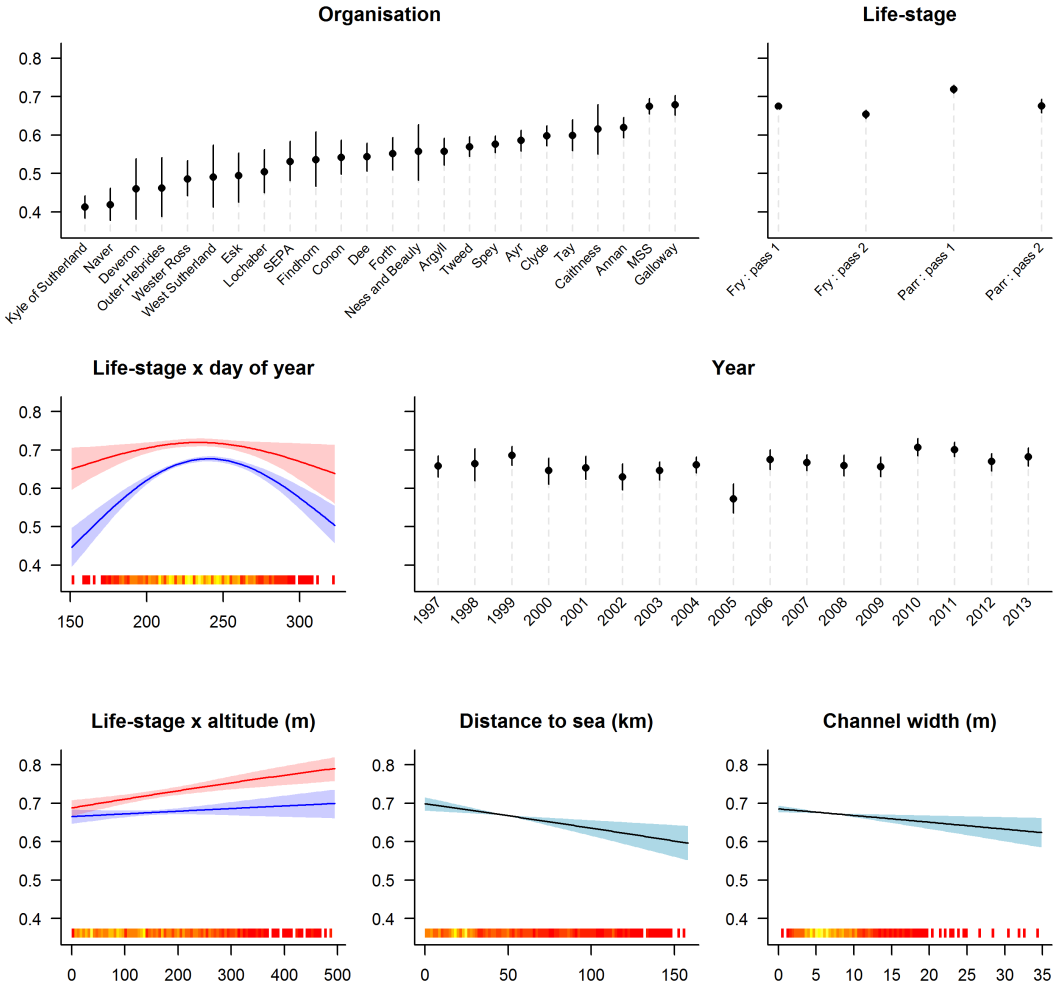


Fig. 3.5 The effect of the explanatory variables on capture probability, standardised to organisation MSS, life-stage fry, pass 1, year 2006, and the median values of the remaining variables. Organisation names have been abbreviated. For the continuous explanatory variables, the line colour indicates life-stage (blue: fry, red: parr, black: both). Bootstrap 95% pointwise confidence intervals are shown as shaded blue or red areas or vertical bars. A 'rug' indicates the distribution of available data on the x-axis (red: few values, yellow: many values).

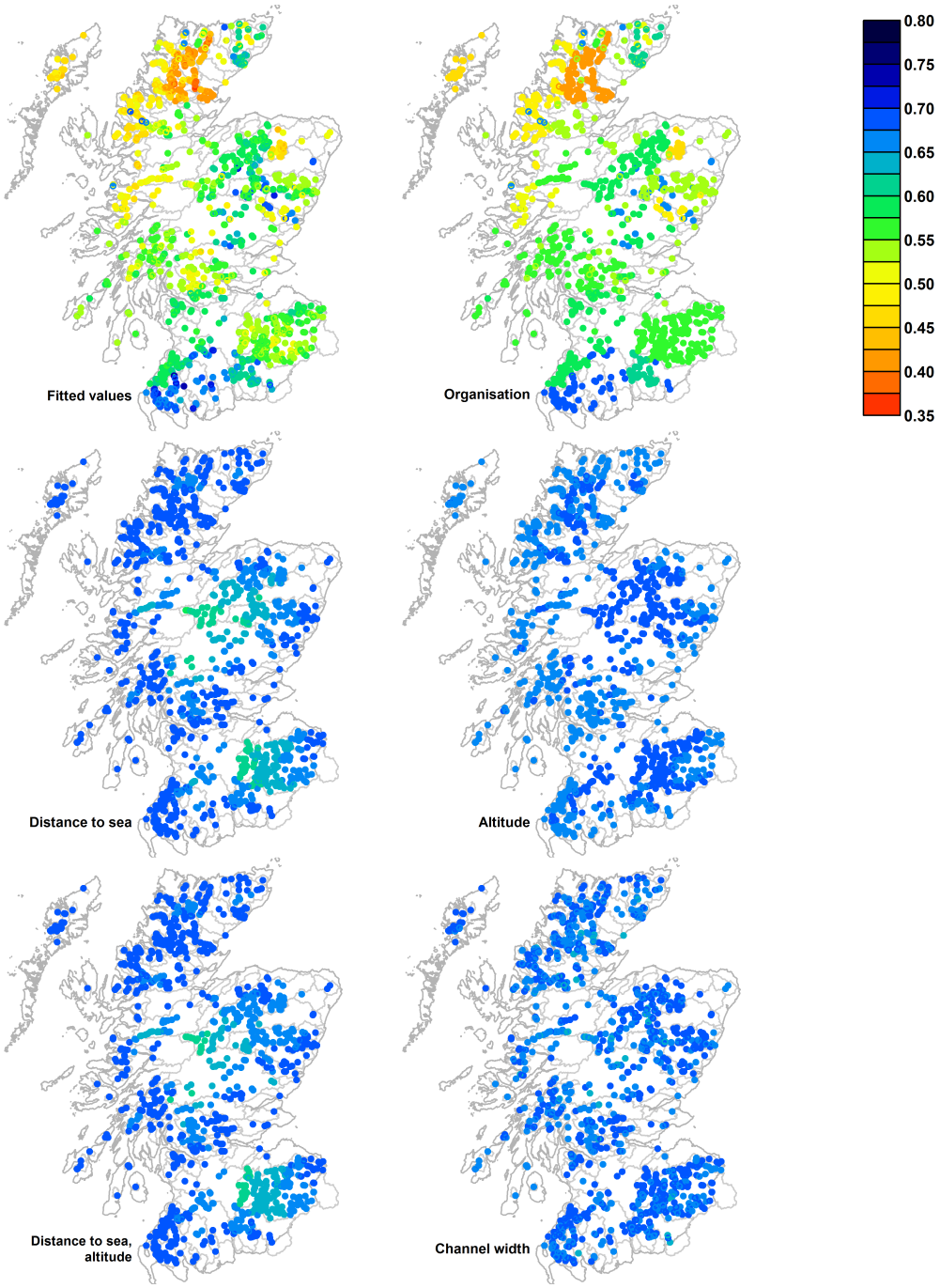


Fig. 3.6 The spatial distribution of modelled capture probabilities (Fitted values, a), the partial effects of organisation (b), distance to sea (c), altitude (d), and channel width (f), and the combined effect of distance to sea and altitude (e). The fitted values and the partial effects are all conditioned on life-stage fry, pass 1, year 2006, and the median day of year (233). The partial effects are also conditioned on the median values of the remaining variables and, for distance to sea, altitude and channel width, on organisation MSS. Organisation has a larger range of partial effect sizes than the other explanatory variables and so panels a and b are coloured using the upper scale, while panels c, d, e and f are coloured using the lower scale.

Table 3.1 The relative importance of each explanatory variable and each interaction term in the final capture probability model as indicated by changes in BIC_{adj} and p-values (based on an F-test) associated with removing individual terms from the final model. The values for life-stage, pass, day of year and altitude measure the overall effect of these variables (i.e. the main effect and the interactions). The p-values are conditioned on a between-sample overdispersion $\Psi_{between}$ of 3.59 estimated from the final model. The degrees of freedom (d.f.) of each term are also given. The p values for the between-sample effects are likely to be 'too significant', but permutation tests confirmed a significance of < 0.001 in each case.

Term	d.f.	ΔBIC_{adj}	p-value
organisation	23	551.6	< 0.0001
life-stage	5	200.4	< 0.0001
s(day of year)	4	117.6	< 0.0001
year	16	83.3	< 0.0001
pass	2	54.0	< 0.0001
distance to sea	1	18.9	< 0.0001
altitude	2	12.0	< 0.0001
channel width	1	7.4	0.0004
altitude:life-stage	1	7.0	0.0004
s(day of year):life-stage	2	4.4	0.0002
life-stage:pass	1	1.8	0.0045

As in the previous section (3.5) there are a wide choice of univariate smoothers to choose from including reduced rank GMRF smoothers and, in this application, thin plate regression splines were used. Preliminary analysis showed that there was a propensity for smoothers to over-fit to the data producing very wiggly relationships, and so the pragmatic decision was taken to restrict the smoothers to a maximum 3 degrees of freedom in order to avoid overfitting.

Catchment was fitted as a random effect. Hydrometric area was modelled as a regional smoother with a maximum of 24 degrees of freedom, chosen as a compromise between model complexity and speed of model fitting.

Rather than perform a model selection procedure, automatic model selection was achieved through the use of null space penalization (Marra and Wood, 2011). Null space penalization works by constructing an extra penalty for each smooth which penalizes the space of functions of zero wiggliness according to its existing penalties,

that is the linear (normally unpenalized) parameters are lightly penalised, and therefore smooth components can be removed completely. This allowed for a form of automatic function selection (Scheipl et al., 2013b) for hydrometric area. The simplicity of this procedure is appealing: HACode is included as a random effect and as a spatial effect; because a penalty is applied to all model components, the different structural forms for HACode compete and only the ones that best fit the data are retained in the model. The same approach was used for year, where it was included as both a smooth effect and a random effect. Note that it is quite possible that multiple structural forms are retained for a single variable.

Finally, because an offset is used to account for the area fished and the overall capture probability, see (3.19), the covariates are directly modelling density λ. This is made clear by the following

$$\begin{aligned}\log E(\text{counts}) &= \log\Big(\text{density} \times \text{area} \times \text{capture probability}\Big)\\ &\log\Big(\text{density}\Big) + \log\Big(\text{area} \times \text{capture probability}\Big)\\ &\log\Big(\text{density}\Big) + \text{offset}\end{aligned}$$

Hence all covariates and parameters estimated directly model density. For this reason, all models are presented in terms of log fish density.

3.6.2 Model fitting

The first step in model fitting is to create the capture probabilities by sample and life-stage. This is done as follows:

```
efdens <- getSummary(finalm, ef)
efdens <- left_join(efdens,
                    unique(ef[!names(ef) %in% c("pass","n","pass23","bigfit")]))
efdens$offset <- log(efdens$Area * efdens$p)
```

A *large* model was specified based on expert knowledge. The model was chosen so that it was larger than would be expected in order to provide an upper bound for the potential model space. This was specified as follows

```
big <- c("LifeStage",
         "s(year, k = 5, by = LifeStage)",
         "s(fyear, bs = 're')",
         "s(HACode, k = 24, bs = 'gmrf', xt = list(penalty = Qhma))",
         "s(HACode_re, bs = 're')",
         "s(fCATCH_ID, bs = 're')",
         "s(Water_W, k = 3, by = LifeStage)",
         "s(Elevation_, k = 3, by = LifeStage)",
         "s(Distance_s, k = 3, by = LifeStage)",
         "s(sinSlope, k = 3, by = LifeStage)",
         "s(Upcatch_km, k = 3, by = LifeStage)",
         "s(Urban, k = 3, by = LifeStage)",
         "s(woodland, k = 3, by = LifeStage)",
         "s(Marsh, k = 3, by = LifeStage)",
         "s(doy, k = 3, by = LifeStage)"
         )
bigf <- formula(paste("y ~", paste(big, collapse = " + ")))
# create some extra covariates
efdens$HACode_re <- factor(efdens$HACode)
efdens$fHACode <- factor(efdens$HACode)
efdens$Trustfyear <- interaction(efdens$Trust, efdens$fyear)
efdens$fCATCH_ID <- factor(efdens$CATCH_ID)
```

and was fitted directly in mgcv using

```
bigg <- gam(bigf, offset = offset, data = efdens, family = nb,
            select = TRUE)
```

The number of degrees of freedom was restricted to a maximum of 3 for all continuous variables, and each was allowed a different form by life-stage. The HA spatial smoother was restricted to a maximum 24 degrees of freedom (chosen as

a compromise between flexibility and efficiency; preliminary fits showed that 24 degrees of freedom was more than enough to capture the spatial variability). Year, catchment and HA were also included as random effects. Parameter estimation was by minimising the GCV score. A negative binomial distribution was assumed to allow for over-dispersion.

There is only one model specified in this procedure and it corresponds to the upper bound on the explorable model space. The final model however is associated with a restricted model space with much fewer degrees of freedom. In order to simplify the presentation of the fitted penalised model an approximate model is constructed in which smoothers with zero degrees of freedom are removed and smoothers with degrees of freedom less than or equal to one are replaced by linear terms.

3.6.3 Results

A summary of the final model is given in tables 3.2 and 3.3. Based on these summaries the regional smoother of hydrometric area was removed from the model, and several smooth terms were reduced to linear terms (gradient, urban, marsh and day of year). The model was refitted and the simplification procedure repeated resulting in the removal of a linear interaction between gradient and life-stage and marsh, and channel width was reduced to a linear effect. The resulting simplified final model for density is specified in R as follows

Table 3.2 Summary of fixed effect terms in large model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.020	0.098	-10.400	0
LifeStageParr	-1.010	0.032	-31.900	0

Table 3.3 Summary of smooth terms in the large model

	edf	Ref.df	Chi.sq	p-value
s(year):LifeStageFry	3.300	4	77.700	0.0002
s(year):LifeStageParr	1.520	4	11.400	0.117
s(fyear)	9.480	16	61.700	0.0003
s(HACode)	0.012	23	0.018	0.474
s(HACode_re)	15.700	43	1,278.000	0.232
s(fCATCH_ID)	85.700	151	1,873.000	0.084
s(Water_W):LifeStageFry	1.220	2	33.200	0.00003
s(Water_W):LifeStageParr	0.917	2	16.800	0.0005
s(Elevation_):LifeStageFry	1.860	2	73.100	0.00000
s(Elevation_):LifeStageParr	1.080	2	27.600	0.0003
s(Distance_s):LifeStageFry	0.944	2	40.400	0.00002
s(Distance_s):LifeStageParr	1.920	2	156.000	0
s(sinSlope):LifeStageFry	0.817	2	6.820	0.018
s(sinSlope):LifeStageParr	0.751	2	3.940	0.046
s(Upcatch_km):LifeStageFry	1.870	2	235.000	0
s(Upcatch_km):LifeStageParr	0.628	2	2.490	0.094
s(Urban):LifeStageFry	0.972	2	49.700	0.00000
s(Urban):LifeStageParr	0.014	2	0.013	0.341
s(woodland):LifeStageFry	1.820	2	21.000	0.001
s(woodland):LifeStageParr	1.890	2	34.800	0.00001
s(Marsh):LifeStageFry	0.637	2	2.280	0.102
s(Marsh):LifeStageParr	0.688	2	3.480	0.081
s(doy):LifeStageFry	0.918	2	49.900	0.0005
s(doy):LifeStageParr	0.841	2	12.400	0.012

```
final <- c("LifeStage",
         "sinSlope",
         "Water_W + Water_W:LifeStage",
         "Urban + Urban:LifeStage",
         "doy + doy:LifeStage",
         "s(year, k = 5, by = LifeStage)",
         "s(fyear, bs = 're')",
         "s(HACode_re, bs = 're')",
         "s(fCATCH_ID, bs = 're')",
         "s(Elevation_, k = 3, by = LifeStage)",
         "s(Distance_s, k = 3, by = LifeStage)",
         "s(Upcatch_km, k = 3, by = LifeStage)",
         "s(woodland, k = 3, by = LifeStage)"
         )
finalf <- formula(paste("y ~", paste(final, collapse = " + ")))
```

and was fitted directly in mgcv

```
finalg <- gam(finalf, offset = offset, data = efdens, family = nb,
            select = TRUE)
```

A summary of the final model is given graphically in Figures 3.7 and 3.8. Fry densities were higher than parr densities (Fig. 3.7). The smooth year effect (Fig. 3.7) was had a modal response for fry, peaking around the year 2000, while for parr the effect was lowest in 2000 and highest in 2005. Opposite trends between fry and parr was also evident for altitude (Fig. 3.7) with fry showing a decreasing effect with increasing altitude and parr showing higher densities at lower and higher altitudes, with the lowest densities occurring around 200m. Density decreased with distance to sea (Fig. 3.7) for both fry and parr, but densities were lower at high distances to sea for parr than for fry. Upstream catchment area (Fig. 3.8) had a small impact on parr densities, but functioned to reduce fry numbers when upstream catchment area was small. For % woodland (Fig. 3.8), the effect was similar between fry and parr both showing a modal effect with the highest densities at 50% and the lowest densities at high woodland cover. The remaining effects (gradient, channel width, % urban and day of year, Fig. 3.8) all showed declining densities with higher covariate values.

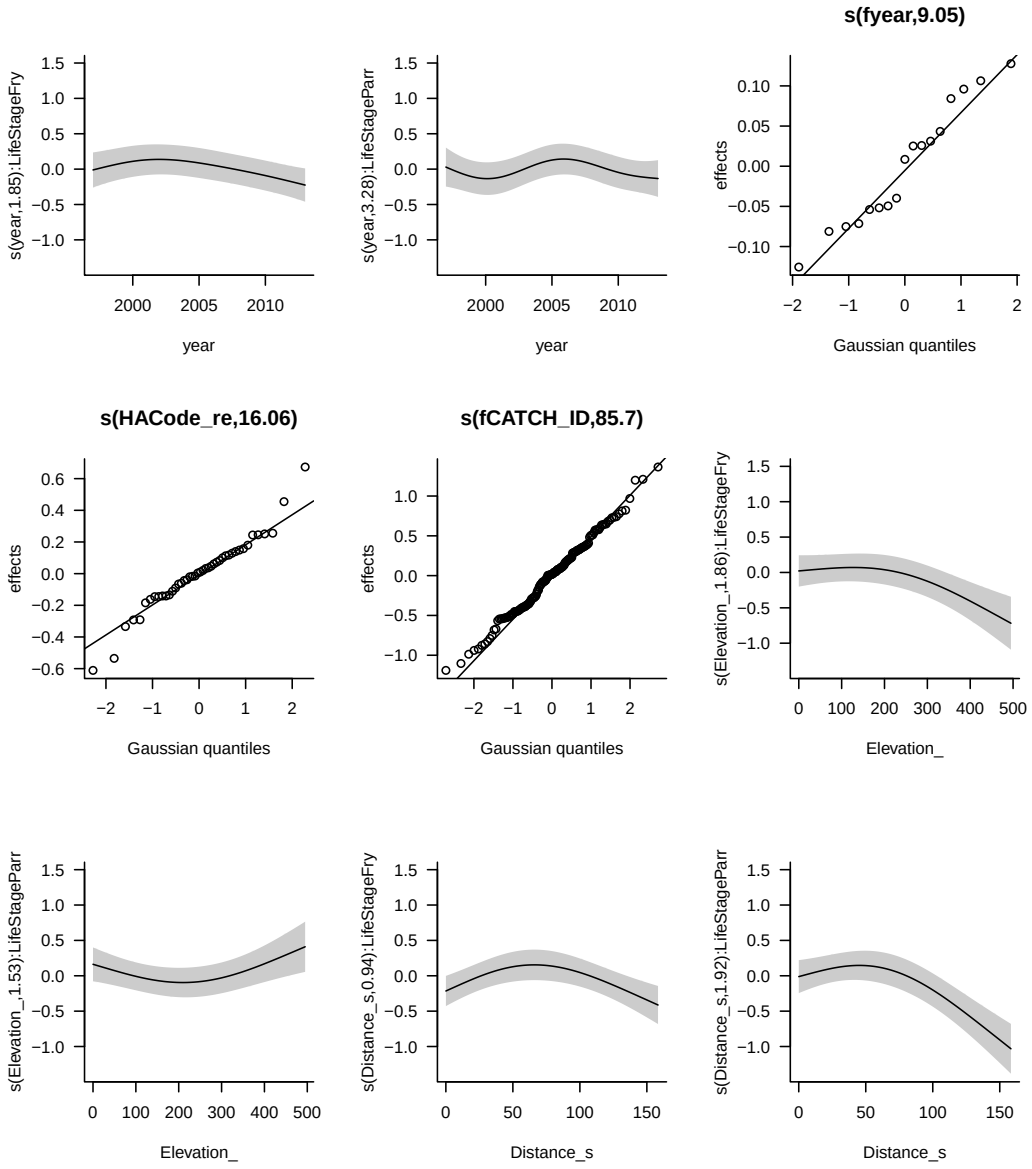


Fig. 3.7 Final density model summary plots. Shown are the effects on the log scale of each effect included in the final model. 95% point-wise confidence intervals are shown by shaded regions. For smooth effects, the effective degrees of freedom is included in the y-axis label. Each random effect is shown by a qq-plot to allow both an assessment of the magnitude of the effect and how close to Gaussian it is in distribution.

3.6.4 Error propagation

The procedure for carrying error from the capture probability model to the density model proceeds as follows:

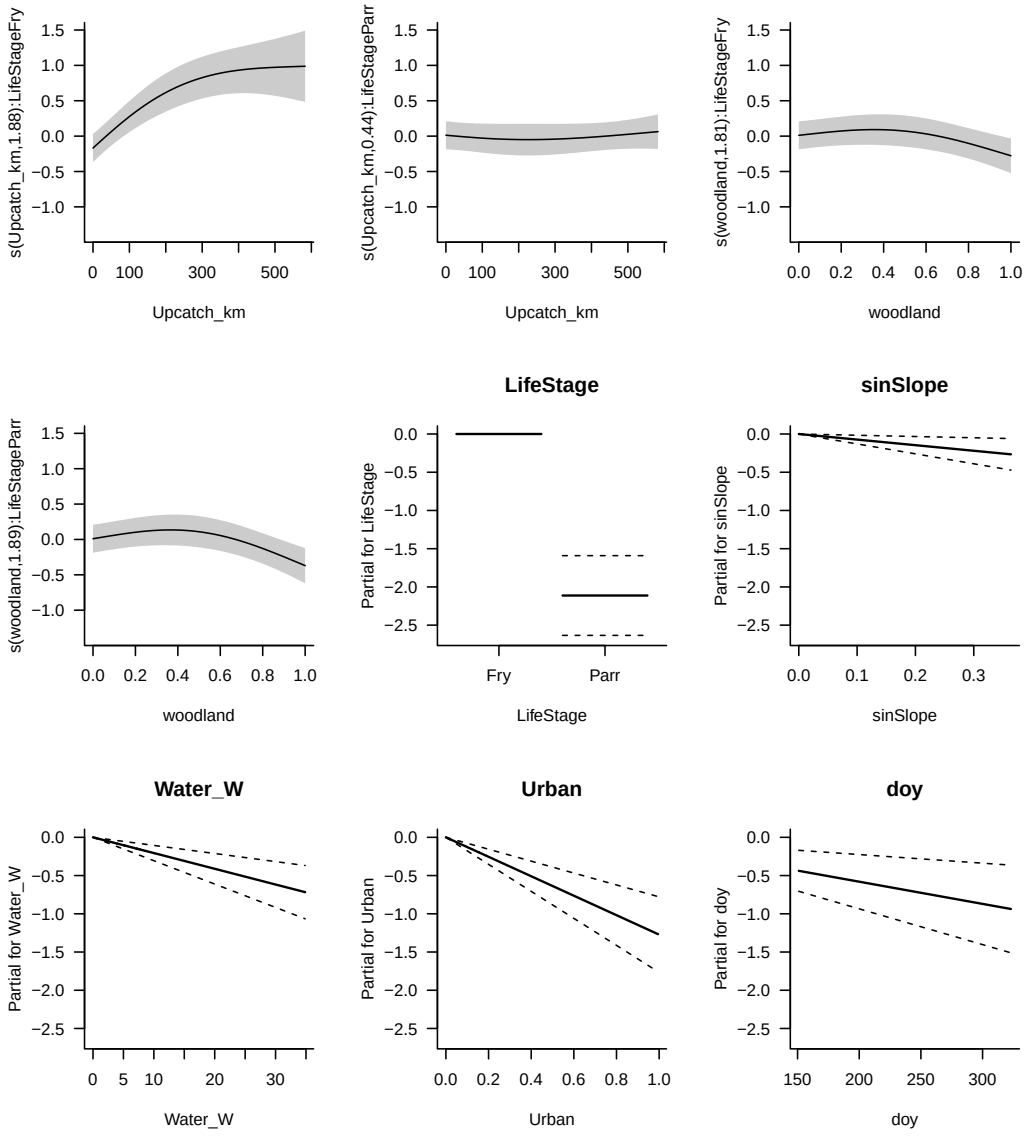


Fig. 3.8 Final density model summary plots. Shown are the effects on the log scale of each effect included in the final model. 95% point-wise confidence intervals are shown by shaded regions. For smooth effects, the effective degrees of freedom is included in the y-axis label. 95% confidence intervals for categorical and linear effects are shown as dotted lines.

1. model selection procedure for capture probability model

2. model selection for density model, condtional on capture probability model

3. simulate from capture probability model

4. refit density model to each simulation

5. simulate from the previous density model fit

Steps 1 and 2 have so far been completed, so in this section I will proceed with steps 3 to 5. Simulating from the final capture probability model is implemented using the following function, note that the variance matrix is inflated by the estimate of Ψ_{between} estimated from the final model.

```
simp <- function(model, data, nsim = 1000) {
  simb <- MASS::mvrnorm(nsim, coef(model), model $ Vb * 3.58)
  simpijk <- 1/(1 + exp(-model $ Gsetup $ X %*% t(simb)))

  tmp1 <- data %>% select(sampleID, LifeStage, pass) %>% as.data.frame(.)
  apply(simpijk, 2, function(p) {
    tmp1 $ pijk <- p
    tmp2 <- tmp1 %>% spread(pass, pijk)

    1 - (1-tmp2[["1"]]) * (1-tmp2[["2"]]) * (1-tmp2[["3"]])
  })
}
```

and 1000 simulations taken from the capture probability model

```
psim <- simp(finalm, ef, nsim = 1000)
```

and to each simulation the final density model was refitted and a parametric bootstrap simulation from the parameter estimates stored

```
gcoef_sim <- matrix(NA, length(coef(finalg)), 1000)
for (i in 1:1000) {
  efdens$simoffset <- log(efdens$Area * psim[,i])
  simg <- gam(finalf, offset = simoffset, data = efdens, family = nb,
              select = TRUE)
  gcoef_sim[,i] <- MASS::mvrnorm(1, coef(simg), simg$Vp)
}
```

The revised model fits with appropriate confidence intervals are plotted in figure 3.9. Catchment, hydrometric area and Year had substantial effects on fish density,

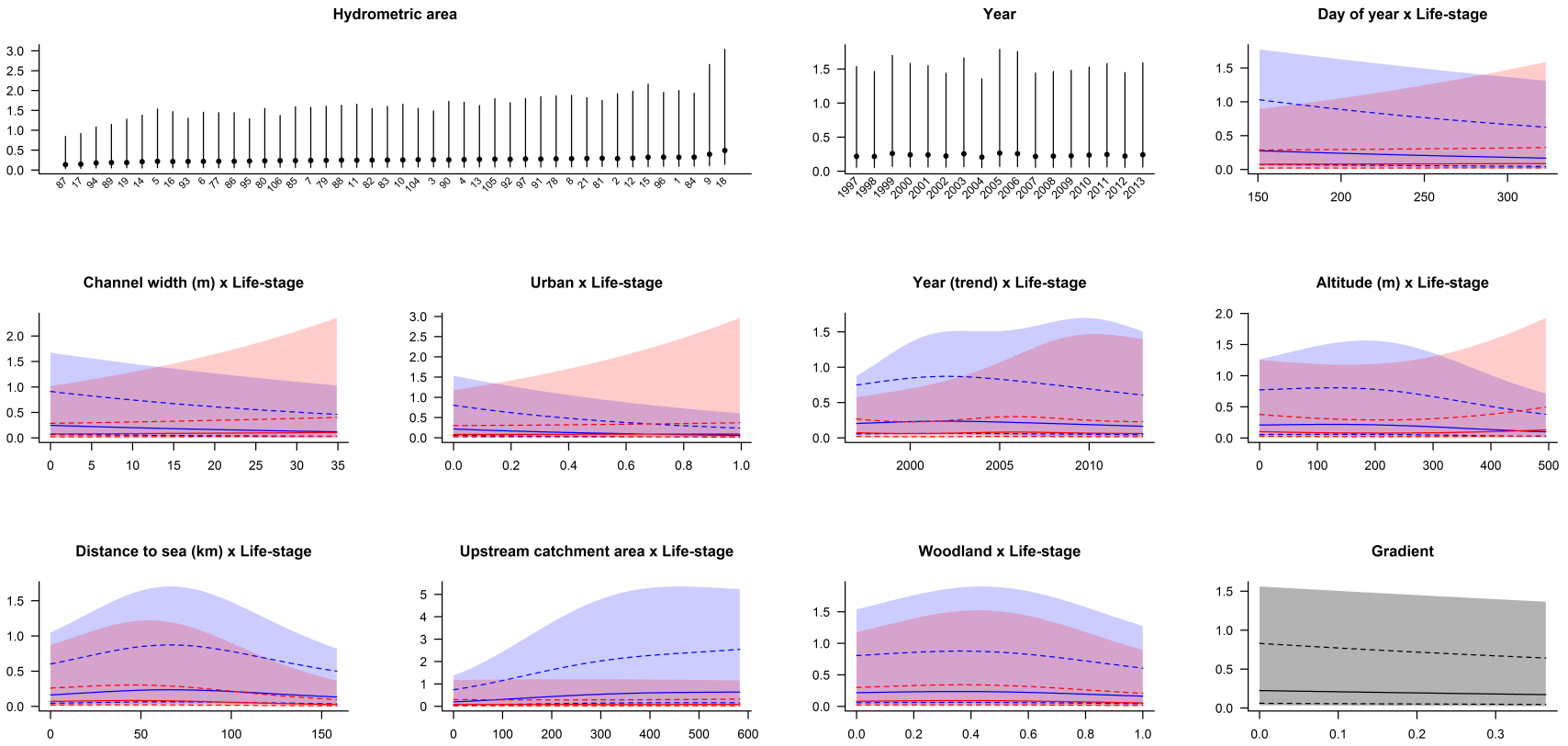


Fig. 3.9 Relationships between fish density and covariates for Fry (blue) and Parr (red). Plots are conditioned on Catchment Tay, HA Tay, Year 1996, and median value for all remaining covariates. Hydrometric area codes are presented for brevity. 95% pointwise confidence intervals incorporating error from the capture probability model are shown as shaded regions, with those without are shown as dotted lines.

with year being retained as both a random effect and a long-term trend (Figure 3.9). The greatest habitat effects were associated with Upstream catchment area, which exhibited an asymptotic trend for Fry (blue) but a negligible effect for parr. The remaining effects were of a similar magnitude and their impacts on density are summarised in the previous section.

The effect of propagating the error from the capture probability model into the density model was to approximately double the size of the confidence intervals, and in some cases, such as the smooth year trend 3.9), the increase in uncertainty was not equal over the covariate range, and resulted in large uncertainties in the recent years where the conditional model had narrowing confidence intervals.

3.7 Discussion

There were two developments in this chapter; the development of software (Section 3.4) to fit, inspect and test models of capture probability, and the proposed use of a two stage model for capture probability and density, with boostrapping used to propagate errors (Section 3.3).

Previous attempts to develop habitat models have generally followed one of three approaches: (1) simplified approaches considering only single pass electrofishing data (Wyatt, 2005) or assuming constant capture probability (SNIFFER, 2011) (2) approaches involving site specific estimates of capture probability (Lanka et al., 1987; Rosenfeld et al., 2000) or (3) Hierarchical Bayesian approaches where capture probability and habitat covariates are considered simultaneously (Rivot et al., 2008; Wyatt, 2002, 2003). The two stage approach in proposed in this chapter, together with the extension to mgcv, and the various GMRF models made available provide a substantial improvement over the previous approaches. The first approach is clearly

misleading since capture probability and habitat variability in abundance could be confounded. The second approach does not make best use of available information, particularly in circumstances where no fish are caught and thus no estimate of capture probability can be obtained. The final approach often results in models which are time consuming to specify and fit, thus making model comparison and selection difficult. Although the two stage approach may not provide better estimates with respect to the Bayesian hierarchical models in principle it may do in practice, through the rapid specification, fitting and selection of models. Hence, a wider model space for both capture probability and fish density can be explored.

An approach was explored here to carry the error from the capture probability model into the density model via a parametric bootstrap procedure. The approach succeeded in propagating the error and giving confidence intervals that were appropriate for the variation in the data. However, there still remain several problems with this approach. Firstly, because the final density model was selected conditional on the final capture probability model, refits of the final density model to simulated re-sampled capture probabilities are likely to have non-significant terms. That is, if the final density model had been selected with the error in the capture probability model taken into account, the slected model would have likely had fewer terms. There are two potential solutions to this: one could do model selection at each bootstrap re-sample (similar to Buckland et al. (1997) where the goal was to include model selection error), but this would introduce decisions to be made about how to combine the different models - for example should an ensemble approach be used?. Another, and simpler, approach is to reassess the final model after the capture probability error is incorporated. This might involve bootstrap hypothesis tests on the fixed parameters to see if they remain significantly different from zero with the added uncertainty. A more sophisticated approach would be to use the two stage approach for model selection and inform a joint hierarchical model. In order for this to solve the issue of identifying non

significant terms in the density model, the joint model would need to have had the same ability to shrink out smooth terms as can be done in mgcv, unfortunately this would not be trivial.

Many large-scale models of salmon abundance have implicitly ignored the effect of capture probability, focusing on single pass data or data from the first pass of multi-pass fishings (Wyatt, 2005), or assuming a constant capture probability (SNIFFER, 2011). These approaches are therefore inconsistent with the findings of a number of studies which have shown that capture probability varies with a range of factors including species and life stage (Borgstrom and Skaala, 1993), sampling protocol and personnel (Niemelä et al., 2000) and habitat characteristics (Kennedy and Strange, 1981). This study demonstrated that implicit (use of 1 pass data) or explicit (a single estimated capture probability) assumptions of constant capture probability are inappropriate simplifications. Capture probability varied substantially with organisation (which has a geographic component), with the time of electrofishing, and to a lesser degree with other covariates. Failure to correct for differences in capture probability would generate spatial bias and subsequently misleading spatiotemporal models of fish density.

Models of salmon fry density were successfully fitted to the available electrofishing data, based on an analysis of diagnostics rather than a comparison to other methods. Although some of the covariates (e.g., channel width) could have been affected by data processing errors (particularly snapping to river), there is no reason to assume that these estimates should be biased, and as such, these errors are only likely to introduce noise into the observed relationships. Furthermore, many of the important covariates which had a strong effect on abundance would be robust to these issues e.g., altitude and distance to sea. It is therefore likely that the resulting models provide a reasonable description of the spatial variability in fry abundance and the relative importance of covariates.

Having accounted for between catchment differences in the distribution of habitat covariates, the spatial covariates, HA and Catchment, describe residual variation in the abundance data. Taken together these covariates could be considered an average relative catchment performance metric over the monitoring period, so long as observed differences did not reflect uncharacterised natural variability in habitat that affects productivity and sampling was representative of the catchment as a whole. If the large scale smoothed variability indicated by HA was considered a natural trend (unaffected by human impacts), then Catchment alone could be considered an indicator of local performance.

The concept of reference or optimal condition underpins legislation such as the EU Water Framework Directive and is important for informing Conservation Limits (CLs) as it potentially identifies spatial variability in spawning requirements for different catchments. Using the current model a reference condition could be inferred by selecting the best Year, setting Catchment and HA to median values and setting the effects of pressures (% Urban and potentially % Conifer if this represents commercial forestry) to zero. An alternative approach would involve modelling of electrofishing data obtained from strategic stocking of sites over a broad geographic and environmental range. A further option could involve defining reference condition for a time where there were high adult salmon returns and lower human impact. Such an approach would require historical data from relatively un-impacted (reference) sites.

3.8 Recommendations for future work

A number of developments could improve the ability of models to characterise and predict spatio-temporal variability in fish abundance at the National scale. It has

only been possible to fit models for single species. It is suggested that future work considers interactions between species and life stages and develops capacity to fit within river variation over multiple river catchments with the aim of producing a single spatial model for salmonids.

The pragmatic choice to restrict the smoothers by limiting the degrees of freedom to 3 was easily implemented. A more flexible approach is to allow more degrees of freedom (more spline basis functions) but penalise with a fixed smoothing parameter, this way the effective degrees of freedom could be restricted to be approximately 3, while allowing a large model space and therefore a wider variety of curve shapes. The tools to implement this have been developed in the ef package (via penalised likelihood), however the link between the effective degrees of freedom and the smoothing parameter value was not developed. More work will be required to implement a smoother with effectve degrees of freedom equal to, say, 3, using say, 10 basis functions, in a general setting.

Although the existing datasets were spatially extensive, there is poor replication between Organisation and HA (or other spatial metrics). It is therefore possible that the effects of factors such as, equipment or methods, are confounded with spatial effects. This could be further investigated if data providers were to fish outside of their geographic areas in locations with similar GIS characteristics. There was also poor data coverage in the case of certain unusual (and broadly correlated) covariate combinations. Strategic data collection in areas of high upstream catchment area - low distance to sea (and vice versa) and low channel width - low distance to sea would improve spatial models and help to separate the effects of these variables.

The development of a capture probability model allows for the integration of single and multi-pass electrofishing data on the assumption that they were both collected in a consistent way (that is, using the same effort, equipment and staff). If single

pass data was included in future models, this would greatly increase the size and coverage of available data given the quantity of one pass data collected by fisheries trusts. To ensure that the capture probability models remain valid, it is recommended that fisheries trusts continue to generate multi-pass electrofishing data. Where there is only 1 pass data available for particular catchments it is recommended that this is supplemented with multi-pass data from which to estimate capture probability.

Finally, the selection of covariates in the current chapter was pragmatic given the need for large scale spatial coverage and the limited time available for the project. However, future work should consider the inclusion of metrics of hydrological impact, hydrochemistry and river temperature data given the importance of these variables in controlling fish distribution and abundance (Armstrong et al., 2003).

4

Modelling river temperature

4.1 Introduction

4.1.1 Motivation

Stream temperature plays an important role in the ecological health of streams and rivers. For salmon, increased stream temperature in summer has the greatest impact (Malcolm et al., 2008a) through increased stress, decreased growth and increased mortality. The implications of reduced temperature are less severe and affect early growth and development which can affect life history traits such as fecundity and maximum size (see, Malcolm et al., 2008a, 2003).

One important regulator of stream temperature is the shade provided by trees. This occurs in a variety of ways, from riverside trees which obscure the local water surface from the sun, to larger scale effects of forestry, which shade to a lesser degree, but cover large stretches of streams.

Due to these impacts there is increasing interest in the influence of shading on stream temperature. In North America, most research has focused on understanding the effects of forest harvesting, primarily to find ways of reducing the impact on maximum stream temperatures (Beschta and Taylor, 1988; Gomi et al., 2006; Macdonald et al., 2003; Mellina et al., 2002). In the UK there is an increasing interest in the use of woodland to limit increases in stream temperatures through climate change climate change (Hannah et al., 2008; Hrachowitz et al., 2010; Malcolm et al., 2008a). In both the US and UK contexts there is a requirement to estimate (with uncertainty) the stream temperature changes associated with variations in riparian cover in order to inform management.

4.1.2 Previous approaches

Previous approaches for assessing temperature effects associated with changing forest cover can be grouped by the type of control used to infer effects. One approach is to compare observations between adjacent reaches of the same river which have different landuse characteristics (Malcolm et al., 2008a; Zwieniecki and Newton, 1999), such as moorland or woodland, making the assumption that differences in temperature are due to differences in landuse; a related approach is to use similar catchments rather than the same river (Brown et al., 2010; Webb and Crisp, 2006). Where sites differ in physical characteristics, for instance, elevation and gradient, one approach is to model the effects of differences using linear regression (Hrachowitz et al., 2010). An alternative is to keep the site constant while changing the landuse, through tree felling or tree planting. A simple (and not very powerful) approach is to compare average daily summary metrics pre- and post felling (e.g., Stott and Marks, 2000), while a more robust experimental design is the use of a control site in a before-after-control-intervention (BACI) design. Early examples of this approach

have graphically compared observations (Feller, 1981), while more recent studies attempt to model a baseline response (Gomi et al., 2006), but are still visually assessing the effect of felling.

This last method by Gomi et al. (2006) is often called a paired or multi catchment design. Statistical modelling of temperatures from paired catchment designs date back to Brown and Krygier (1970), in which the problem of autocorrelation in time series of environmental data is highlighted. To circumvent this, simple approaches using Mann-Whitney U tests to compare mean monthly temperatures (Brown et al., 2010) or modelling them using linear regression (Fellman et al., 2014) are still common. These approaches circumvent the issue of autocorrelation because much of the autocorrelation exists at short time scales (hours and days), therefore, taking monthly averages ensures that each observation is practically independent.

The approach developed by Gomi et al. (2006), extending previous work by Watson et al. (2001), estimates the relationship between stream temperature at a control site and a study site using linear regression with auto-regressive residuals. The relationship is fitted under natural conditions before any felling has taken place. This baseline relationship is then used to predict the stream temperature at the study site after felling. Any difference in the predicted (baseline) temperature and the recorded temperature is then interpreted as a result of the felling.

The prefelling model takes the form

$$y_i = a + bx_i + c_1 \sin\left(\frac{2\pi \text{doy}_i}{365}\right) + c_2 \cos\left(\frac{2\pi \text{doy}_i}{365}\right) + ar_k(\text{day}_i) \qquad (4.1)$$

where y_i is the ith observation of daily mean stream temperature at the impact site and x_i the same for the control site, for $i = 1, \ldots, n$. The day of the year associated with the observations is doy_i and the number of days since the start of the study is

day$_i$. The kth order auto-regressive process $ar_k(j)$ can be defined as

$$ar_k(j) = u_j = \phi_1 u_{j-1} + \ldots + \phi_k u_{j-k} + \epsilon_j \tag{4.2}$$

where $j = 1, \ldots$ corresponds to consecutive days, and the error term ϵ_j is assumed to be iid normal

$$\epsilon_j \sim \mathrm{N}(0, \sigma) \tag{4.3}$$

that is the residuals are modelled as an AR(k) process, and the seasonal trend as a sinusoid. A felling effect is then implied by using this model to predict daily mean temperatures after the impact and inspecting the residuals

$$y_i^* - \hat{y_i^*} \tag{4.4}$$

There is no formal assessment of the significance of a felling effect, although it is inferred if the residuals are greater than that expected by prediction intervals for the stationary distribution of the fitted AR(k) process.

The attractiveness of this approach is that it does not make any parametric assumptions about the effect of felling. However, a limitation of this, and other existing methods, is that there is no quantitative assessment of the magnitude of the felling effect or for the speed of recovery as trees are replanted. The approach was extended by Leach et al. (2012) by modelling the residuals as an autoregressive integrated moving average process, but still relies on forecasting a no-felling-effect model to infer felling effects, albeit with a more complex Monte-Carlo procedure. More recent studies (Guenther et al., 2012; Moore et al., 2013) revert back to the simpler method of Gomi et al. (2006).

For computational reasons, most current model based assessments of felling use daily summary data such as the daily mean, minimum, or maximum temperature. Often, up to 96 daily observations (an observation every 15 minutes) are reduced to one or two summary measures, which are then modelled and interpreted in turn. This gives limited insight into the effect of felling on the cyclical within-day (or diel) changes in temperature. For example, they would not be able to estimate the length of time that a stream is above a critical temperature, but are restricted to statements about daily maximums or daily means.

4.1.3 Chapter structure

This chapter develops new methods for quantifying the effect of clear felling on stream temperature and the speed of recovery. It assumes that there are temperature time series from two streams (a paired catchment study), a control stream with no felling activity, and an impacted stream with data from both before and after a clear felling event. The methods can be used with daily temperature summaries, but can also be used to estimate the effects of felling on diel variability in temperature. The methods are illustrated using data from two burns in Loch Ard. The models used are structural additive regression (STAR) models. GMRFs smoothers were used initially, however computational difficulties lead to a more efficient approach being adopted.

Section 4.2 describes the data from Loch Ard.

Section 4.3 uses a subset of the data to formulate a simple model that motivates the subsequent methodological development.

Section 4.4 develops models that formally assess the effect of felling. These models use cyclic smoothers (see Chapter 2) to characterise seasonal variation. Felling is included as a parametric function. To account for short term climatological effects,

an AR1 process (see chapter 2) is used. I illustrate these fits using daily mean temperatures.

Section 4.5 develops methods to model diel variation in temperature by extending the daily model from 4.4 through the use of functional data techniques. It goes on to show how these models can be used to predict the effect of felling for different temperature regimes, and hence show the range of effects from cold to warm days. This is important for assessing the maximum impact on stream temperature, and thus the risks for salmon under different scenarios.

4.2 Data

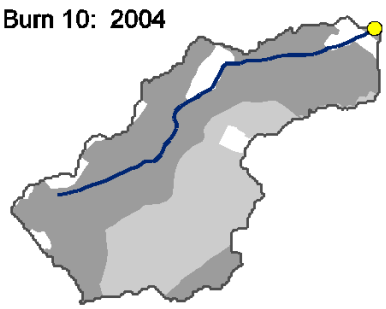
4.2.1 Study site

Stream temperature was monitored in two Burns (Burns 2 and 10) in the Loch Ard area of West Central Scotland. Figure 4.1 shows the location of the site and where the stream temperature was monitored on Burn 10 (yellow dot). Burn 10 is a forested site within an area of managed woodland while Burn 2 is an upland moor site.

4.2.2 Felling event

Figure 4.1b shows a map depicting a planned felling event that took place in Burn 10 in 2004. The felling occurred to the south of the river and left a buffer zone of trees between the river and the clear felling area. A buffer zone will reduce the impact on stream temperature but some may remain, for example if the clear felling lowers the horizon this would provide indirect sunlight for longer time period. It is not known

(a) Loch Ard location in Scotland, UK



(b) A close up of the Burn 10 catchment showing the river in blue, the sampling location (yellow dot), mature forest (dark grey), the area of trees removed during the 2004 felling (light grey), white areas have no tree planting.

Fig. 4.1 Study site

exactly when the felling took place, but evidence such as time series of nutrients in the water corroborate the year of the event.

4.2.3 Temperature monitoring

Stream temperature was monitored in the two catchments from 1999 to 2013. The full time series of data is shown in Figure 4.2.

A range of temperature dataloggers with various reporting resolutions (typically 0.01 - 0.05 ^{o}C) were deployed over the study period. This can be seen in Figure 4.2: data prior to 2008 shows as discretised because observations are rounded internally by the logger to increase the storage capacity of the logger. Reporting frequency also varied from hourly to every 15 minutes. However, the data loggers across the two sites were always the same make and model at any one time, but the make and model did change through time. Additionally, each data logger undergoes a calibration against a 'gold standard', prior to deployment. This improves the accuracy of the

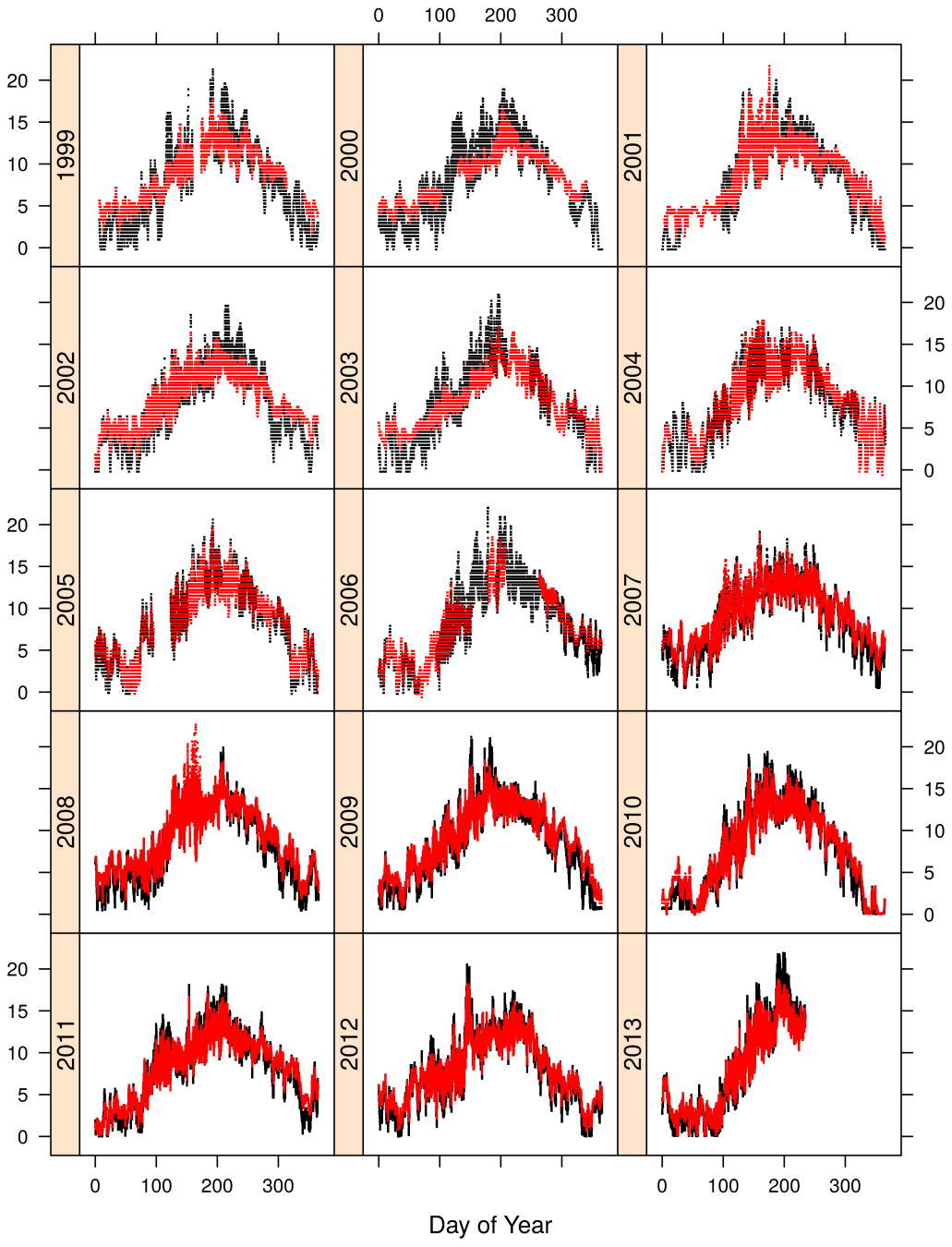


Fig. 4.2 Time-series of temperature observations from Burns 2 (black) and 10 (red), by year. The felling event took place in 2004.

recorded observations beyond their officially stated accuracy to within 0.02 ^{o}C of the true temperature.

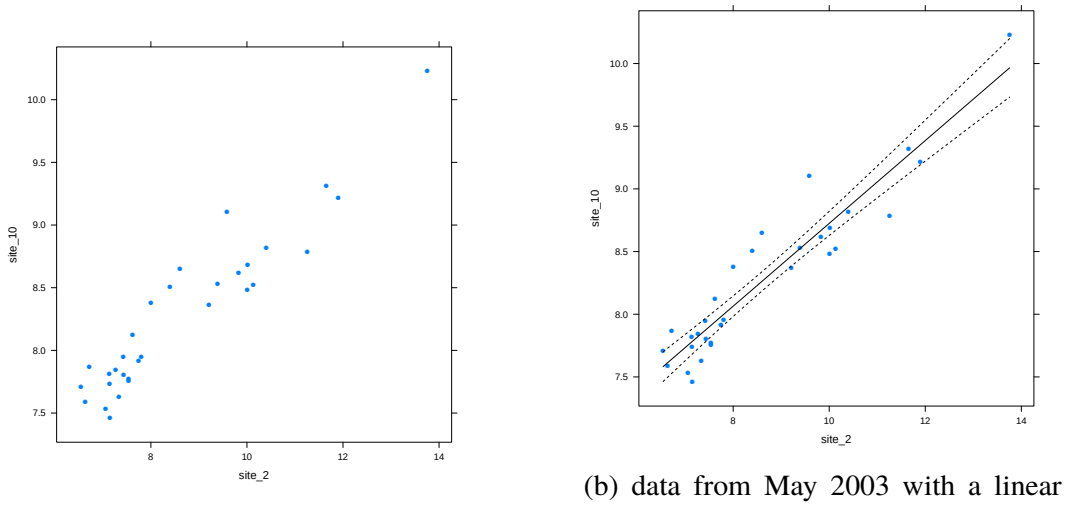
4.2.4 Data cleaning

Prior to analysis the data was screened to ensure that as far as possible the data was a true representation of the temperature in the stream. Data was removed in winter when the stream froze and the thermistor (the temperature sensor) was exposed to the air or insulated by snow; during dry spells in the summer when the water level dropped below the level of the thermistor; and during periods where the data-logger was clearly mis-calibrated (all temperatures were too high) in which case the time series of that particular logger deployment was removed. The cleaned dataset is presented in Figure 4.2.

4.3 Simple motivating model

4.3.1 Two streams

To motivate the modelling approach, and to introduce notation, consider two streams within the same river catchment with characteristics similar to those in Scotland. Because they share a similar location, and hence climate, you might expect them to be similar in terms of stream temperature. However, due to differences such as terrain, altitude, orientation and forestation, their temperatures could be quite different. They will, however, behave similarly, as the major factor controlling stream temperature is radiation from the sun. It should then be possible to model the behaviour of one



(a) Data from May 2003

(b) data from May 2003 with a linear model fitted

Fig. 4.3 Temperature observations from Burns 2 and 10 in May 2003.

stream in terms of another, and this is the basis of paired stream analyses such as Gomi et al. (2006).

I will begin by looking at the temperatures of our two study streams, Burn 2 and Burn 10, but restrict the investigation to a short period of time in order to ignore seasonal and temporal effects. Figure 4.3a plots the temperature in Burn 2 against Burn 10, according to measurements made during May 2003, the year prior to the felling event.

Table 4.1 Model summary for a simple linear fit of Burn 10 to Burn 2 using data from May 2003

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	5.4210	0.1905	28.45	0.0000
site2	0.3304	0.0216	15.32	0.0000

The temperatures are clearly correlated: for every degree Burn 2 changes Burn 10 changes on average by 0.33 ^{o}C and occupies a temperature range of 6.5 to 14.0 ^{o}C, compared to 7.4 to 10.5 ^{o}C for Burn 10. Figure 4.3a suggests that a linear model could be a good way to characterise the relationship between the temperatures in the

two streams. One way to proceed is to consider the linear model

$$y_i = a + bx_i + \epsilon_i \tag{4.5}$$

where $i = 1, \ldots, 31$ denotes the days of the month under study

$$\epsilon_i \sim \mathrm{N}(0, \sigma) \tag{4.6}$$

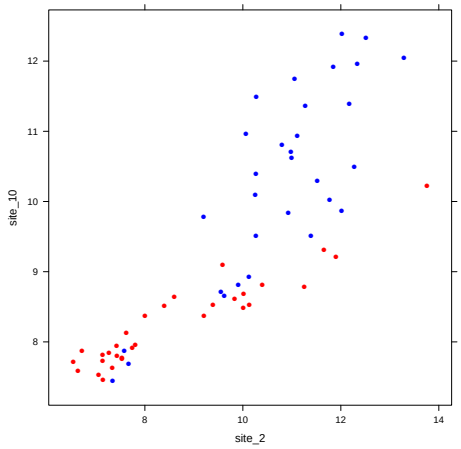
It is possible to consider such a model due to the high accuracy and precision of the observations made by the logger allows us to consider the temperatures in Burn 2 to be known approximately without error. This means that the residual variation is not strictly speaking an observation error, but is a process error for the hypothetical process linking Burn 2 with Burn 10.

Prior to analysis all temperatures have been centred around the mean daily temperature which for the time series is 7 ^{o}C. Hence, y_i is the ith (paired) observation of centred daily mean temperatures in Burn 10, and x_i is the same for Burn 2. I will ignore the fact x_i is also observed with error, and assume for now that the residuals are iid normal with variance sigma. This model can be written in matrix notation as follows

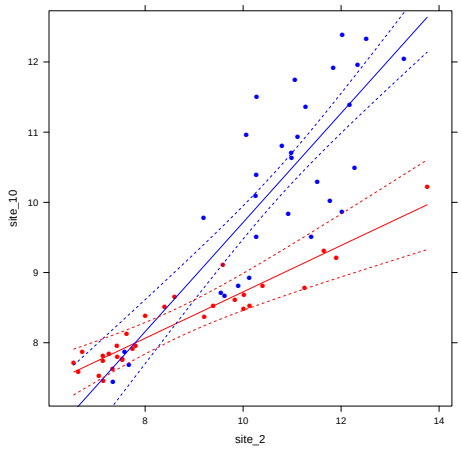
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.7}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, and X is a two column matrix formed from a column of 1s for the intercept a and $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. The parameters in the linear model are easily estimated using the maximum likelihood estimator

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{4.8}$$



(a) Data from May 2003 (red) and 2004 (blue)



(b) data from May 2003 (red) and 2004 (blue) with a linear model fitted seperately to each year.

Fig. 4.4 Temperature observations from Burns 2 and 10 in May 2003 and May 2004.

to be $a = 5.45$ and $b = 0.33$. This means that for every 1 ^{o}C change from the average 7 ^{o}C daily mean in Burn 2, Burn 10 changes by 0.33 ^{o}C. This is interesting as Burn 2 is surrounded by moorland, whereas Burn 10 is forested; the lack of river shading in Burn 2 seems to result in higher stream temperatures.

Table 4.2 Model summary for a simple linear fit of Burn 10 to Burn 2 using data from May 2003 and 2004 including a felling effect.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	5.4210	0.5223	10.38	0.0000
site_2	0.3304	0.0591	5.59	0.0000
felling	-3.4869	0.9584	-3.64	0.0006
site_2:felling	0.4476	0.0949	4.71	0.0000

4.3.2 A felling event

The felling event which is the focus of this chapter took place in the subsequent year, 2004, near Burn 10 (Figure 4.1b). The data plotted in Figure 4.4 indicates that there was an effect of the clear felling on the daily mean temperatures. One way to quantify the effect of this felling on stream temperature in Burn 10 is to introduce a parameter

for felling

$$y_{ij} = \begin{cases} a + bx_{ij} + \epsilon_{ij} & \text{if } j = 2003 \\ a + bx_{ij} + \underbrace{a_f + b_f x_{ij}}_{\text{felling effect}} + \epsilon_{ij} & \text{if } j = 2004 \end{cases} \tag{4.9}$$

where $i = 1, \ldots, 31$ denotes the days of the month under study, and $j = 2003, 2004$ are the years considered. The residual variance is assumed to be constant.

$$\epsilon_{ij} \sim \mathrm{N}(0, \sigma) \tag{4.10}$$

Figure 4.4 shows stream temperatures from 2003 and 2004 with the fitted lines from (4.9). The model can be written in the matrix form as before, hence the same maximum likelihood estimator of the parameters apply, however, this model has a design matrix $\boldsymbol{X}$ with 4 columns. The estimates of the felling parameters are a_f = -3.49 (SE 0.96) and b_f = 0.45 (SE 0.09); the values in brackets are the standard errors derived from the estimate of the variance of b which is given by the 4 x 4 matrix

$$\boldsymbol{V} = \frac{1}{\sigma^2}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{4.11}$$

The effect of felling on the intercept and slope in this model has a high probability of being non-zero and hence we conclude that there is a significant felling effect. Note that the slope has changed from 0.33 to 0.78 - so the removal of trees has made the temperatures in Burn 10 more similar to that in Burn 2 (because it is closer to 1). We can also predict, with confidence intervals, the effect of felling experienced during May of 2004 from the following

$$\text{felling} = a_f + b_f \boldsymbol{x} = \boldsymbol{X}_f \boldsymbol{\beta} \tag{4.12}$$

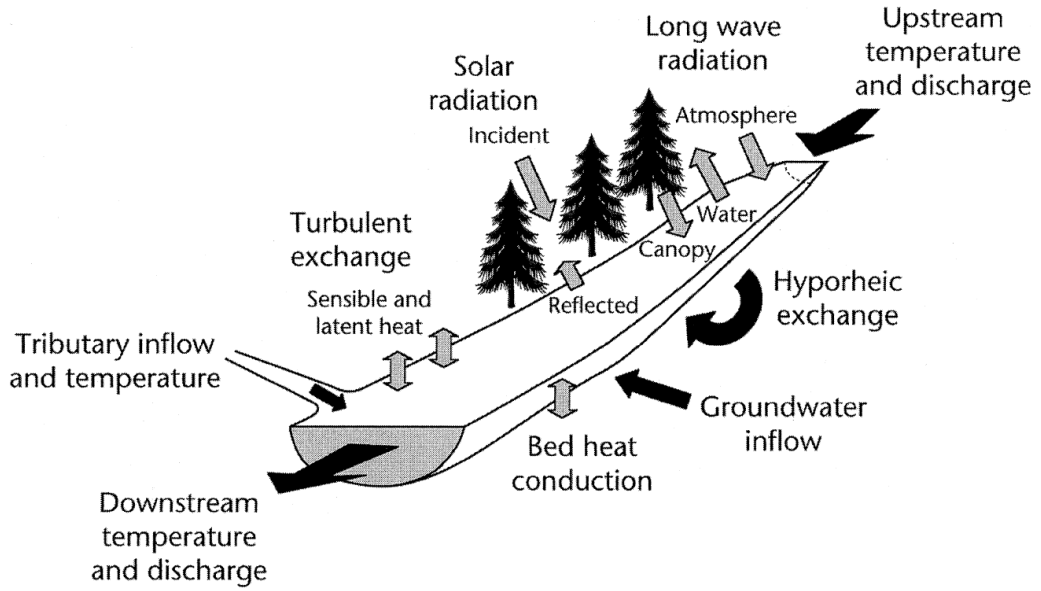


Fig. 4.5 Factors controlling stream temperature from (Moore et al., 2005). The greatest effect of forestry is the reduction of incident solar radiation through shading, however river canopies also affect net long-wave and evaporation. Energy fluxes associated with water exchanges are shown as black arrows.

Confidence intervals can be derived for the felling effect by linear transformation of the variance

$$\frac{1}{\sigma^2}\boldsymbol{X}_f\boldsymbol{V}\boldsymbol{X}_f' \tag{4.13}$$

where $\boldsymbol{X}_f$ is the design matrix from the model with the first two columns set to zero. The reason for the change in temperature is due to the removal of shading (see Figure 4.5), increasing the amount of radiation received from the sun and hence increasing the stream temperature (a_f) and the effect of increased sunshine (b_f).

Shade reducing input from the sun is one aspect of the effect of forestation on stream temperature. The other is that shade (through tree cover) also reduces heat loss, and so on cold days the removal of trees would be expected to result in greater decreases in temperature. To investigate this effect the model is extended to cover all the months in the year

$$y = \beta_a + \beta_b x + \beta_a(\text{moy}) + \beta_b(\text{moy})x + \epsilon \tag{4.14}$$

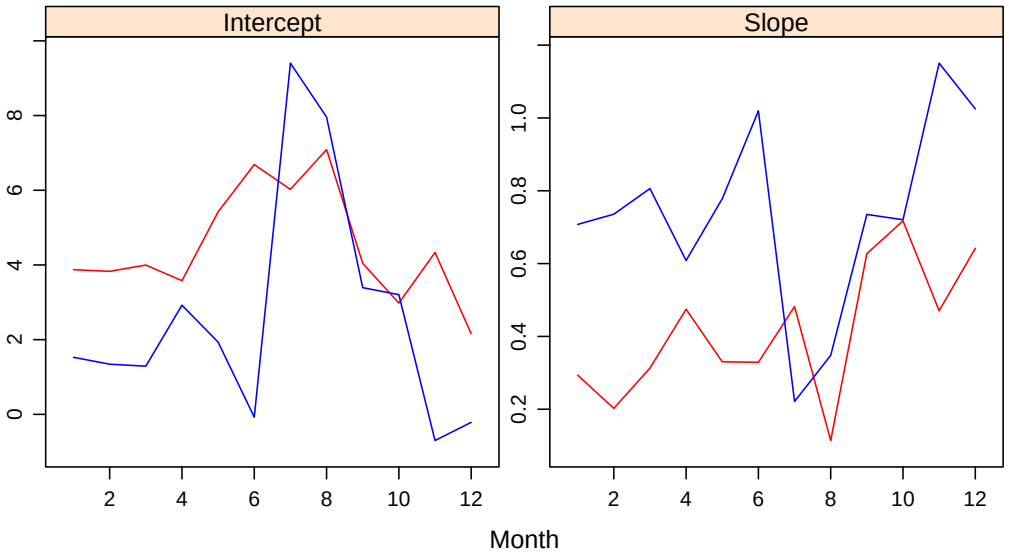


Fig. 4.6 Estimated intercepts and slopes by month for 2003 (red) and 2004 (blue).

The fits from the model are shown in Figure 4.6, and shows there is apparent seasonality in the intercept and potentially in the slope. The felling effect also appears to vary seasonally.

This basic approach forms the basis that will be used to estimate the effect of felling in the more complex situation where the relationship between sites varies continuously and where the effect of felling is estimated for multiple years, allowing for the possibility that the felling effect gradually reduces through time as vegetation and trees regrow, returning levels of shading back to that prior to felling.

4.4 Modelling daily felling effects

Preliminary analysis in Section 4.3 showed that a linear relationship holds between sites when viewed on a month by month basis (Figure 4.6). The models developed in this section for a felling effect are the same in spirit as the simple example: define a baseline model and a model for a felling effect in which the pre-felling data is given

by the baseline state which is then augmented by a felling effect to model the post felling data.

This section further extends this simple idea to allow for a felling effect which decays over time in several ways. GCV is used as the fitting criterion at each stage and also serves to highlight were additions to the felling decay model provide improved fits.

4.4.1 Smoother choice

To investigate the utility of each approach the following model was fitted to the full time series of data

$$y = a + x + g_a(doy) + \epsilon \tag{4.15}$$

with 6 forms for the function g_a: 1) CRW1, 2) harmonic, 3) one value for each week, 4) reduced rank CRW1 with 52 bases, 5) reduced rank harmonic with 52 bases, and 6) reduced rank harmonic with 100 bases. 52 was chosen to restrict the smoothness to a weekly level. Figure 4.7 shows the fitted smoothers. The results from CRW1, reduced CRW1, harmonic, and reduced rank harmonic with 100 bases are very similar; the factor model is very wiggly as expected, and the reduced harmonic with 52 bases is also very wiggly, however, when the number of bases is increased to 100, the model is able to shrink down to a smaller size. The edf of the least wiggly models are between 22 and 27 (Table 4.3), with the harmonic models attaining the lowest smoothing parameter and among the lowest GCV. Interestingly, the reduced harmonic model with 52 basis functions does not find the optimum smoother, but the reduced harmonic smoother with 100 basis functions does. The lowest GCV actually came from the reduced rank CRW1 model.

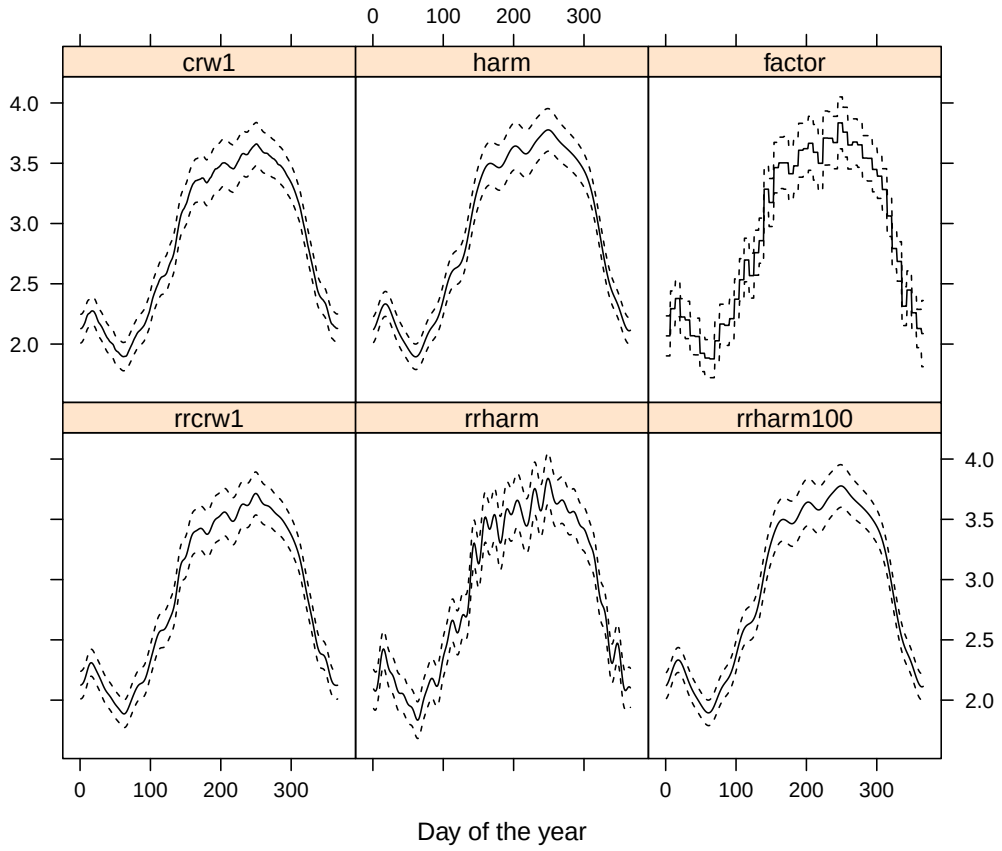


Fig. 4.7 Fits of several cyclic smoothers to the full dataset using model (4.15). CRW1: cyclic 1st order random walk; harm: harmonic; rr denotes reduced rank versions; factor: an independent parameter for each week and cc: cyclic cubic spine. The dotted lines represent pointwise 95% confidence intervals.

Table 4.3 The model degrees of freedom (edf) and generalised cross validation (GCV) score for various cyclic spline models fitted to the full time series using model (4.15)

model	edf	GCV
crw1	27.258	0.54106
harm	21.848	0.54009
factor	52	0.54478
rrcrw1	25.589	0.54004
rrharm	50.948	0.54308
rrharm100	21.854	0.54009

4.4.2 Simple model

The structural form of a relationship between two streams is likely to be seasonal as shown in Figure 4.6. I use the following STAR model which is a generalisation of that used by Gomi et al. (2006)

$$y = \beta_a + \beta_b x + g_a(\text{doy}) + g_b(\text{doy})x + \epsilon \tag{4.16}$$

y is the mean daily temperature at Burn 10 centered by 7 ^{o}C; 7 ^{o}C is the average stream temperature across the whole year. Similarly, x is the mean daily temperature at Burn 2 also centred by 7 ^{o}C. The function $g_a(\text{doy})$ is the seasonal non-linear effect of doy (day of the year) for an average day. The term $\beta_b + g_b(\text{doy})$ can be interpreted as the seasonal non-linear effect for every degree difference from an average day. This model allows for a seasonal difference between Burn 2 and Burn 10 and a seasonal difference in how warmer or cooler temperatures in Burn 2 translate to Burn 10. The slope can also be thought of as how reactive to changes in temperature Burn 10 is, compared to Burn 2.

I use reduced rank CRW1 smoothers to model the functions g_a and g_b which are assumed to be common across years. First let's consider a full rank GMRF smoother. Let the vector $\boldsymbol{\gamma}_a = (\gamma_{a,1}, \ldots, \gamma_{a,365})$ represent the daily seasonal effect of day of the year described by the function $g_a(\text{doy})$. Then, in matrix notation, the effect on any

given day is given by $\boldsymbol{Z}_i\boldsymbol{\gamma}_a$, where $\boldsymbol{Z}_i$ is a (row) vector of length 365, with a 1 in the column for the appropriate day and zeros elswhere, so if the ith observation took place on the 40th day of the year, $Z_{i,40}$ is 1, and $Z_{i,j}$ for $j \neq 40$ is 0. This allows a design matrix to be specified for the days on which paired observations took place, so that the function g_a can be evaluated at an arbitrary subset of days of the year through

$$g_a(\text{doy}) = \boldsymbol{Z}_a\boldsymbol{\gamma}_a \tag{4.17}$$

For the interaction term $g_b(\text{doy})x$, a similar approach can be taken. Let the centered Burn 2 daily mean temperatures be denoted by x_{ij}. Then the interaction term effect on any given day i, in any year j, is $x_{ij}\boldsymbol{Z}_i\boldsymbol{\gamma}_b$, so that the Burn 2 temperatures can be absorbed into the design matrix

$$g_b(\text{doy})x = \boldsymbol{Z}_b\boldsymbol{\gamma}_b \tag{4.18}$$

The intercept and slope terms β_a and β_b can also be writted in matrix notation. The vector equation for the centered daily mean temperatures in Burn 10, $\boldsymbol{y}$ is

$$\boldsymbol{y} = \boldsymbol{Z}_a\boldsymbol{\gamma}_a + \boldsymbol{Z}_b\boldsymbol{\gamma}_b + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.19}$$

4.4.2.1 Model fitting

These models can be fitted as full rank GMRFs using the mgcv extension given in Chapter 2. However, for computational efficiency, rather than specifying a full rank GMRF and reduce the dimension, I simply use 52 Fourier basis functions directly. I choose 52 to allow for a maximum of weekly level variability, whilst still keeping the size of the model to a manageable size. This choice of basis resonates well with the Gomi et al. (2006) model which can be thought of as an unpenalised spline using the

first 3 Fourier basis functions. Because I am expressly seeking to model a sinusoidal function I use the harmonic penalty in conjunction with the Fourier bases.

Because there has been a change of basis, the penalty also changes, as we saw before. Consider a new basis which can be transformed via a linear transformation $\boldsymbol{T}$ back to the original bases, i.e. $\boldsymbol{Z}_{\text{old}} = \boldsymbol{T}\boldsymbol{Z}_{\text{new}}$, The penalty matrix for the new bases is related to the original penalty $\boldsymbol{Q}_{\text{old}}$ by

$$\boldsymbol{Z}'_{\text{old}}\boldsymbol{Q}_{\text{old}}\boldsymbol{Z}_{\text{old}} = \boldsymbol{Z}'_{\text{new}}\boldsymbol{T}'\boldsymbol{Q}_{\text{old}}\boldsymbol{T}\boldsymbol{Z}_{\text{new}} \tag{4.20}$$

that is $\boldsymbol{Q}_{\text{new}} = \boldsymbol{T}'\boldsymbol{Q}_{\text{old}}\boldsymbol{T}$

Consider these transformations have already taken place, from the GMRF type model to one with Fourier bases. Now, equation (4.19) defines the model, where the matrix Z is built up of 52 Fourier bases functions and the penalties for the parameters are derived from the harmonic penalty on the functions g.

The coefficients $\boldsymbol{\gamma}_a$ and $\boldsymbol{\gamma}_b$ are then estimated in `mgcv` where the transformed penalties are included as before as

$$\lambda_a\boldsymbol{\gamma}'_a\boldsymbol{Q}_{harm}\boldsymbol{\gamma}_a + \lambda_b\boldsymbol{\gamma}'_b\boldsymbol{Q}_{harm}\boldsymbol{\gamma}_b \tag{4.21}$$

The model estimates of the intercept and slope from model (4.19) are given in Figure 4.8. This model will be referred to as 'model 1' in the following sections. The smoothers contain a lot of detail: they contain local (high frequency) time variations: we are interested in longer range effects. The use of multiple years of data should reduce problems with over-fitting, however, missing data in some years is perhaps causing over-fitting problems.

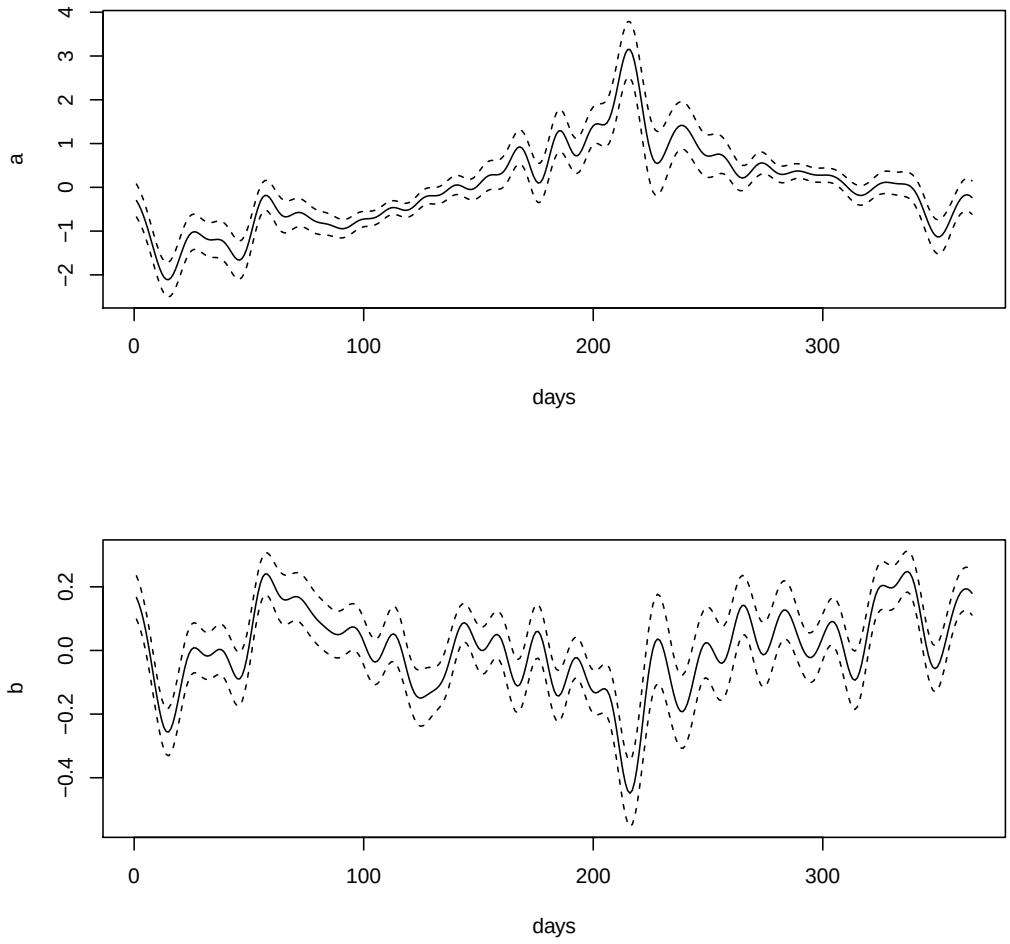


Fig. 4.8 Fitted smooth intercept and slope functions from Model 1 fitted to daily mean temperatures over the full time series (1997 to 2013) including both the prefelling and the felling periods. Only the smooth effects are shown here as the purpose of these plots is to investigate the degree of local-scale variation in the fitted smoothers.

4.4.3 Adding AR1

4.4.3.1 Modelling options

There are several ways to isolate long term stationary random trends. I implement a simple approach where local trends are assumed to be well described by a first order auto-regressive process (AR1).

It is possible to model the AR1 process in the model as an extra structural smoother. Because an AR1 process is a GMRF, this precision matrix can be used directly as a penalty matrix

$$\boldsymbol{Q}_{ar1}(\phi) = \begin{pmatrix} 1 & -\phi & & & \\ -\phi & 1+\phi & -\phi & & \\ & & \ddots & & \\ & & -\phi & 1+\phi & -\phi \\ & & & -\phi & 1 \end{pmatrix} \tag{4.22}$$

The model extended to include local variation is

$$y = \beta_a + \beta_b x + g_a(\text{doy}) + g_b(\text{doy})x + ar1(\text{doy}) + \epsilon \tag{4.23}$$

The problem with this approach is that the size of the model increases with the number of observations so that fitting time becomes limiting. The effect in this case is rather extreme as the AR1 process requires as many parameters as days. The other problem is that the AR1 can be confounded with the observation error (i.e. the signal to noise ratio of the AR1 is not strong). This can be remedied by fixing the regression error to a very small value, however, this can result in problems with parameter estimation if using GCV and would require switching to a criterion closely related to AIC called UBRE (UnBiased Risk Estimator). However, in all these cases, the AR1

process is considered as part of the model and contributes to the model degrees of freedom.

Since I am not interested in the AR1 itself, an alternative, used in many applications, is to model the residuals as AR1 directly. In this model the AR1 process is imposed on the residuals by weighting the residuals in the fitting process using the square root of the precision matrix above, that is, $\boldsymbol{Q}_{ar1} = \boldsymbol{D}'_{ar1}\boldsymbol{D}_{ar1}$ where,

$$\boldsymbol{D}_{ar1}(\phi) = \begin{pmatrix} 1 & -\phi & & \\ & 1 & -\phi & \\ & & \ddots & \\ & & & 1 & -\phi \end{pmatrix} \tag{4.24}$$

This model can be written, conditional on the AR1 parameter ϕ in matrix notation as

$$\boldsymbol{y} = \boldsymbol{Z}_a\boldsymbol{\gamma}_a + \boldsymbol{Z}_b\boldsymbol{\gamma}_b + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}_{ar1}(\phi)\boldsymbol{\epsilon} \tag{4.25}$$

4.4.3.2 Model fitting

Model fitting is achieved using a doubly nested iterative procedure. In the inner most loop, the regression coefficients are estimated conditional on ϕ and λ. The smoothing parameters λ are estimated by optimisation conditional on the AR1 parameter ϕ, and in the outer most loop ϕ is estimated. In practice, the smoothing parameters λ are estimated using the `gam` function in `mgcv` using generalised cross validation, and code was written to optimise over ϕ using the GCV score as the objective function.

The model estimates of intercept and slope are presented in Figure 4.9, the estimate of the AR1 correlation was 0.83. This model is hereafter referred to as 'model 2'. It

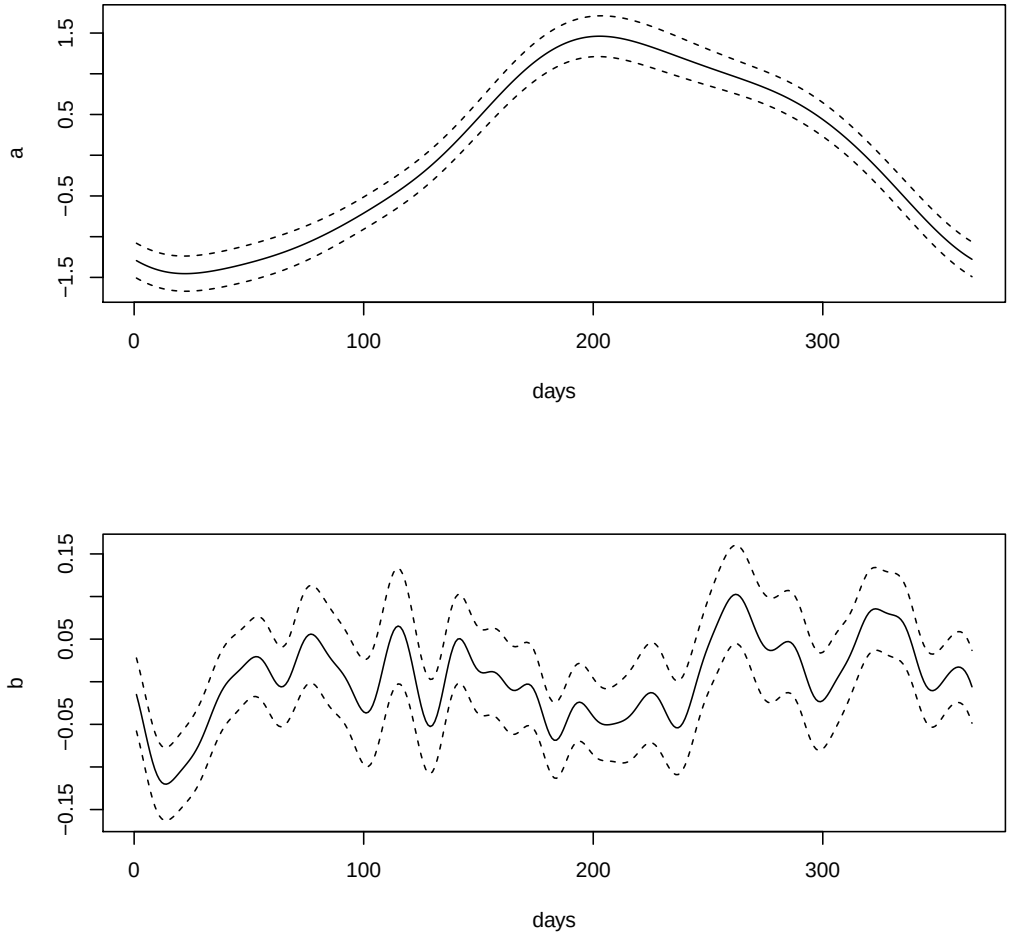


Fig. 4.9 Fitted smooth intercept and slope functions from Model 2 fitted to daily mean temperatures over the full time series (1997 to 2013) including both the prefelling and the felling periods.

can be seen that the intercept is much smoother. The slope, however, is smoother than the previous fit but still retains a lot of variability.

4.4.4 Introducing a felling effect

4.4.4.1 Felling models

The effect of felling will be modelled according to

$$\beta_f x + d_a(\text{doy}) \tag{4.26}$$

that is, felling alters seasonal differences for an average daily temperature, while the multiplicative effect of warming or cooling changes by a constant amount. The function $d_a(\text{doy})$ is modelled as a 2nd order cyclic random walk, hence the felling effect on the sampled days can be written

$$\boldsymbol{F}\boldsymbol{Z}_a\boldsymbol{\gamma}_c + \boldsymbol{F}\boldsymbol{x}\beta_c \tag{4.27}$$

where the matrix $\boldsymbol{F}$ is a square matrix and the ith diagonal entry is 1 if the ith observation comes from a felling period, and zero otherwise. It is clearly possible to combine $\boldsymbol{F}\boldsymbol{Z}_a = \boldsymbol{Z}_c$ and $\boldsymbol{F}\boldsymbol{x} = \boldsymbol{x}_f$, say. Hence the full model, including a felling effect and an AR1 process is

$$\boldsymbol{y} = \boldsymbol{Z}_a\boldsymbol{\gamma}_a + \boldsymbol{Z}_b\boldsymbol{\gamma}_b + \boldsymbol{Z}_c\boldsymbol{\gamma}_c + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{4.28}$$

where

$$\boldsymbol{X\beta} = \begin{pmatrix} \mathbf{1} & \boldsymbol{x} & \boldsymbol{x}_f \end{pmatrix} \begin{pmatrix} \beta_a \\ \beta_b \\ \beta_c \end{pmatrix} \tag{4.29}$$

For identifiability reasons there is no seasonal relationship in the felling effect for the slope. Preliminary analyses showed that when the felling slope effect was allowed to be seasonal, there were cases when all the non-linearity went into the felling slope effect rather than the baseline slope effect. Since the fits above show a seasonal effect on the slope in the prefelling period, it was assumed that the baseline relationship was the more complex and the felling slope effect was constrained to be a constant.

As a first step, the following simple constant felling effect model was fitted to the data. The corresponding structure of the matrix $\boldsymbol{F}$ is simply

$$F_{ii} = \begin{cases} 0 & \text{if year}_i < 2004 \\ 1 & \text{if year}_i \geq 2004 \end{cases} \tag{4.30}$$

This model will be referred to as 'model 3'. The fitted felling effect is plotted in Figure 4.10. This shows that the impact of felling is to increase temperature in winter and decrease temperature in summer. However, these predictions are made for an average day and deviations from it, since an average day is 7 ^{o}C. A winter day is likely to be 5 ^{o}C cooler than this, hence in winter the effect of felling corresponds to the lower dashed line. Similarly in summer, the temperatures are typically 5 ^{o}C or more warmer, so that the effect of felling should be read from the upper dashed and dotted lines. In practice then, the effect of felling would be better interpreted by predicting the effect of felling for an average temperature curve throughout the year.

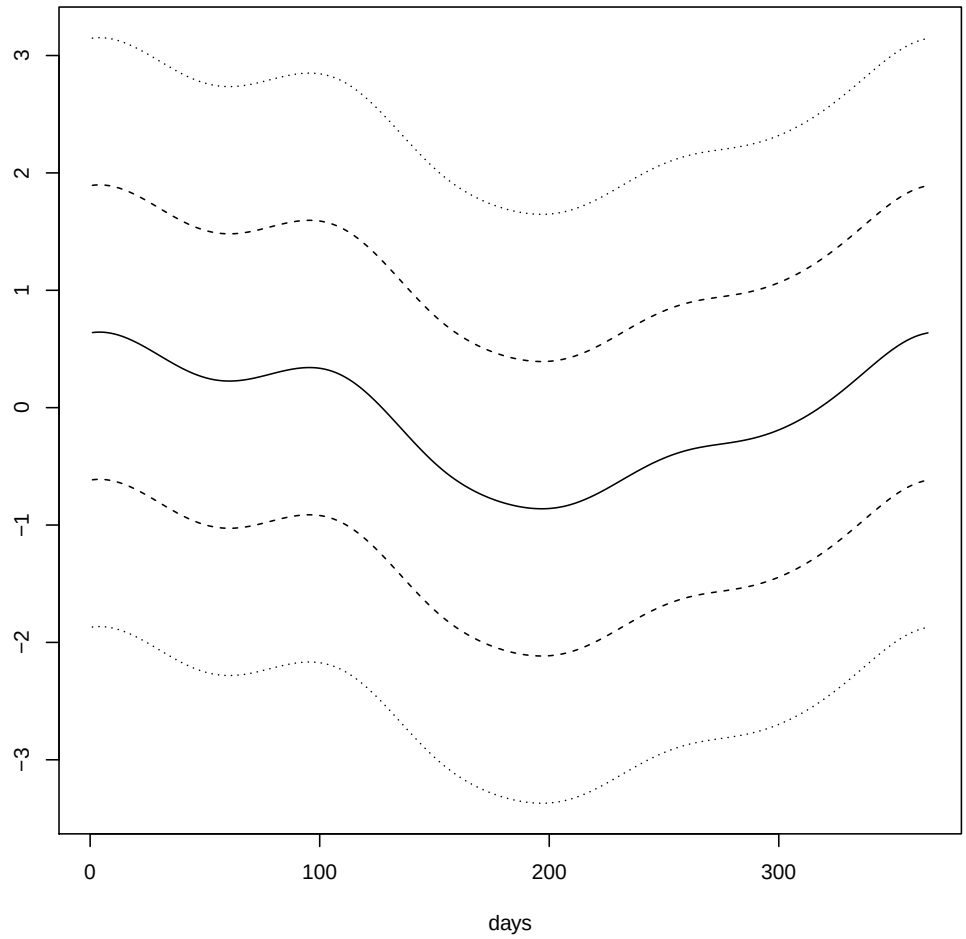


Fig. 4.10 The effect of felling on Burn 10 assuming model 3. The lines show the effect of felling on an average (7 ^{o}C) day (solid line), for +/- 5 ^{o}C (dashed lines), and for +/- 10 ^{o}C (dotted lines)

The residuals from this model are shown in Figure 4.11. The residual pattern includes AR1 noise, so may have local trends but should on the whole be stationary. It appears that there may be positive residuals beginning two years after the felling event and lasting for around two years. This could be due to climatological variation. The model described here does not deal with year to year variability and there may be some trends such as these.

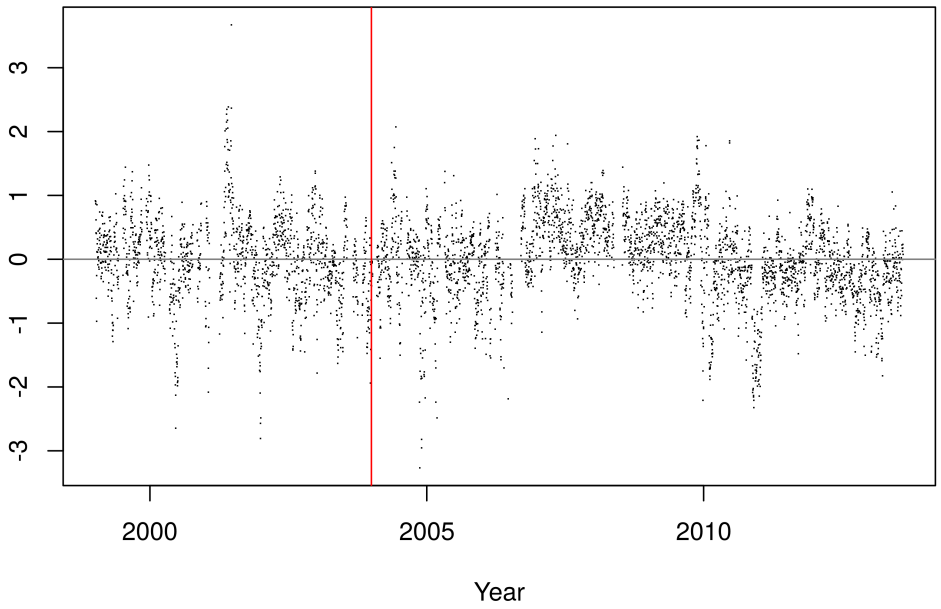


Fig. 4.11 The timeseries of residuals from the fit to Model 3. The red line indicates where the felling effect parameters enter the model.

There are several options to extend the basic model. One option is where the felling effect can decay over time,

$$F_{ii}(\rho) = \begin{cases} 0 & \text{if year}_i < 2004 \\ 1 & \text{if year}_i = 2004 \\ r_\rho(i) & \text{if year}_i > 2004 \end{cases} \tag{4.31}$$

where $r_\rho(i)$ is a smooth monotonic function decaying from 1 to zero and the index parameter $\rho > 0$ controls the length of recovery. Larger values of ρ produce longer recovery periods. For example, recovery can be modelled as an exponential decay

$$r_\rho(i) = \exp\left\{-\frac{1}{\rho}\left(\frac{\text{doy}_i}{365} + \text{year}_i - 2005\right)\right\} \tag{4.32}$$

Other possibilities for decay include cumulative distribution functions, half normal or half t distributions, or monotonic functions. It is sensible to choose a function that has a gradient of zero when $r_\rho(i) = 1$.

4.4.4.2 Model fitting

I consider five possible felling models in total, called model 3 to model 7. Note that model 1 is the model without a felling effect or an AR1 error structure, and model 2 is the model without a felling effect with an AR1 error structure. Model 3 is a constant felling effect

$$r_\rho(i) = 1 \tag{4.33}$$

Model 4 is an exponentially decaying felling effect

$$r_\rho(i) = \exp\left\{-\frac{1}{\rho}\left(\frac{\text{doy}_i}{365} + \text{year}_i - 2005\right)\right\} \tag{4.34}$$

Model 5 allows for a change in the start time of felling and a variable decay through

$$F_{ii}(\boldsymbol{\rho}) = \begin{cases} \exp(-\frac{1}{\rho_1}(\text{year}_i - \mu)^2) & \text{if year}_i < \mu \\ \exp(-\frac{1}{\rho_2}(\text{year}_i - \mu)^2) & \text{if year}_i \geq \mu \end{cases} \tag{4.35}$$

where $\boldsymbol{\rho} = (\rho_1, \mu, \rho_2)$ are parameters that govern how quickly the felling effect increases to its maximum impact, where the maximum is, and how slowly the felling effect decays. Model 6 extends model 4 to allow a different rate of decay for the intercept d_a and the slope β_f. Model 7 extends model 5 in the same way, that is model 7 maintains the same onset of felling for the intercept and slope but allows them a different rate of decay.

Each model is fitted using the same iterative procedure described for model 2, with the additional felling parameters being estimated in the outer iteration. Table 4.4 summarises the GCV scores of each model.

Table 4.4 The generalised cross validation (GCV) score for each model fitted to the daily mean temperatures.

	GCV score
Model 1	0.50145
Model 2	0.14357
Model 3	0.12808
Model 4	0.12807
Model 5	0.12805
Model 6	0.12806
Model 7	0.12797

The final model for a daily felling effect is thus model 7, although it should be noted that the GCV scores of models 3 to 7 are very similar and thus provide fits that are very similar in quality. There is strong evidence for an effect of felling, however, there is less evidence for a decline in the impact of felling. This is in line with expectations. Figure 4.2 shows the daily mean temperature time series for Burns 2 and 10. The yearly temperature range experienced by Burn 2 is fairly constant throughout the time series, whereas for Burn 10, there is a distinct increase in the range of temperatures experienced during the year of felling, and this change does not appear to reverse in the following 10 years.

The decay of the felling effect from model 7 is shown in Figure 4.12. It is possible that because the exact date of felling was not known, felling may have begun in late 2003. Model 5 and model 7 both have improved GCVs over similar models which fix the onset of felling to the start of 2004. Additionally, there is evidence that the slope in the relationship between the two burns does not recover, while the mean difference does very slightly. Even so, there still remains the same trends in residuals (Figure 4.13) identified from the constant felling model (Figure 4.10). This is not that surprising given that all models estimated a very small amount of recovery.

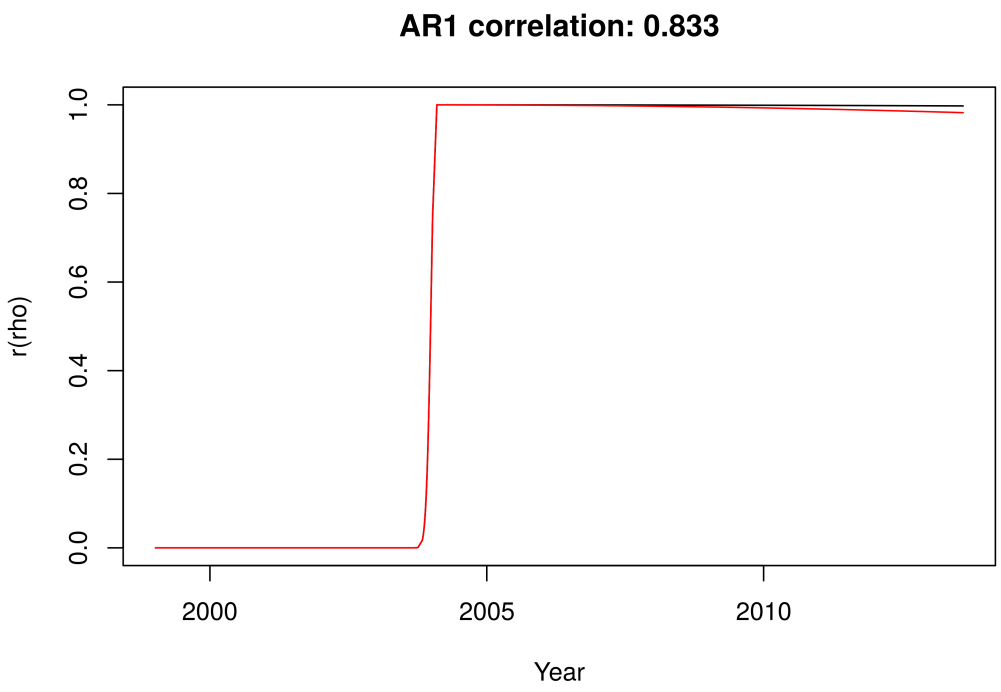


Fig. 4.12 The predicted decay of felling effect for the intercept (black) and slope (red) for Model 7.

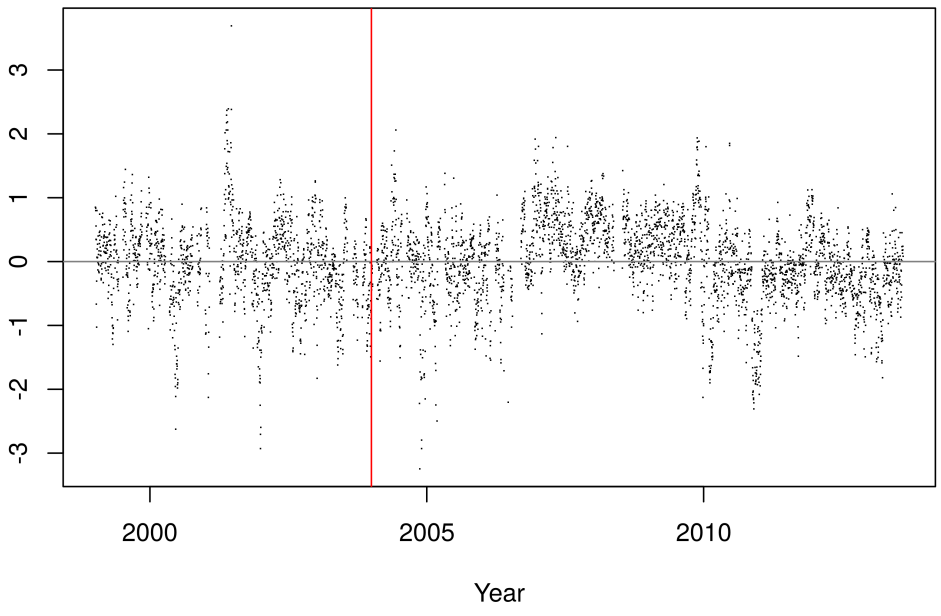


Fig. 4.13 The timeseries of residuals from the fit to Model 7. The red line indicates where the felling effect parameters enter the model.

4.5 Modelling sub-daily variation in felling effects

4.5.1 Functional data representation

Major computation issues are encountered when attempting to model sub-daily data over multiple years due to the extra level of cyclicity required in the model. This is the prime reason that most large scale models are based on daily summary data. Representing continuous data as functional data provides a potential way to model at sub-daily scales, with a little extra computational cost.

The idea behind functional data analysis (Ramsay and Silverman, 2005) is that a data vector, which can be thought of as a curve, can be considered as a single observation. Hence, for stream temperature we would consider the daily temperature curve as an observation. I have already shown how a continuous process can be modelled from discrete observations through the use of splines. For example, for a vector of temperatures within a given day $\boldsymbol{y}$, a possible model is

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{4.36}$$

where $\boldsymbol{Z}$ is a matrix of smooth orthogonal basis functions associated with the time of day each observation in $\boldsymbol{y}$ took place. The estimates $\hat{\boldsymbol{\gamma}}$ can be considered as a transformation from the data space into the parameter space,

$$\hat{\boldsymbol{\gamma}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \tag{4.37}$$

where $\boldsymbol{H}$ is a projection matrix, often referred to as the *hat* matrix. Because the orthogonal bases describing the parameter space are smooth functions in the data space, this means that each point in the parameter space corresponds to a smooth function in the data space. Therefore, curves in the data space (i.e. curves describing

smooth variations in temperature throughout the day) correspond to points in the parameter (or function) space.

To see the problem from a different angle, consider the case where $\boldsymbol{Z}$ has as many basis functions as there are data. This becomes a one to one transformation projecting the observation from the data space into a so-called function space. A common example of such a projection is where the basis functions are sines and cosines - the well known Fourier transformation. Analyses based on Fourier transformed data are often called spectral analyses, the data space is referred to as the time domain, and the function space as the frequency domain.

These techniques are also widely used in signal compression and image compression (Strang, 2009), for example an early Joint Photographic Experts Group (JPEG) algorithm for compressing images is based on using the first 8 Fourier basis coefficients to reduce 64 bits of information to 8. This is very similar in nature to the methods for reduced rank smoothers described in Chapter 2. In this section, the goal is to reduce the size of a daily temperature curve, composed of up to 96 observations, to a manageable size without losing too much information.

4.5.2 Basis functions

There are many, many possible orthogonal basis functions to choose from when attempting to describe a curve by a small number of values. An example of an inefficient choice of basis (in this context) is the use of Fourier basis functions to describe a square wave: in order to describe this curve with arbitrary accuracy it takes infinite degrees of freedom. A much better (or more efficient) choice of basis would be a zero-order B spline. The contrary example is that zero order B splines

are hugely inefficient for describing sinusoidal curves, where Fourier bases are as about efficient as you can get.

Daily temperature curves have a sinusoidal nature, but are not so simple that they can be described by the first two Fourier bases: a cos(x) + b sin(x). But how to find suitable basis functions? One way to find a good set of basis functions is functional principal components (for details see Ramsay and Silverman, 2005; Silverman, 1996) which find combinations of bases that explain the largest amount of variability, much like the eigenvalue decomposition does in reduced rank approximation (Chapter 2). However, whereas in rank reduction the eigen vectors provide new bases, in FPCA new bases are provided by eigen functions. Functional principal components is a method for finding the most efficient basis representation for a given set of bases. Although the mathematical procedures differ between standard PCA and FPCA, the result is the same: linear combinations of the bases. And choosing the first few components provides a means to efficiently approximate a curve using a small number of bases

$$\boldsymbol{y} \approx \boldsymbol{Z}\boldsymbol{U}\boldsymbol{\gamma}' \tag{4.38}$$

where $\boldsymbol{Z}$ can be a one to one transformation matrix or a restricted set of bases, that is, $\boldsymbol{Z}$ is $n \times p$ where $p \leq n$, and $\boldsymbol{U}$ describes the linear combinations found by FPCA and is $p \times k$ where k is typically small. In the application here, $k = 4$, chosen to be large enough to capture a substantial part of the variability.

In this application I use functional principal components to find the best basis functions to describe diel temperature variation. The new basis functions are shown in Figure 4.14. These are the results of a functional PCA applied to daily temperature curves in which observations were taken every 15 minutes resulting in 96 observations per day. The daily temperature curves were estimated separately for each day using a large number of Fourier basis functions (47) penalised by the harmonic accelerator

penalty so that there were on average 20 degrees of freedom. Penalisation was necessary as data sampling frequency varied and so it was not possible to converge the data without loss to a functional representation based on a common set of bases for the whole dataset. A compromise was to use a large number of bases, but to penalise so that the models were identifiable.

The results of the FPCA have much in common with PCAs used in morphometric studies: the percentage explained by component 1 is large, however the other (lower order) components are often the interesting ones which explain variation in shape rather than size. Likewise, the 1st functional principal component explains more than 97% of the variability with a basis function that models changes in the daily mean while also providing an increase in the temperature range at higher daily temperatures. The remaining components model aspects of the shape of the daily temperature curve. The 2nd component allows increases in the amplitude of the daily temperature, while the third and fourth components account for variation in phase shift and rates of heating and cooling (Figure 4.14).

This process provides a projection from temperature data to a four parameter functional summary, which act much like a daily mean and min and max, but which also allows the diel temperature curve to be reconstructed. This process is very similar to signal compression and reconstruction techniques used to digitise and save audio and visual data. The steps taken to compress an image, say, are as follows:

1. Transform pixel values $\boldsymbol{y}$ into a functional (more appropriate) representation $\boldsymbol{\gamma} = \boldsymbol{Z}^{-1}\boldsymbol{y}$. In this step no information is lost, $\boldsymbol{\gamma}$ is the same size as $\boldsymbol{y}$.

2. Retain only the most useful bits of $\boldsymbol{\gamma}$ by retaining the first 8 (for example) lowest resolution basis functions. In JPEG compression, Fourier basis functions are used, and so the higher frequency basis functions are dropped. In this step

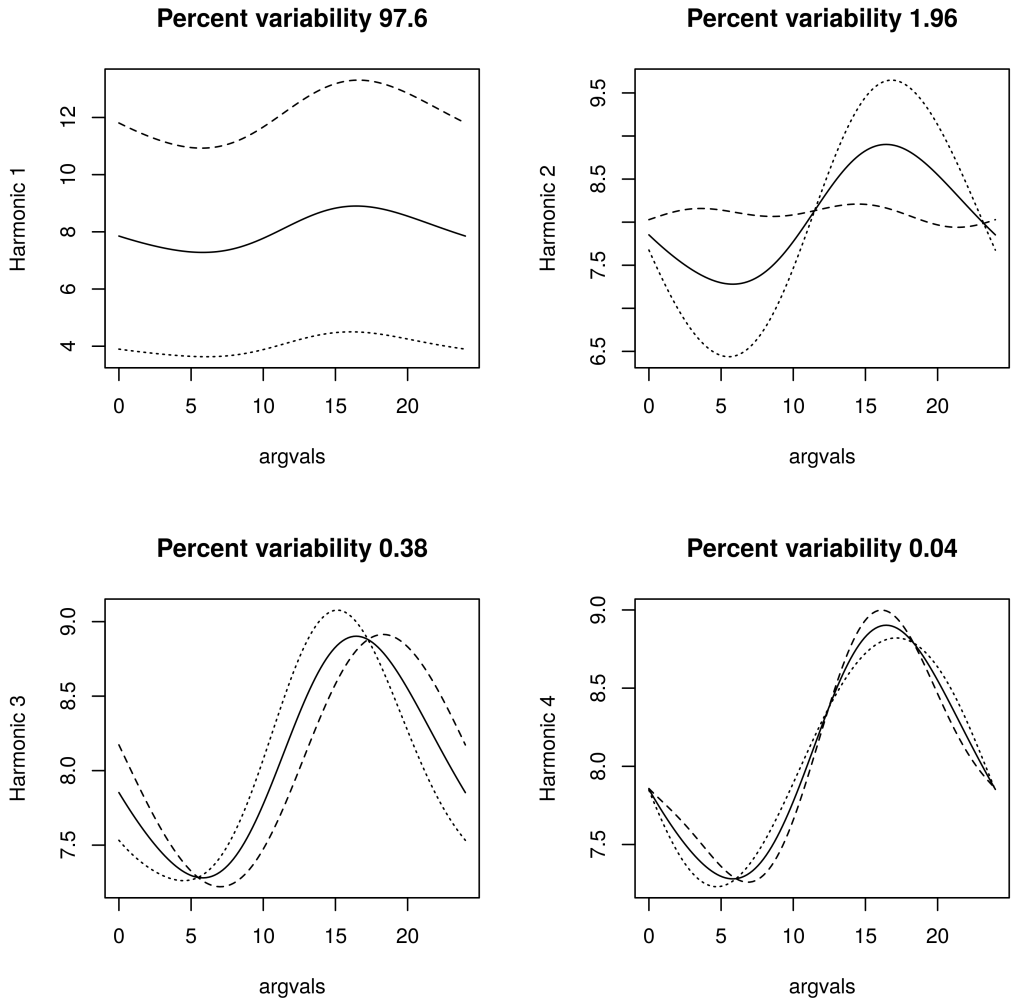


Fig. 4.14 A graphical depiction of the results of the functional PCA conducted on the raw temperature data. The components are presented in order of decreasing percentage variance explained. The dashed and dotted lines show the addition and substraction of the PCA basis functions to the mean daily curve. The solid line shows the daily mean curve and is repeated in each panel.

some information is lost, and the resulting *compressed* information is stored in a vector $\hat{\boldsymbol{\gamma}}$, say.

3. To reconstruct the image from the compressed information the reverse of the original transformation takes place using, $\hat{\boldsymbol{y}} = \boldsymbol{Z}_c\hat{\boldsymbol{\gamma}}$, where $\boldsymbol{Z}_c$ is the first 8 basis functions of $\boldsymbol{Z}$.

This describes what is known as a lossy compression, and to increase the resolution of a compressed image more combinations of basis functions are retained. The art of image compression is in finding efficient basis functions that are suitable for a wide range of images. In more recent versions of the JPEG compression algorithm, e.g., JPEG2000, wavelet rather than Fourier basis functions are used.

4.5.3 Model fitting

The regression models developed in Section 4.4 are fitted separately to each functional principal component (FPC) shown in Figure 4.14. The modelling process is related to the compression algorithm given previously. Model parameters are estimated for each component and predictions are made for each component. This step occurs between stage 2 and 3 of the compression algorithm. Model predictions are then combined as in stage 3 to give a predicted daily temperature curve and predictions of daily felling effect curves. The idea is to fit the models developed in the previous sections to each FPC in turn and then convert the model predictions back into sub-daily temperatures. A key feature here is that the FPC are, by design, orthogonal, so it is possible to model each PFC as an independent time series.

Figure 4.15 shows the transformed (and compressed) data. There are apparent felling effects in each component, the higher order components are less correlated than the 1st and 2nd components.

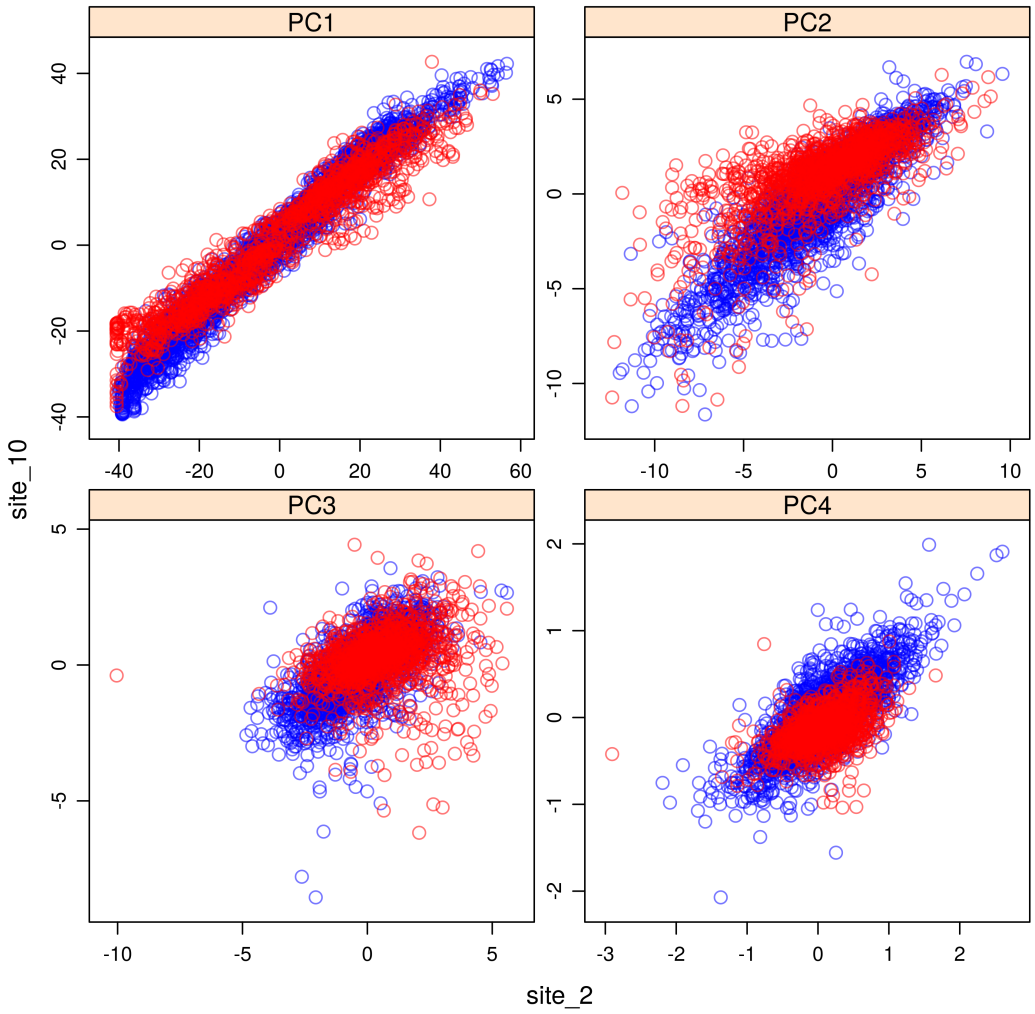


Fig. 4.15 Scatter plots of the four functional principal component scores. Burn 2 is plotted against Burn 10 with data from the prefelling period shown in red, and data post felling in blue

Table 4.5 The generalised cross validation (GCV) score for each model fitted to each set of functional principal component scores

	PC1	PC2	PC3	PC4
Model 3	3.042	1.214	0.473	0.043
Model 4	3.039	1.214	0.473	0.043
Model 5	3.038	1.213	0.472	0.043
Model 6	3.038	1.214	0.471	0.043
Model 7	3.037	1.213	0.470	0.043

The various felling decay rate models are applied and the GCV of each are presented in Table 4.5. The best model in each case was model 7, indicating that the felling effect decayed differently for the intercept and the slope felling terms and that there is evidence that onset of the felling effect was not the 1st January 2004. The findings in the daily felling model are echoed here, as there was not a great deal of improvement in fit from a simple model with a constant felling effect.

The final model fits on the component scales are given in Figures 4.16 and 4.17. Figure 4.16 shows the model estimates of the intercept and slope smoothers g_a and g_b for each component, as in Figure 4.8, the intercept smoothers are less variable than the slope smoothers, and tend to show a stronger seasonal pattern. Given that the first component is related to the daily mean, Figures 4.8 and 4.16a show similar patterns: the slope of the relationship between Burn 2 and Burn 10 is variable but doesn't show a seasonal trend, this implies that the range of temperatures in Burn 10 induced by a given range of temperatures in Burn 2 does not change seasonally. However, with this steady relationship, the temperature difference between the sites increases during the summer and decreases during the winter. Figure 4.16b shows the same for the component that was identified with the daily temperature range. This figure shows the daily range having a different seasonal pattern for the intercept g_a which shows peak differences in the spring and autumn. This is possibly indicative of the orientation of the sites, Burn 2 may receive more solar input during the spring, where Burn 10 receives more in the autumn. The intercept smoother for the 3rd component

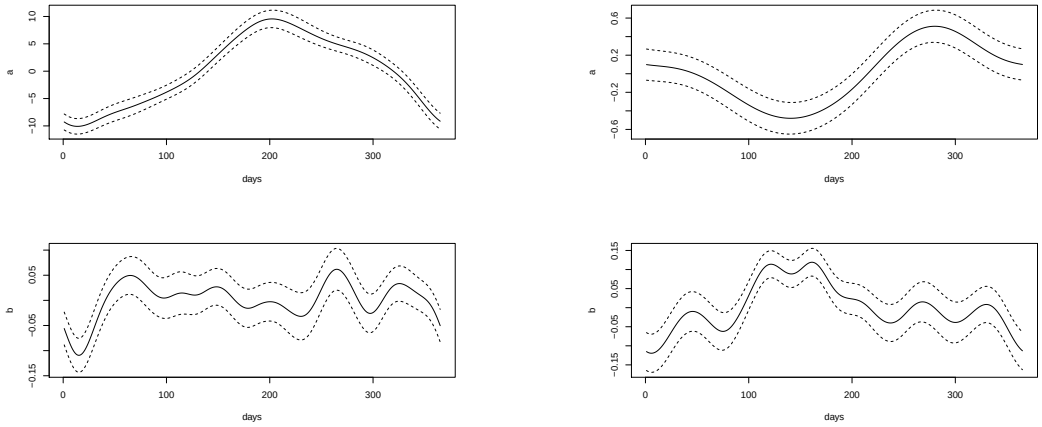
(the phase shift) is given in Figure 4.16c, and this shows a steady relationship with the exception of the spring when there is a wave effect. The reason for this effect is not clear. The smoothers for the 4th component (heating and cooling rate) are given in Figure 4.16d. The 4th component intercept is shows a similar but less pronounced trend as the intercept for the 3rd, while the slope of the 4th component shows two peaks around the spring and the autumn.

Figure 4.17 shows the predicted decay of the felling effect for each component with the estimated AR1 correlation given above each plot. In general the autocorrelation reduces with higher order components and the felling effects to not substantially decay, the exception is the decay of the intercept effect for the 3rd component.

It is difficult to interpret these effects in isolation. In order to predict the effect of felling on actual daily temperatures, the modelled basis coefficients need to be recombined using the transformation

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{U}\hat{\boldsymbol{\gamma}} \tag{4.39}$$

Using the same idea in (4.12) it is possible to extract the predicted felling effects. The fitted felling effects for each component were combined into 4-vectors and each transformed into a diel effect of felling on temperature for the days used in model fitting. Figure 4.18 summarises these fitted felling effects by plotting the average bi-monthly felling effect (Jan-Feb, Mar-Apr, etc.). Positive effects on temperature are distinguished from negative effects by colour. The general pattern shows that the clear felling has increased temperatures in summer and decreased temperatures in winter, and that the range of the response within a day is greatest towards the end of the first half of the year.



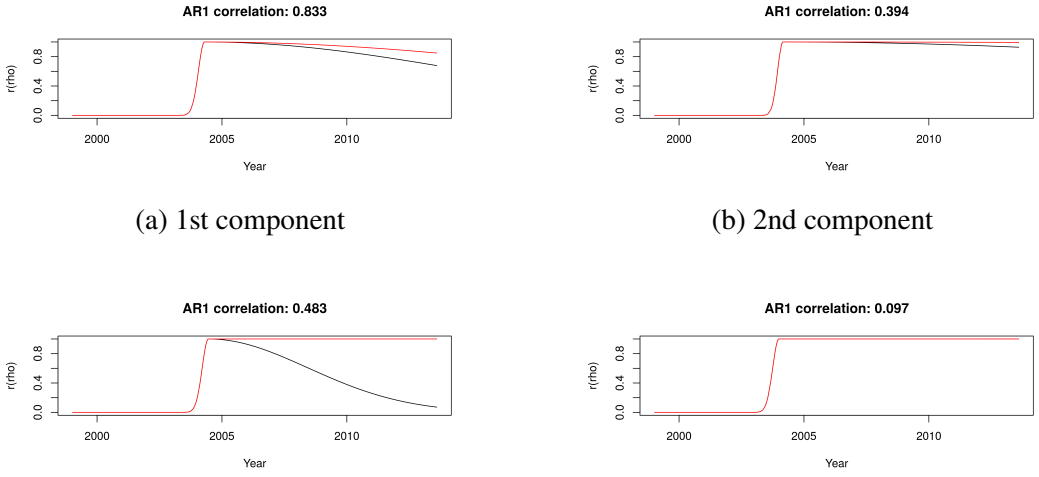
(a) 1st component

(b) 2nd component

(c) 3rd component

(d) 4th component

Fig. 4.16 Fitted intercept and slope effects from Model 7 for the baseline prefelling effect for each functional component. Dotted lines give approximate point-wise 95% confidence intervals.



(a) 1st component

(b) 2nd component

(c) 3rd component

(d) 4th component

Fig. 4.17 Modelled decay of the estimated felling effect (model 7). The decay for the intercept model is shown in black, the red line shows the decay related to the change in slope. The title gives the estimate of the AR1 parameter for each model component.

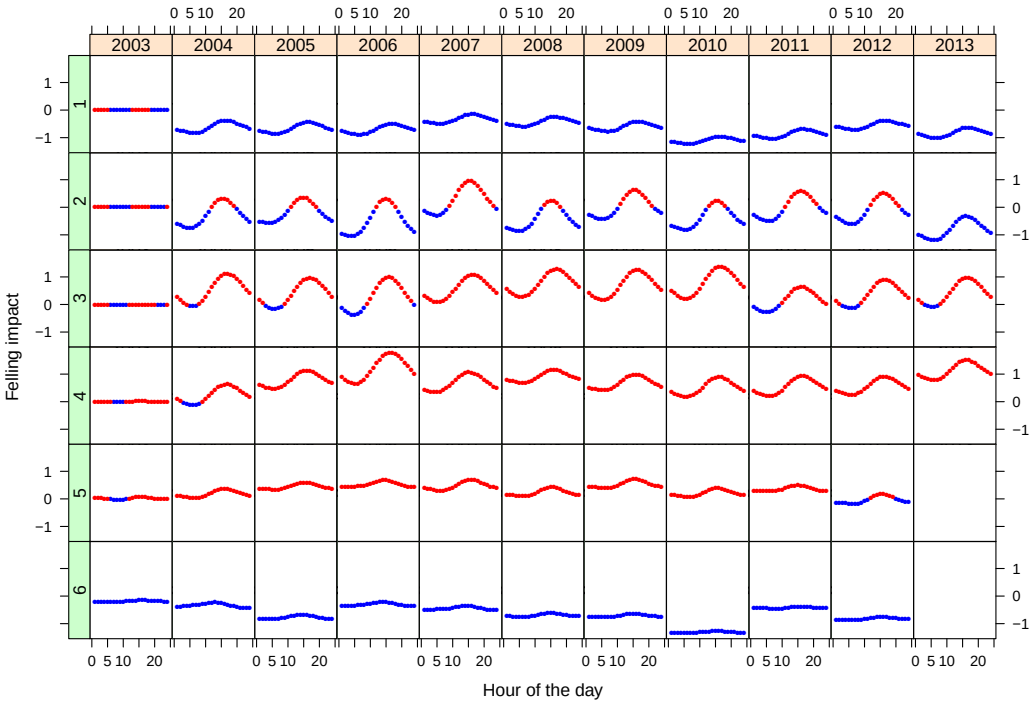


Fig. 4.18 Fitted felling effects for 1: Jan-Feb, 2: Mar-Apr, 3: May-Jun, 4: Jul-Aug, 5:Sep-Oct, 6:Nov-Dec. Positive effects on temperature are distinguished from negative effects by colour: blue - negative, red - positive.

Further aspects of predicting the effects of felling are explored in the following section.

4.5.4 Predicting effects for different temperature regimes

In this section I cover the prediction of temperature from the models fitted in the previous section and present a nice way to interpret the effect of felling in stream temperature.

In order to write an equation that goes from raw temperatures in Burn 2 to predicted raw temperatures in Burn 10, I need to introduce a few new matrices. I will state the equations first, then go through each part. The equation I am aiming to explain is

$$\hat{\boldsymbol{y}} = \boldsymbol{BPZ}(\boldsymbol{\alpha})\hat{\boldsymbol{\gamma}} \tag{4.40}$$

which can be split into parts. $\boldsymbol{Z}(\boldsymbol{\alpha})$ is a 'super' design matrix which replicates the design matrix accross the four copmonents. It is written as a function of $\boldsymbol{\alpha}$, the components of Burn 2 temperature to make explicit that this is where new data for predictions would enter. $\hat{\boldsymbol{\gamma}}$ are the stacked fitted parameters from each model, so that $\hat{\boldsymbol{\beta}} = \boldsymbol{Z}(\boldsymbol{\alpha})\hat{\boldsymbol{\gamma}}$ are the fitted Burn 10 components. $\boldsymbol{P}$ is a permutation matrix to change the ordering of $\hat{\boldsymbol{\beta}}$ from component-wise to day-wise, and finally $\boldsymbol{B}$ is a matrix to convert the components back into daily temperatures. The following paragraphs explain these steps in more detail.

It is easiest to first project the raw temperature observations $\boldsymbol{x}$ onto the specially designed 4-d function space found by the functional PCA, to get the functional principal components $\boldsymbol{\alpha}$

$$\boldsymbol{\alpha} = \boldsymbol{P}^{-1}\left(\boldsymbol{A}'\boldsymbol{A}\right)\boldsymbol{A}'\boldsymbol{x} \tag{4.41}$$

here, the vector $\boldsymbol{x} = (x_{11}, \ldots, x_{1n_1}, x_{21}, \ldots, x_{2n_2}, \ldots, x_{d1}, \ldots, x_{dn_d})'$, where the element x_{ij} denotes the jth temperature observation on the ith day, with $j = 1, \ldots, n_i$ and $i = 1, \ldots, d$. The permutation matrix $\boldsymbol{P}^{-1}$ reorders the output so that $\boldsymbol{\alpha} = (\alpha_{11}, \ldots, \alpha_{d1}, \alpha_{12}, \ldots, \alpha_{d2}, \ldots, \alpha_{14}, \ldots, \alpha_{d4})' = (\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2', \boldsymbol{\alpha}_3', \boldsymbol{\alpha}_4')'$, where the element α_{ik} denotes the kth temperature component on the ith day and $\boldsymbol{\alpha}_k$ denotes the vector of daily kth components. The matrix A has a block diagonal structure where

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_1 & & & \\ & \boldsymbol{A}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{A}_d \end{pmatrix} \tag{4.42}$$

where each $\boldsymbol{A}_j$ is a $n_j \times 4$ matrix

$$\boldsymbol{A}_j = \boldsymbol{Z}_{xj}\boldsymbol{U} \tag{4.43}$$

where $\boldsymbol{Z}_{xj}$ is a $n_j \times 47$ design matrix where each row is the original 47 Fourier basis functions evaluated at the time of the day where the x_{ij} temperature observation took place, and U is a 47×4 linear combination matrix that was found by the FPCA.

Because four models were fitted independently, the design matrices can be combined into a block diagonal matrix and predictions from the four models can be written

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \\ \hat{\boldsymbol{\beta}}_3 \\ \hat{\boldsymbol{\beta}}_4 \end{pmatrix} = \hat{\boldsymbol{\beta}} = \boldsymbol{Z}(\boldsymbol{\alpha})\hat{\boldsymbol{\gamma}} = \begin{pmatrix} \boldsymbol{Z}_1(\boldsymbol{\alpha}_1) & & & \\ & \boldsymbol{Z}_2(\boldsymbol{\alpha}_2) & & \\ & & \boldsymbol{Z}_3(\boldsymbol{\alpha}_3) & \\ & & & \boldsymbol{Z}_4(\boldsymbol{\alpha}_4) \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\gamma}}_1 \\ \hat{\boldsymbol{\gamma}}_2 \\ \hat{\boldsymbol{\gamma}}_3 \\ \hat{\boldsymbol{\gamma}}_4 \end{pmatrix} \tag{4.44}$$

where $\boldsymbol{\beta}$ and related vectors are defined for Burn 10 as $\boldsymbol{\alpha}$ is for Burn 2.

The fitted components can now be converted into temperatures. First the parameter vector needs to be reordered using $\boldsymbol{P}\hat{\boldsymbol{\beta}}$ so that the fitted components can be grouped into d 4-vectors $\boldsymbol{\beta}_j$, one for each day. Then a matrix $\boldsymbol{B}$ is constructed much like (4.42) in which the spline design matrices $\boldsymbol{Z}_{yj}$ can be evaluated wherever a prediction is required in the day: a suitable choice is one per half hour, so that $\boldsymbol{B}_j$ (the equivalent elements of (4.42) but for the matrix $\boldsymbol{B}$) are identical 48×4 matrices.

Consider now, the prediction of a felling effect for a sequence of days, $j = 200, \ldots, 210$, in the study period. The *observed* components at Burn 2 are

$$\boldsymbol{\alpha} = (\alpha_{200,1}, \ldots, \alpha_{210,1}, \alpha_{200,2}, \ldots, \alpha_{210,2}, \ldots, \alpha_{200,4}, \ldots, \alpha_{210,4})' \tag{4.45}$$

This allows us to predict the temperature in Burn 10 for these days via (4.47/cmnotefix). If interest is in isolating the felling effect, this can be done by constructing a design matrix $\boldsymbol{Z}_f(\boldsymbol{\alpha})$ which has all columns not associated with felling set to zero, and so the felling effect on days 200 to 210 is given by

$$\hat{\boldsymbol{y}}_f = \boldsymbol{B}\boldsymbol{P}\boldsymbol{Z}_f(\boldsymbol{\alpha})\hat{\boldsymbol{\gamma}} \tag{4.46}$$

A nice way to summarise the total thermal regime is via cumulative temperature plots (see Malcolm et al., 2008a). The plots are effectively the cumulative distribution functions for temperature, usually over a month or a season. These can be useful in showing the amount of time spent above or below a particular temperature and so allow an ecologist to assess if unfavourable stream temperatures have occurred.

Cumulative temperature curves (CTCs) can also be used to assess the impact of felling by comparison with the predicted baseline temperature. As the variance of the model parameters are available it is possible to include uncertainty in the CTCs.

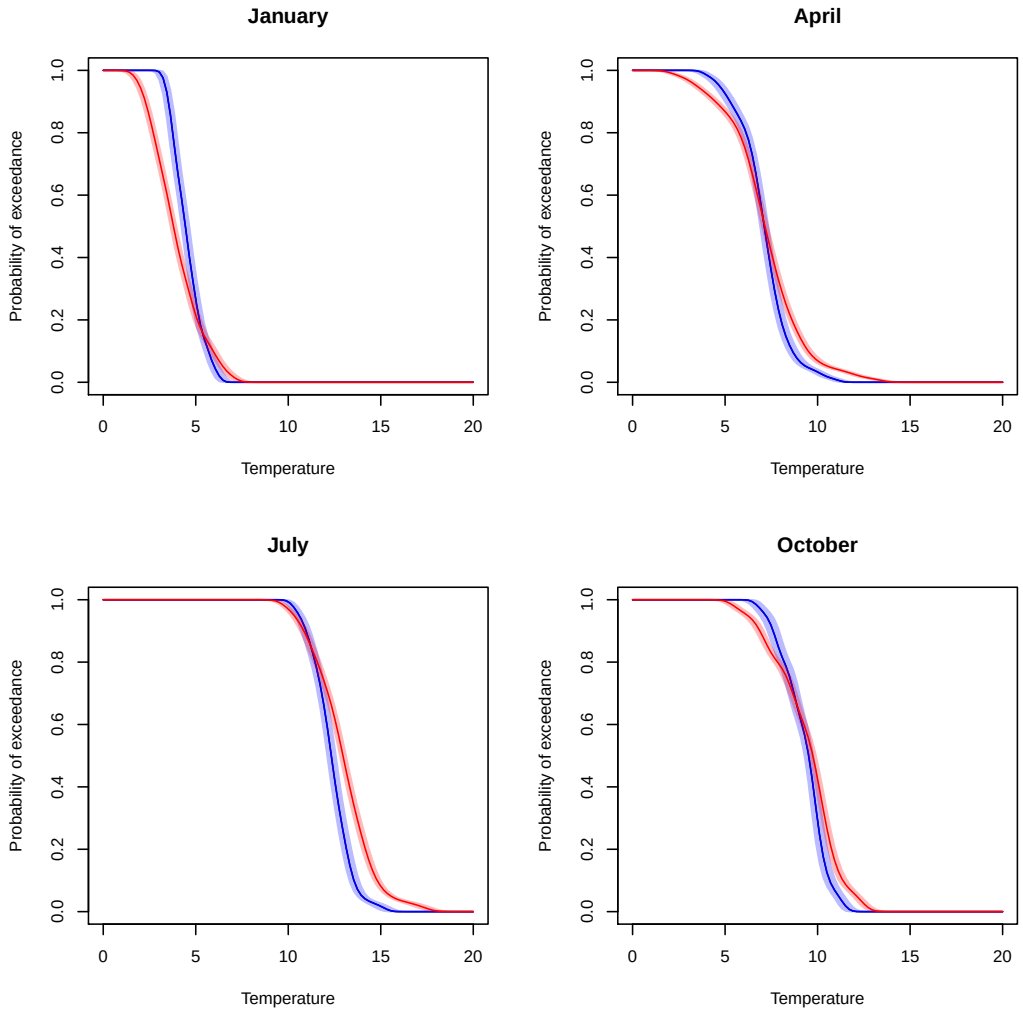


Fig. 4.19 Predicted cumulative temperature duration curves (CTCs) for winter (January), spring (April), summer (July) and autumn (October). Plots show the proportion of time above the temperature given on the x-axis. The red line denotes the CTC under the impact of felling, while the blue line denotes the CTCs with the felling effect removed. Uncertainty was simulated by assumiing the model parameters had a multivariate normal distribution with variance as estimated by mgcv. A CTCs was simulated for each resample resulting in an estimated 95% uncertainty interval for the monthly CTCs given by the shaded polygons.

The effect of felling given month was investigated by simulating baseline temperatures and the associated impacted temperatures at Burn 10 and calculating the CTC for the baseline and impacted temperatures for each simulation. For example, to calculate the CTC for July, daily temperature curves (functional component values) for every July day across all years 1000 impacted and non-impacted burn 10 temperature curves were simulated. The simulated monthly temperature curves were then summarised accross year and day into a single CTC for each simulation, resulting in 1000 CTCs. Figure 4.19 shows the median, 2.5th and 97.5th percentiles of the simulated CTCs over the period 2004-2013 for the months, January, April, July and October.

It can be seen from the CTC plots that the modelled effect of felling is to increase the risk of attaining high and low temperatures, in general. The overall range of temperatures experienced by Burn 10 is predicted to have increased under felling.

4.6 Discussion

This chapter sought to extend the paired catchment method used by Gomi et al. (2006) to explore the effect of tree felling on stream temperature. The developments were the use of cyclic smoothers to generalise beyond the use of a sine and cosine, the direct modelling of a felling impact, and the expansion from modelling daily summaries to modelling daily temperature curves. The results in this chapter show the benefits of increased complexity in the modelling procedure. Although, the increased complexity did not provide big improvements in fit, the increased structure in the model did allow for a more comprehensive investigation of the impacts of and magnitude of the felling that is not possible when fitting the models found in e.g. Gomi et al. (2006).

The core modelling strategy was to follow and extend the approach of (Gomi et al., 2006), in which an unimpacted stream is treated as a control site and the impacted site regressed against the control to measure differences. A potential weakness in this approach is that both stream temperatures are observed with error, and hence treating the control site as error free could introduce a bias that is related to the amount of observation error in the control site. Fortunately, due to the calibration procedures conducted prior to deploymeny, the temperature data loggers can measure temperature to high precision (to within 0.02 degrees) hence treating the control burn as error free is not likely to impact on predctions that in the order of 0.5 to 1 .

4.6.1 Cyclic smoothers

The simple model used by Gomi et al. (2006) is attractive due to its simplicity. However, a simple sinusoidal model is not appropriate for all applications, for example in high latitudes winters tend to be longer or the impact of melt water can induce asymmetric cyclic trends. So, from a general point of view, the use of cyclic smoothers is a sensible extension. Furthermore, since the ultimate goal was to apply this model to the components of diel temperature curves, this generality was required to deal with potentially non-sinusoidal relationships.

To this end, the use of splines with the harmonic penalty performed well. This is not to say that the use of simpler cyclic cubic splines would not also be appropriate. However, exploring the use of more general ways to penalise through the use of the differential equations (Section 4.2), served to highlight the potential of more general differential equations for finding penalties that incorporate dynamics specific to modelling temperature variation, e.g., temperature trends depend on latitude due to the spin of the earth. Furthermore, the connection between differential equations (and linear differential operators) and GMRF penalties, brings the dynamics of the

quadratic penalty to life. There are a number of recent examples where splines have been developed as solutions to differential equations. Lindgren et al. (2011) in a seminal paper showed there was a link between continuous spatial random fields such as the Matern process and GMRFs, and Yue et al. (2014) use this idea to develop univariate adaptive splines by finding a GMRF that is the solution to a differential equation.

The harmonic penalty could be used to include potential year to year cyclic variability by defining a spline bases over a number of years, while the penalty is defined with a period of 365 days. One may consider a similar model that fits a cyclic smoother, repeated every year, with an additional freely varying smoother to deal with the year to year variability. But what happens in this case is that the free smoother absorbs all the variability and the cyclic smoother becomes unidentifiable. So, the harmonic penalty appears to have a particular advantage in this regard. However, in this application, the use of such a smoother may endanger the estimability of a felling effect, as the felling impact trend could be absorbed into the year to year variability.

There is, however, the potential to define a particular GMRF that will approximately penalise towards a given function, this would be achieved by finding a suitable differential equation to describe the given function. The technique known as partial differential analysis (Ramsay and Silverman, 2005) potentially allows for this, but the details of how this would be applied are not yet clear. If such a procedure could be incorporated into the model fitting process, it may be possible to penalise towards an estimated baseline seasonal effect. This would seem to be a way to both find a baseline seasonal effect while allowing smooth cyclical year to year variability.

4.6.2 Modelling the impact of felling

The non-linear addition of a decaying felling effect, in the example presented here, did not have a large impact on the overall fit, however, the research question was to design a model that could investigate the magnitude of felling, and apply it to a suitable dataset. Although the felling effect was not large, it did result in demonstrable differences in the cumulative temperature duration curves, for example. Furthermore, the fact that the felling effect was not large, but nonetheless detectable is still of scientific interest. The small impact of felling was completely due to the nature of the felling event. The felling event *was* substantial, but did not completely expose the river and left a *buffer* zone to mitigate the impact. This meant that the impact was likely due to reduced shading on a much larger scale (for instance, the removal of a large bank of distant trees) which would require complete regrowth of the forest before recovery took place.

On the other hand, a preliminary application of the model developed here has been applied to another felling event, in which trees were removed right up to the edge of the river. This event had a large impact on temperatures, followed by a substantial decay, but not full recovery. The difference here is that the rapid recovery is likely to be due to fast growing bush and brush on the banks of the river which quickly reduce the extreme effects of the felling. Unfortunately, there are several issues with this dataset that mean it was not possible to further develop the analysis as fully as the one in this thesis. Firstly, there was only one year of pre-felling data, and hence the conclusion on the size of the felling effect is potentially confounded with year to year environmental variability. And secondly, the felling event temperature loggers used had a non-negligible reporting resolution such that the observation error in the control burn would need to be explicitly taken into account, and thus requires a change in methodology.

The model that allowed the onset of felling to be estimated from the data had separate decay rates for the intercept and slope terms. This leaves open the possibility to further improve the felling model through increased flexibility. One option is the use of monotonic splines to model the decay, or to devise more parametric forms. However, the fitting procedure became much less efficient as the number of parameters increased in the outer iteration. The problem is that the GCV surface became very flat with more complex models in the example presented here, presumably because the evidence for a decay in felling was only slight. This would not be such a problem if the model fitting in the inner loop was fast, but it was not. This speed of fitting was traced to the application of the weight matrix for the AR1 process where the $n \times n$ matrix of weights multiplies the model design matrix. The code used to fit this model in the mgcv package used standard matrix objects from the base package. But since the weight matrix was sparse (AR1 weight matrices have non-zeros on the two main diagonals), sparse matrix methods could be used. Therefore, the use of the Matrix package to improve fitting speed could allow the investigation of more complex felling models.

The development of the felling model is therefore very promising, as it allows quantification of both recovery rate and reductions in temperature that can be attributed to different felling activities, and how this changes with time after felling. There are also promising avenues for further development.

Trends in residuals

Several attempts were made to address the problems with trends in the residuals 4.13. An abrupt step change for the felling model was tried, and models were checked for convergence by trying various starting values, and altering the parameter values manually to see if the residual problem could be fixed. Unfortunately, none of these

exercises were fruitful. The main problem is the existence of long term trends over years, and sudden, large, short term changes in the residuals such as those around the start of 2010. These trends could not be fitted given the model design. During model development long term smoothers were investigated and it would be possible to add smoothers over years to the current models to deal with the residual patterns, but this was decided against because any such smoother would be confounded with both the assessment of the felling effect and its decay over time. In addition, any trends fitted in this way have no interpretation. A good solution would be to identify a plausible process or processes behind the trends and find a covariate to introduce into the model to control for it.

Following discussions with staff at the Marine Scotland Science Freshwater Laboratory several possible processes were identified that could lead to such variability. There are many covariates that have the potential to influence river temperature. Physical characteristics such as channel orientation, elevation and the steepness of valley sides, can have impacts that operate on a daily level, (channel orientation and valley steepness) and on a seasonal level (elevation). The physical features of a temperature monitoring site can induce predictable seasonal differences between sites and is the subject of ongoing research (Garner et al., 2015; Hannah et al., 2008; Jackson et al., 2015; Malcolm et al., 2008b). The model was designed to account for such differences via the seasonal smoothing terms. But the main driver of variation in concurrent differences in temperature across sites is likely to be climatological. This is the likely cause of the residual variability in the model. A major driver of stream temperature is discharge (the volume of water flowing). For a given river, higher discharge means faster flow and more water. Both of these things make it harder for the sun to warm it, and hence tends to reduce differences in temperature between sites. The major cause of high discharge is precipitation, so wet weather will reduce the differences between sites. On the contrary, dry weather results in low

discharge shallower rivers, so therefore easier to warm and hence differences between sites will be enhanced. Furthermore these climatological impacts can interact with the physical nature of a site: precipitation at higher altitudes can be colder than at lower altitudes, and steeper valley sides transfer water more quickly to the river than sites with flatter profiles. In summary, it is clear that both climate variables, such as river discharge or rainfall metrics, and the physical characteristics of the sites would be excellent covariates to include in the models presented here, and could very well improve the residual trends exhibited. Unfortunately, these data are either not in a suitable form to be readily used or are the basis of ongoing doctoral research and are not available at the present time.

4.6.3 Daily temperature curves

The extension from using daily summaries, such as the daily mean, or daily range, to modelling the daily temperature curves gives greater insight into the effects of felling on stream temperature. Temperature duration curves are used primarily in hydrology, but have found uses in ecology to demonstrate the distribution of temperatures that stream inhabitants face. It is not simply the mean temperature that is important but the full range of temperatures experienced. Hence the modelling approach developed which provides a much fuller picture about the effect of felling on stream temperature, has greater value as a tool to describe changes in environment due to the removal and addition of trees.

The model could be extended to consider models where components could affect each other in a multivariate regression. If only linear relationships (no splines) were considered this model would closely resemble canonical correspondence analysis (Ter Braak, 1986), which could be conducted with standard software. However, the use of splines takes the analysis beyond the capability of commonly available tools.

In the multivariate setting, further extensions would be to allow the AR1 processes to be correlated across the components. All of these extensions would require more parameters to be estimated in the outer optimisation, however, and so at present these ideas are not practical.

In the estimation of the error in the cumulative temperature duration curves the error comes from the linear and smooth effects only, and there is no uncertainty due to the estimation of the AR1 parameter, the felling decay or the smoothing parameters estimated in mgcv. In order to incorporate these estimation errors it one could optimise jointly over all model parameters and get the joint Hessian matrix, but this would require the specification of a full joint model. This was not possible with the current approach and is a topic for further investigation.

4.6.4 General remarks

I found in this application that the use of GMRF models in large applications became computationally limiting. The fitting procedure used is similar to other two stage methods such as the use of Gibbs step block updates in Markov chain Monte Carlo methods for inference with large GMRF models (Rue and Held, 2005). In this method the inner optimisation (block update) works by conditioning on parameters that make the model linear or of standard form. The inner step, being of standard form should in theory be fast, allowing an outer optimisation of the non-linear / non-standard parameters. However, these methods also suffer from the curse of dimensionality. The solution used in these applications is to use efficient numerical libraries for manipulation and storage of large sparse matrices. These techniques could improve fitting time in the models developed here.

Finally, I hope this chapter serves to highlight the wide flexibility that can be achieved using linear models. Although the AR1 process and felling decay model introduced non-linearity, there was considerable flexibility in the linear part of the model. I also hope that this chapter raised awareness of the central role that GMRFs, through their connection to linear differential equations, have in penalised smoothing problems. A good example of this is the RW models and the harmonic model. It is enlightening to note that GMRFs are the very same matrices used to model discretised versions of mechanical systems (Strang, 2009). For example, the RW1 model is the same model used for a line of masses linked by springs, and the cyclic RW1 model is that for a *circle* of masses linked by springs; in these examples the stiffness of the springs is proportional to the smoothing parameter. These same ideas extend to modelling the structural stability of bridges and buildings; all the information on the dynamics of these physical systems is contained in the equivalent penalty matrix.

5

General discussion and future directions

Throughout this thesis I have tried to make clear three important points:

1. The central role that GMRFs play in quadratic penalties
2. The variety and flexibility of quadratically penalised linear models (i.e. STAR models)
3. The usefulness of STAR models for freshwater management and wider ecological applications.

5.1 Introduction

GMRF models are necessarily better than a simple GAM. GAMs are highly flexible and efficient modelling tool, capable of fitting cyclic, multidimensional surfaces. However, there are special circumstances where the use of GMRFs is warranted,

such as a regional smoother or a correlated smoother, or where a specific penalty is required such as the combination of penalties that results in a sinusoid for modelling periodic time series. It is easy to construct and modify these GMRFs from first principles as I have shown in this thesis, and after the construction of the penalty, their use is, in practice, no more complex than working with a standard GAM model. The combination of ease of use and the straightforward creation of complex penalties make GMRF models worthwhile learning about and applying, and in my opinion elevate them above simple GAMs.

One of the most enlightening things I have learnt about linear models during the writing of this thesis is the notion of vector spaces. We typically think of data $\boldsymbol{y}$ (with n observations) as living in a Cartesian coordinate system: a 2-D data point can be plotted on an x-y plot, and a 3-D point in an x-y-z cube. In terms of a vector space, the basis vectors for data are given by the $n \times n$ identity matrix. A design matrix $\boldsymbol{Z}$ for a spline formed from p coefficients $\boldsymbol{\gamma}$, is a link between a function space and the data space. The coordinate system of the function space is defined by basis vectors that are the spline components. Moving between the two spaces is simple: from the function space to the data space it is $\boldsymbol{X\gamma}$, and from the data space to the function space it is $(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$, i.e. the least squares regression estimate.

But what is a good spline basis? This is where GMRFs shine in my opinion. GMRFs live in the data space. But they are intimately linked with function spaces as demonstrated by the eigen transformations used for reduced rank approximation. In Section 4, I showed that it is possible to find a linear transformation of a GMRF structure matrix $\boldsymbol{Q}$ that produced a diagonal penalty. This procedure has moved the GMRF from the data space to a function space, so that a reduced set of basis functions could be selected to reduce the overall size of the model. The insight here is that we can use GMRFs to define function spaces with specific properties.

The properties of various GMRF models were shown in Chapter 2. All the GMRF models presented here arise from various types of linear combinations of standard normal variables. Consider the general GMRF density

$$f(\boldsymbol{\gamma}|\boldsymbol{Q}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{Q}\boldsymbol{\gamma}\right) \tag{5.1}$$

Writing $\boldsymbol{Q}$ in terms of a linear combination $\boldsymbol{L}$ we get

$$f(\boldsymbol{\gamma}|\boldsymbol{Q}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{L}'\boldsymbol{L}\gamma\right) \tag{5.2}$$

$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{L}\boldsymbol{\gamma})'(\boldsymbol{L}\boldsymbol{\gamma})\right) \tag{5.3}$$

so $\boldsymbol{L}\boldsymbol{\gamma}$ is multivariate normal with identity variance matrix. The matrix $\boldsymbol{L}$ can be thought of as being at the heart of GMRFs. The first order random walk model is $\boldsymbol{L} = \boldsymbol{D}$, where $\boldsymbol{D}$ computes the differences between time steps. More generally $\boldsymbol{D}$ could compute the differences between values at nodes in a network, for example, to define a smoother on a river network. Furthermore, If model fitting is extended to allow the matrix $\boldsymbol{L}$ to depend on parameters, then models such as the first order autoregressive process (AR1) can be defined as GMRFs. In this case, $\boldsymbol{L}$ can be constructed directly from the definition of an AR1 (see Section 4.4.2).

It is possible to take this further. The difference matrix $\boldsymbol{D}$ is really doing calculus: it is approximating the derivative. This means that GMRFs, and hence penalties, can be constructed based on differential equations. In section 4.4.1, I used the harmonic penalty which arises from the equations of simple harmonic motion:

$$f'' + \omega^2 f = 0 \tag{5.4}$$

Not presented in this thesis, but mentioned in Section 2.2.1.2 are orthogonal polynomials. Curiously, classical orthogonal polynomials are solutions to differential

equations of the form

$$A(x)f'' + B(x)f' + \omega f = 0 \tag{5.5}$$

where $A(x)$ is at most a quadratic and $B(x)$ is at most a linear polynomial. So, potentially even orthogonal polynomials can be thought of as GMRFs. More generally, any differential equation with known coefficients could be used to construct a penalty. A nice example of this idea is Wood et al. (2008). The point is that by defining GMRFs based on a linear difference matrix provides a link from the comprehensible data space to a wide variety of function spaces.

5.2 STAR in practice

In this thesis the models were fitted in R (R Core Team, 2015) using the optimisation routines in mgcv (Wood, 2006). In order to fit GMRF models, it was necessary to alter the code in the mgcv package. This allowed the full range of GMRF models to be fitted within, what is to many, a familiar modelling environment. The benefits of extending mgcv is not restricted to parameter estimation. Model prediction, plotting, and other model summaries developed for mgcv all become automatically available for GMRF based STAR models.

5.2.1 Salmon density

The usefulness of STAR for modelling salmon fry abundance was clearly demonstrated. In addition the use of a two stage model was hugely beneficial as it allowed a range of models to be explored for both capture probability and density. Although the two stage model does not make optimal use of the available information in the way that a

joint model of capture probability and density would, it does allow a wide range of models to be explored in a short time frame.

In addition to writing the extensions to mgcv, additional software was produced specifically for the capture probability likelihood. This was optimised using a recently developed library which can perform automatic differentiation (RSTAN, Stan Development Team, 2014a) and therefore runs very quickly. The formula interface for mgcv was used, and so the same notation for defining STAR models was used to define the reduced rank unpenalised models for capture probability. This allowed the incorporation of spatial effects into the design matrix for capture probability.

To summarise the advantages of the approach, I would emphasis the ease of model specification and the robust fitting due to the `ef` package developed as part of this thesis utilising the tools available in mgcv and STAN. However, it would be good to more formally link the two stages of model fitting which would allow the possibility of feedback from the density model to the capture probability model when model selection is taking place. The bootstrapping approach used here has a clear advantage over a simple two stage approach. Further improvements are to include correlated random effects to improve the joint modelling of lifestages and species. The GMRFs for this type of modelling exist and are documented in the Methods chapter, but fitting such models will require the use of more specialised software such as INLA (Rue et al., 2009).

There are emerging methods for automatic function selection in STAR models (Scheipl et al., 2013a). This was attempted in part by including both a random effect and a slope term for year in the density model. The idea of Scheipl et al. (2013a) is to fit several functional possibilities for a relationship in the one model. For example, for a regional effect, one might fit a random effect and a spatial effect and the one that

is most supported by the data would be retained while the remaining unsupported forms would be penalised out the model. This idea could be used in developing standard analyses for density modelling for use in management. The use of a standard model space reduces subjective input, placing more importance on the signals in the data than on what may be expected by the researcher. There is still a requirement for subjective input in choosing a set of covariates and possible functional forms. This is crucial in order to provide a sensible model space to work in.

5.2.2 Stream temperature

There were several difficulties when constructing a suitable model for this section. The first was that the data set is extensive and there was a substantial amount of data cleaning to be done. The next issue was how to model the data, the problem being that there were so many options. There were more false starts and changes in approach than I care to remember; even now I am thinking of alternative ways to approach the problem. But after several attempts, it became clear that the model should have the following properties:

- The model needed to be flexible, but not too flexible
- The model needed to quantify the effect of felling and allow it to decay over time
- The model should allow for year to year variability.

In the end, I chose to use a simple construct, and add complexity piece by piece. So the first model design is that felling effects occur in addition to a base line relationship between the control and the impacted site. In order to allow for a flexible seasonal relationship, while ensuring that the felling effect was identifiable, I forced a common

seasonal pattern for each year. The decay of the felling impact was included as a multiplicative term, but it was understood that, conditional on the decay rate, the model was a STAR model and hence could be fitted iteratively with mgcv, treating the decay rate as an external parameter to be optimised over. Without additional information, however, it proved to be difficult to allow for year to year variability, while keeping the underlying model continuous through time and not sacrificing identifiably of the felling effect. So, in the end, all seasonal variation was forced into the residuals by way of a restrictive felling decay model. The main reason that a spline model for the felling decay was not chosen was so that year to year variability was kept separate from the estimate of felling and recovery. The consequence, however, was very likely an over estimate in the residual variation, and hence too large confidence intervals on model estimates.

The issues raised above aside, I still consider the model a useful contribution. The use of functional principal components (FPCA, Ramsay and Silverman, 2005; Silverman, 1996) to find a set of 'good' basis functions to model daily variation was a great success, and allowed the complete thermal regime to be modelled. The consequence is a model that can make predictions about the consequences to the full thermal environment that fish inhabit which is of greater use to ecologists and managers.

It also provides an alternative, perhaps, to the eigen decomposition approach to finding reduced rank basis functions. The FPCA approach finds a set of reduced bases that are most appropriate for the data, in much the same sense that conducting a PCA of a range of covariates finds a reduced set of covariates.

5.3 Future directions

The ability to use the interface provided by mgcv made the specification of models very efficient, and for small data sets (< 1000 observations) model fitting was very fast. However as the dataset and model increased in size, computation became markedly slow, for example, model 7 fitted in Chapter 4 took around 1 hour to complete the outer optimisation. The main reason for this is that each step in the inner optimisation to find the appropriate smoothing parameters conditional on the non-linear felling and AR1 parameters, was slow, taking up to 2 minutes. There are two potential ways to improve this: 1) optimise directly over the smoothing parameters and the non-linear parameters in one step, or 2) improve the speed of estimation of smoothing parameters.

Optimising directly over the smoothing parameters and the non-linear parameters is appealing, but would be difficult to make general. Although, you could argue that the algorithm is tailored to the application as it stands, so why not invest time in improving it? Well, this would depend on the frequency that the particular method would be used.

The second option is to improve the speed of smoothing parameter estimation in general. This is the more plausible option. The slow fitting speed was traced to the use of linear algebra techniques based on the assumption of dense matrices. Typically GMRFs are sparse matrices, containing a large number of zero entries. There are a variety of sparse matrix methods available and in R they can be accessed through the package Matrix. Methods such as INLA (Rue et al., 2009) which make wide use of GMRF models, rely on sparse matrix methods (see Rue and Held, 2005). So, in order to use mgcv as a tool for fitting GMRF models, it would be necessary to convert existing mgcv code to use the Matrix package (Bates and Maechler, 2015).

There are several improvements that could be made to the modelling approach presented in Chapter 3. I will cover what I consider to be the most important.

The electrofishing capture probability model likelihood is not an exponential family model, and so cannot be fitted using mgvc. An interesting future development would be the use of penalised likelihood estimation (Green, 1987). The main problem with this is the estimation of the effective degrees of freedom (EDF) of the model. The notion of effective degrees of freedom is an interesting one and is discussed by Mallows (1973), Efron (1986), Buja et al. (1989) and more recently Janson et al. (2013). A geometric interpretation of EDF is that it corresponds to the amount of movement in the model space induced by movement in the data space. These *movements* are related to the residual error and prediction error (Efron, 1986; Mallows, 1973). Briefly, for linear unpenalised models the EDF is the dimension of the model space, but for penalised models the EDF is generally smaller than the model space because the fit is constrained, and hence for a given movement in the data space, the corresponding parameter estimates do not move as freely as in the unpenalised case. For curved model surfaces (non-linear models and generalised linear models) the EDF can be more or less than the dimension of the model space, and this depends on the curvature of the model space: convex surfaces give smaller degrees of freedom, and concave surfaces give more. The last two conclusions can be derived from the geometry. These facts mean that quantities such as AIC and GCV are approximate for non-linear models (including for GLMs, see Claeskens and Hjort, 2008). In order to gain a suitable approximation a bootstrap procedure has been suggested for logistic regression by Efron (1986)and more widely by Claeskens and Hjort (2008) for GLMs. These methods could be adapted to estimate the EDF for the capture probability model for a given set of smoothing parameters, and hence allow the capture probability model to be fitted with penalised model components. This has not been tested, however, and because each step in the optimisation would require

a bootstrap procedure to estimate the EDF, model fitting may take some time. On the other hand, since smoothing parameter estimation can be thought of as a form of automatic model selection (Scheipl et al., 2013a), then the need for a step-wise procedure would be reduced.

The methods in chapter 4 serve as a good starting point for further investigation into the use of the functional representation of data (Ramsay and Silverman, 2005) to model diel temperature variation. I will briefly discuss one possible future development area that I think is promising.

A limitation in the model presented for river temperature was the lack of an approach to deal with year to year variability in the seasonal trend. The seasonal model presented in Section 2.2.1.5 is a potential candidate. It would not be possible to use the full GMRF model, but investigation of the use of a reduced rank version would be very interesting.

An alternative (possibly overcomplicated) approach is to use a procedure called principal differential analysis (Ramsay and Silverman, 2005, Chapter 19). Principal differential analysis (PDA) estimates a system of n linear differential equations that fit the provided data. The idea would be to estimate a common seasonal trend and derive the principal differential components of the trend. The differential components could then be used to construct a linear differential operator (L) to define a GMRF which would penalise towards this common trend. It would then be possible to allow for year to year variability about this seasonal trend. The details of such an approach have not been fully considered, but it is likely that such a fitting procedure would have to consist of two steps: 1) fit a seasonal trend and find the appropriate penalty associated with it, and 2) estimate the model parameters conditional on the penalty.

5.4 Closing remarks

The flexibility in STAR models comes from the beauty and variety of linear algebra (Strang, 2009). When I first started studying linear models in statistics, the use of splines seemed like a cheat's way to include non-linearity into linear models. However, understanding that through linear transformations data could be translated into a new space in which the coordinate systems are smooth functions such as the Fourier sines and cosines, and that it was possible to find such new basis functions through eigenvalue decompositions, exposed the fact that the nominal classification of a model as simply linear undersells its potential flexibility.

References

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998. URL http://link.springer.com/chapter/10.1007/978-1-4612-1694-0_15.

J. D Armstrong, P. S Kemp, G. J. A Kennedy, M Ladle, and N. J Milner. Habitat requirements of Atlantic salmon and brown trout in rivers and streams. *Fisheries Research*, 62(2):143–170, May 2003. ISSN 0165-7836. doi: 10.1016/S0165-7836(02)00160-1. URL http://www.sciencedirect.com/science/article/pii/S0165783602001601.

Douglas Bates and Martin Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*. 2015. URL http://CRAN.R-project.org/package=Matrix. R package version 1.2-0.

C. Belitz, A. Brezger, N. Klein, T. Kneib, S. Lang, and N. Umlauf. *Bayesx. Software for Bayesian inference in structured additive regression models, Version 3.0. 1*. 2015.

Robert L. Beschta and R. Lynn Taylor. Stream Temperature Increases and Land Use in a Forested Oregon Watershed1. *JAWRA Journal of the American Water Resources Association*, 24(1):19–25, February 1988. ISSN 1752-1688. doi: 10.1111/j.1752-1688.1988.tb00875.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1752-1688.1988.tb00875.x/abstract.

Marta Blangiardo, Michela Cameletti, Gianluca Baio, and Håvard Rue. Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology*, 4:33–49, March 2013. ISSN 1877-5845. doi: 10.1016/j.sste.2012.12.001. URL http://www.sciencedirect.com/science/article/pii/S1877584512000846.

T. Bohlin, J. Pettersson, and E. Degerman. Population density of migratory and resident brown trout (Salmo trutta) in relation to altitude: evidence for a migration cost. *Journal of Animal Ecology*, 70(1):112–121, 2001. ISSN 1365-2656. doi: 10.1111/j.1365-2656.2001.00466.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2656.2001.00466.x/abstract.

Reidar Borgstrom and Oystein Skaala. Size-dependent catchability of brown trout and Atlantic salmon parr by electrofishing in a low conductivity stream. *Nordic Journal of Freshwater Research*, 68:14–21, 1993.

George W. Brown and James T. Krygier. Effects of Clear-Cutting on Stream Temperature. *Water Resources Research*, 6(4):1133–1139, August 1970. ISSN 1944-7973. doi: 10.1029/WR006i004p01133. URL http://onlinelibrary.wiley.com/doi/10.1029/WR006i004p01133/abstract.

L. E. Brown, L. Cooper, J. Holden, and S. J. Ramchunder. A comparison of stream water temperature regimes from open and afforested moorland, Yorkshire Dales, northern England. *Hydrological Processes*, 24(22):3206–3218, June 2010. ISSN 08856087. doi: 10.1002/hyp.7746. URL http://doi.wiley.com/10.1002/hyp.7746.

S. T. Buckland, K. P. Burnham, and N. H. Augustin. Model Selection: An Integral Part of Inference. *Biometrics*, 53(2):603, June 1997. ISSN 0006341X. doi: 10.2307/2533961. URL http://www.jstor.org/stable/2533961?origin=crossref.

Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510, 1989. URL http://www.jstor.org/stable/2241560.

N. G. Cadigan. Fitting a non-parametric stock-recruitment model in R that is useful for deriving MSY reference points and accounting for model uncertainty. *ICES Journal of Marine Science*, 70(1):56–67, January 2013. ISSN 1054-3139, 1095-9289. doi: 10.1093/icesjms/fss183. URL http://icesjms.oxfordjournals.org/cgi/doi/10.1093/icesjms/fss183.

Frank Louis Carle and Mike R. Strub. A New Method for Estimating Population Size from Removal Data. *Biometrics*, 34(4):621–630, December 1978. ISSN 0006-341X. doi: 10.2307/2530381. URL http://www.jstor.org/stable/2530381.

Gerda Claeskens and Nils Lid Hjort. *Model selection and model averaging*, volume 330. Cambridge University Press Cambridge, 2008. URL http://www.langtoninfo.com/web_content/9780521852258_frontmatter.pdf.

Noel Cressie, Jesse Frey, Bronwyn Harch, and Mick Smith. Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(2):127–150, June 2006. ISSN 1085-7117, 1537-2693. doi: 10.1198/108571106X110649. URL http://link.springer.com/article/10.1198/108571106X110649.

Noel Cressie, Catherine A. Calder, James S. Clark, Jay M. Ver Hoef, and Christopher K. Wikle. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, 19 (3):553–570, March 2009. ISSN 1051-0761. doi: 10.1890/07-0744.1. URL http://www.esajournals.org/doi/abs/10.1890/07-0744.1.

D. B. DeLury. On the Estimation of Biological Populations. *Biometrics*, 3(4): 145–167, December 1947. ISSN 0006-341X. doi: 10.2307/3001390. URL http://www.jstor.org/stable/3001390.

Peter Diggle and Paulo J Ribeiro. *Model-based geostatistics*. Springer, New York, NY, 2007. ISBN 978-0-387-32907-9. URL http://public.eblib.com/choice/publicfullrecord.aspx?p=302148.

Bradley Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470, 1986. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478291.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Paul HC Eilers and Brian D. Marx. Flexible smoothing with B-splines and penalties. *Statistical science*, pages 89–102, 1996. URL http://www.jstor.org/stable/2246049.

Ludwig Fahrmeir, Thomas Kneib, and Stefan Lang. Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, 14(3): 731–762, 2004. URL http://www.uibk.ac.at/statistics/personal/lang/publications/fahrmeir_kneib_lang_statsinica2004.pdf.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-34332-2, 978-3-642-34333-9. URL http://link.springer.com/10.1007/978-3-642-34333-9.

Kurt D. Fausch, Clifford L. Hawkes, and Mit G. Parsons. Models that predict standing crop of stream fish from habitat variables: 1950-85. Gen. Tech. Rep.

PNW-GTR-213, Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, 1988. URL http://www.treesearch.fs.fed.us/pubs/8730.

M. C. Feller. Effects of Clearcutting and Slashburning on Stream Temperature in Southwestern British Columbia1. *JAWRA Journal of the American Water Resources Association*, 17(5):863–867, October 1981. ISSN 1752-1688. doi: 10.1111/j.1752-1688.1981.tb01309.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1752-1688.1981.tb01309.x/abstract.

Jason B. Fellman, Sonia Nagorski, Sanjay Pyare, Andrew W. Vermilyea, Durelle Scott, and Eran Hood. Stream temperature response to variable glacier coverage in coastal watersheds of Southeast Alaska: STREAM TEMPERATURE RESPONSE TO VARIABLE GLACIER COVERAGE. *Hydrological Processes*, 28(4):2062–2073, February 2014. ISSN 08856087. doi: 10.1002/hyp.9742. URL http://doi.wiley.com/10.1002/hyp.9742.

David A. Fournier, Hans J. Skaug, Johnoel Ancheta, James Ianelli, Arni Magnusson, Mark N. Maunder, Anders Nielsen, and John Sibert. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27(2):233–249, April 2012. ISSN 1055-6788. doi: 10.1080/10556788.2011.597854. URL http://dx.doi.org/10.1080/10556788.2011.597854.

Grace Garner, Iain A Malcolm, Jonathan P Sadler, Colin P Millar, and David M Hannah. Inter-annual variability in the effects of riparian woodland on microclimate, energy exchanges and water temperature of an upland scottish stream. *Hydrological Processes*, 29(6):1080–1095, 2015.

Jason D. Godfrey. Site condition monitoring of Atlantic salmon SACs. Technical report, SFCC, 2005.

Takashi Gomi, R. Dan Moore, and Amod S. Dhakal. Headwater stream temperature response to clear-cut harvesting with different riparian treatments, coastal British Columbia, Canada: FOREST HARVESTING EFFECTS ON STREAM TEMPERATURE. *Water Resources Research*, 42(8):n/a–n/a, August 2006. ISSN 00431397. doi: 10.1029/2005WR004162. URL http://doi.wiley.com/10.1029/2005WR004162.

Peter J. Green. Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, pages 245–259, 1987. URL http://www.jstor.org/stable/1403404.

S.M. Guenther, R.D. Moore, and T. Gomi. Riparian microclimate and evaporation from a coastal headwater stream, and their response to partial-retention forest harvesting. *Agricultural and Forest Meteorology*, 164:1–9, October 2012. ISSN 01681923. doi: 10.1016/j.agrformet.2012.05.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0168192312001724.

David M. Hannah, Iain A. Malcolm, Chris Soulsby, and Alan F. Youngson. A comparison of forest and moorland stream microclimate, heat exchanges and thermal dynamics. *Hydrological Processes*, 22(7):919–940, March 2008. ISSN 08856087, 10991085. doi: 10.1002/hyp.7003. URL http://doi.wiley.com/10.1002/hyp.7003.

Andrew C. Harvey and James Durbin. The effects of seat belt legislation on British road casualties: A case study in structural time series modelling. *Journal of the Royal Statistical Society. Series A (General)*, pages 187–227, 1986. URL http://www.jstor.org/stable/2981553.

Ray Hilborn and Marc Mangel. *The ecological detective: confronting models with data*, volume 28. Princeton University Press, 1997.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260): 663–685, 1952. ISSN 01621459. URL http://www.jstor.org/stable/2280784.

Markus Hrachowitz, C. Soulsby, C. Imholt, I. A. Malcolm, and D. Tetzlaff. Thermal regimes in a large upland salmon river: a simple model to identify the influence of landscape controls and climate change on maximum temperatures. *Hydrological Processes*, 24(23):3374–3391, November 2010. ISSN 08856087. doi: 10.1002/hyp.7756. URL http://doi.wiley.com/10.1002/hyp.7756.

Richard M. Huggins and Paul SF Yip. Statistical analysis of removal experiments with the use of auxilliary variables. *Statistica Sinica*, 7(3):705–712, 1997. URL http://hub.hku.hk/handle/10722/45345.

FL Jackson, IA Malcolm, and David M Hannah. A novel approach for designing large-scale river temperature monitoring networks. *Hydrology Research*, page nh2015106, 2015.

Lucas Janson, William Fithian, and Trevor Hastie. Effective degrees of freedom: a flawed metaphor. *arXiv preprint arXiv:1312.7851*, 2013. URL http://arxiv.org/abs/1312.7851.

G. J. A. Kennedy and C. D. Strange. Efficiency of Electric Fishing for Salmonids in Relation to River Width. *Aquaculture Research*, 12(2):55–60, April 1981. ISSN 1365-2109. doi: 10.1111/j.1365-2109.1981.tb00010.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2109.1981.tb00010.x/abstract.

William J. Kennedy and James E. Gentle. Statistical computing. *New York*, 1980. URL http://www.maa.org/publications/maa-reviews/statistical-computing.

Thomas Kneib and Ludwig Fahrmeir. Structured Additive Regression for Categorical Space-Time Data: A Mixed Model Approach. *Biometrics*, 62(1):109–118, March 2006. ISSN 0006-341X. URL http://www.jstor.org/stable/3695711.

Robert P. Lanka, Wayne A. Hubert, and Thomas A. Wesche. Relations of Geomorphology to Stream Habitat and Trout Standing Stock in Small Rocky Mountain Streams. *Transactions of the American Fisheries Society*, 116(1):21–28, January 1987. ISSN 0002-8487. doi: 10.1577/1548-8659(1987)116<21:ROGTSH>2.0.CO;2. URL http://dx.doi.org/10.1577/1548-8659(1987)116<21:ROGTSH>2.0.CO;2.

Steffen L. Lauritzen. Time Series Analysis in 1880: A Discussion of Contributions Made by T.N. Thiele. *International Statistical Review / Revue Internationale de Statistique*, 49(3):319–331, December 1981. ISSN 0306-7734. doi: 10.2307/1402616. URL http://www.jstor.org/stable/1402616.

J. A. Leach, R. D. Moore, S. G. Hinch, and T. Gomi. Estimation of forest harvesting-induced stream temperature changes and bioenergetic consequences for cutthroat trout in a coastal stream in British Columbia, Canada. *Aquatic Sciences*, 74(3):427–441, July 2012. ISSN 1015-1621, 1420-9055. doi: 10.1007/s00027-011-0238-z. URL http://link.springer.com/10.1007/s00027-011-0238-z.

Youngjo Lee and John A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 619–678, 1996. URL http://www.jstor.org/stable/2346105.

Youngjo Lee and John A. Nelder. Modelling and analysing correlated non-normal data. *Statistical Modelling*, 1(1):3–16, 2001. URL http://smj.sagepub.com/content/1/1/3.short.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. URL http://biomet.oxfordjournals.org/content/73/1/13.short.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, September 2011. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2011.00777.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2011.00777.x/abstract.

David J. Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, October 2000. ISSN 0960-3174, 1573-1375. doi: 10.1023/A:1008929526011. URL http://link.springer.com/article/10.1023/A%3A1008929526011.

J S Macdonald, E A MacIsaac, and H E Herunter. The effect of variable-retention riparian buffer zones on water temperatures in small headwater streams in sub-boreal forest ecosystems of British Columbia. *Canadian Journal of Forest Research*, 33(8):1371–1382, August 2003. ISSN 0045-5067. doi: 10.1139/x03-015. URL http://www.nrcresearchpress.com/doi/abs/10.1139/x03-015.

Michael W. Mahoney and Petros Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, January 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0803205106. URL http://www.pnas.org/content/106/3/697.

I. A. Malcolm, C. Soulsby, D. M. Hannah, P. J. Bacon, A. F. Youngson, and D. Tetzlaff. The influence of riparian woodland on stream temperatures: implications for the performance of juvenile salmonids. *Hydrological Processes*, 22(7):968–979, March 2008a. ISSN 08856087, 10991085. doi: 10.1002/hyp.6996. URL http://doi.wiley.com/10.1002/hyp.6996.

IA Malcolm, C Soulsby, DM Hannah, PJ Bacon, AF Youngson, and D Tetzlaff. The influence of riparian woodland on stream temperatures: implications for the performance of juvenile salmonids. *Hydrological processes*, 22(7):968–979, 2008b.

Iain A. Malcolm, Alan F. Youngson, and Chris Soulsby. Survival of salmonid eggs in a degraded gravel-bed stream: effects of groundwater–surface water interactions. *River Research and Applications*, 19(4):303–316, July 2003. ISSN 1535-1467. doi: 10.1002/rra.706. URL http://onlinelibrary.wiley.com/doi/10.1002/rra.706/abstract.

C. L. Mallows. Some Comments on CP. *Technometrics*, 15(4):661–675, November 1973. ISSN 0040-1706. doi: 10.2307/1267380. URL http://www.jstor.org/stable/1267380.

Giampiero Marra and Simon N Wood. Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7):2372–2387, 2011.

Sara Martino and H\a avard Rue. Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the inla program. *Department of Mathematical Sciences, NTNU, Norway*, 2009. URL http://www.bias-project.org.uk/gmrfcourse/inla-program.pdf.

Eric Mellina, R Dan Moore, Scott G Hinch, J Stevenson Macdonald, and Greg Pearson. Stream temperature responses to clearcut logging in British Columbia: the moderating influences of groundwater and headwater lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, 59(12):1886–1900, December 2002. ISSN 0706-652X, 1205-7533. doi: 10.1139/f02-158. URL http://www.nrcresearchpress.com/doi/abs/10.1139/f02-158.

R. Moore, D. L. Spittlehouse, and Anthony Story. *RIPARIAN MICROCLIMATE AND STREAM TEMPERATURE RESPONSE TO FOREST HARVESTING: A REVIEW1*. Wiley Online Library, 2005. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1752-1688.2005.tb03772.x/abstract.

R.D. Moore, M. Nelitz, and E. Parkinson. Empirical modelling of maximum weekly average stream temperature in British Columbia, Canada, to support assessment of fish habitat suitability. *Canadian Water Resources Journal*, 38(2):135–147, June 2013. ISSN 0701-1784, 1918-1817. doi: 10.1080/07011784.2013.794992. URL http://www.tandfonline.com/doi/abs/10.1080/07011784.2013.794992.

P. A. P. Moran. A Mathematical Theory of Animal Trapping. *Biometrika*, 38 (3/4):307–311, December 1951. ISSN 0006-3444. doi: 10.2307/2332576. URL http://www.jstor.org/stable/2332576.

Paul A Murtaugh. Performance of several variable-selection methods applied to real ecological data. *Ecology Letters*, 12(10):1061–1068, 2009.

Eero Niemelä, Markku Julkunen, and Jaakko Erkinaro. Quantitative electrofishing for juvenile salmon densities: assessment of the catchability during a long-term monitoring programme. *Fisheries Research*, 48(1):15–22, August 2000. ISSN 0165-7836. doi: 10.1016/S0165-7836(00)00113-2. URL http://www.sciencedirect.com/science/article/pii/S0165783600001132.

David L. Otis, Kenneth P. Burnham, Gary C. White, and David R. Anderson. Statistical Inference from Capture Data on Closed Animal Populations. *Wildlife Monographs*, 62:3–135, October 1978. ISSN 0084-0173. URL http://www.jstor.org/stable/3830650.

Erin E. Peterson, Jay M. Ver Hoef, Dan J. Isaak, Jeffrey A. Falke, Marie-Josée Fortin, Chris E. Jordan, Kristina McNyset, Pascal Monestiez, Aaron S. Ruesch, Aritra Sengupta, Nicholas Som, E. Ashley Steel, David M. Theobald, Christian E. Torgersen, and Seth J. Wenger. Modelling dendritic ecological networks in space: an integrated network perspective. *Ecology Letters*, 16(5):707–719, May 2013. ISSN 1461023X. doi: 10.1111/ele.12084. URL http://doi.wiley.com/10.1111/ele.12084.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org/.

Lawrence E Raffalovich, Glenn D Deane, David Armstrong, and Hui-Shien Tsao. Model selection procedures in social research: Monte-carlo simulation results. *Journal of Applied Statistics*, 35(10):1093–1114, 2008.

J. O Ramsay and B. W Silverman. *Functional data analysis*. Springer, New York, 2005. ISBN 9780387400808 038740080X 9780387227511 0387227512. URL http://www.knovel.com/knovel2/Toc.jsp?BookID=1804.

Etienne Rivot, Etienne Prévost, Anne Cuzol, Jean-Luc Bagliniére, and Eric Parent. Hierarchical Bayesian modelling with habitat and time covariates for estimating riverine fish population size by successive removal method. *Canadian Journal of Fisheries and Aquatic Sciences*, 65(1):117–133, January 2008. ISSN 0706-652X. doi: 10.1139/f07-153. URL http://www.nrcresearchpress.com/doi/abs/10.1139/f07-153.

Jordan Rosenfeld, Marc Porter, and Eric Parkinson. Habitat factors affecting the abundance and distribution of juvenile cutthroat trout (Oncorhynchus clarki) and coho salmon (Oncorhynchus kisutch). *Canadian Journal of Fisheries and Aquatic Sciences*, 57(4):766–774, April 2000. ISSN 0706-652X. doi: 10.1139/f00-010. URL http://www.nrcresearchpress.com/doi/abs/10.1139/f00-010.

Havard Rue and Leonhard Held. *Gaussian Markov Random Fields*. Number 104 in Monographs on Statistics and Applied Probability. Chapman and Hall / CRC press, Boca Raton, 2005. URL http://www.crcpress.com/product/isbn/9781584884323.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, April 2009. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2008.00700.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2008.00700.x/abstract.

Fabian Scheipl, Thomas Kneib, and Ludwig Fahrmeir. Penalized likelihood and Bayesian function selection in regression models. *AStA Advances in Statistical Analysis*, 97(4):349–385, April 2013a. ISSN 1863-8171, 1863-818X. doi: 10.1007/s10182-013-0211-3. URL http://link.springer.com/article/10.1007/s10182-013-0211-3.

Fabian Scheipl, Thomas Kneib, and Ludwig Fahrmeir. Penalized likelihood and Bayesian function selection in regression models. *AStA Advances in Statistical Analysis*, 97(4):349–385, April 2013b. ISSN 1863-8171, 1863-818X. doi: 10.1007/s10182-013-0211-3. URL http://link.springer.com/article/10.1007/s10182-013-0211-3.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464, 1978. URL http://projecteuclid.org/euclid.aos/1176344136.

George Arthur Frederick Seber. *The estimation of animal abundance*. 1982. URL http://tocs.ulb.tu-darmstadt.de/5373212X.pdf.

Bernard W. Silverman. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1):1–24, 1996. URL http://projecteuclid.org/euclid.aos/1033066196.

SNIFFER. River fish Classification tool: Science Work. Technical report, SNIFFER, WFD, 2011.

Stan Development Team. *RStan: the R interface to Stan, Version 2.5.0*. 2014a. URL http://mc-stan.org/rstan.html.

Stan Development Team. *Stan: A C++ Library for Probability and Sampling, Version 2.5.0*. 2014b. URL http://mc-stan.org/.

T. Stott and S. Marks. Effects of plantation forest clearfelling on stream temperatures in the Plynlimon experimental catchments, mid-Wales. *Hydrology and Earth System Sciences Discussions*, 4(1):95–104, 2000. URL https://hal.archives-ouvertes.fr/hal-00304513.

Gilbert Strang. *Introduction to linear algebra*. Wellesley-Cambridge Press, Wellesley, MA, 2009. ISBN 9780980232714 0980232716 9780980232721 0980232724.

Cajo JF Ter Braak. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67(5):1167–1179, 1986.

T. N. Thiele. Sur la compensation de quelques Erreurs Quasi-systematiques. 1880.

James T. Thorson and Cóilín Minto. Mixed effects: a unifying framework for statistical modelling in fisheries biology. *ICES Journal of Marine Science: Journal du Conseil*, page fsu213, December 2014. ISSN 1054-3139, 1095-9289. doi: 10.1093/icesjms/fsu213. URL http://icesjms.oxfordjournals.org/content/early/2014/12/03/icesjms.fsu213.

Arũnas P. Verbyla, Brian R. Cullis, Michael G. Kenward, and Sue J. Welham. The Analysis of Designed Experiments and Longitudinal Data by Using Smoothing Splines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48 (3):269–311, January 1999. ISSN 1467-9876. doi: 10.1111/1467-9876.00154. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9876.00154/abstract.

Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990. URL https://books.google.co.uk/books?hl=en&lr=&id=BS6fGsebFNkC&oi=fnd&pg=PR4&dq=wahba+1990&ots=_CwVuMVBwQ&sig=wvTNuVCw51xJeEdOxBSIumuS2tY.

Fred Watson, Rob Vertessy, Tom McMahon, Bruce Rhodes, and Ian Watson. Improved methods to assess water yield changes from paired-catchment studies: application to the Maroondah catchments. *Forest Ecology and Management*, 143(1):189–204, 2001. URL http://www.sciencedirect.com/science/article/pii/S037811270000517X.

B. W. Webb and D. T. Crisp. Afforestation and stream temperature in a temperate maritime environment. *Hydrological Processes*, 20(1):51–66, January 2006. ISSN 1099-1085. doi: 10.1002/hyp.5898. URL http://onlinelibrary.wiley.com/doi/10.1002/hyp.5898/abstract.

Edmund T. Whittaker. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:63–75, 1922. URL http://journals.cambridge.org/abstract_S001309150000359X.

Christopher K Wikle. Hierarchical bayesian models for predicting the spread of ecological processes. *Ecology*, 84(6):1382–1394, 2003.

Christopher K. Wikle, Ralph F. Milliff, Radu Herbei, and William B. Leeds. Modern Statistical Methods in Oceanography: A Hierarchical Perspective. *Statistical Science*, 28(4):466–486, November 2013. ISSN 0883-4237, 2168-8745. doi: 10.1214/13-STS436. URL http://projecteuclid.org/euclid.ss/1386078874.

Simon Wood. *Generalized additive models: an introduction with R*. CRC press, 2006. URL https://books.google.co.uk/books?hl=en&lr=&id=GbzXe-L8uFgC&oi=fnd&pg=PP1&dq=wood+2006+generalized+additive+models&ots=BMIFu-bnX9&sig=wwiZxLMlE0RLU0n1SPpxlkrPkuw.

Simon N. Wood. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114, 2003. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00374/full.

Simon N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36, January 2011. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2010.00749.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2010.00749.x/abstract.

Simon N. Wood, Mark V. Bravington, and Sharon L. Hedley. Soap film smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70 (5):931–955, November 2008. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2008.00665.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2008.00665.x/abstract.

RJ Wyatt and S Barnard. The transportation of the maximum gain salmon spawning target from the River Bush (NI) to England and Wales. *Environment Agency R & D Technical Report*, (W65):38, 1997.

Robin Wyatt. River Fish Habitat Inventory phase 2 : methodology developemnt for juvenile salmonids. Technical report, Environment Agency, Bristol, 2005.

Robin Wyatt, Roy Sedgwick, and Helen Simcox. River fish habitat inventory phase 3: multi-species models. Technical report, Environment Agency, Bristol, 2007.

Robin J Wyatt. Estimating riverine fish population size from single- and multiple-pass removal sampling using a hierarchical model. *Canadian Journal of Fisheries and Aquatic Sciences*, 59(4):695–706, April 2002. ISSN 0706-652X. doi: 10.1139/f02-041. URL http://www.nrcresearchpress.com/doi/abs/10.1139/f02-041.

Robin J Wyatt. Mapping the abundance of riverine fish populations: integrating hierarchical Bayesian models with a geographic information system (GIS). *Canadian Journal of Fisheries and Aquatic Sciences*, 60(8): 997–1006, August 2003. ISSN 0706-652X. doi: 10.1139/f03-085. URL http://www.nrcresearchpress.com/doi/abs/10.1139/f03-085.

Thomas W Yee and Trevor J Hastie. Reduced-rank vector generalized linear models. *Statistical modelling*, 3(1):15–41, 2003.

Yu Ryan Yue, Daniel Simpson, Finn Lindgren, and Håvard Rue. Bayesian Adaptive Smoothing Splines Using Stochastic Differential Equations. *Bayesian Analysis*, 9 (2):397–424, June 2014. ISSN 1936-0975, 1931-6690. doi: 10.1214/13-BA866. URL http://projecteuclid.org/euclid.ba/1401148314.

Calvin Zippin. An Evaluation of the Removal Method of Estimating Animal Populations. *Biometrics*, 12(2):163–189, June 1956. ISSN 0006-341X. doi: 10.2307/3001759. URL http://www.jstor.org/stable/3001759.

Alain F Zuur, Elena N Ieno, and Graham M Smith. *Analysing ecological data*. Springer, New York; London, 2007. ISBN 9780387459721 0387459723 0387459677 9780387459677. URL http://public.eblib.com/choice/publicfullrecord.aspx?p=371384.

Maciej A. Zwieniecki and Michael Newton. Influence of Streamside Cover and Stream Features on Temperature Trends in Forested Streams of Western Oregon. *Western Journal of Applied Forestry*, 14(2):106–113, April 1999.