UNIVERSITY OF GLASGOW

Using river network structure to improve estimation of common temporal patterns

by

Kelly Marie Gallacher

A thesis submitted for the degree of
Doctor of Philosophy

in the

College of Science and Engineering
School of Mathematics and Statistics

March 2016

Declaration of Authorship

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was executed by myself, unless otherwise stated.

Part of the work presented in Chapter 2 has been presented as a poster at the 28th International Workshop on Statistical Modelling (IWSM) in Palermo, 2013, with the title "Statistical methods comparing seasonality for water quality" and is included in the corresponding journal of conference proceedings.

Signed: _____

Date: _____

"By perseverance the snail reached the ark."

Charles Spurgeon (1834-1892)

Abstract

Statistical models for data collected over space are widely available and commonly used. These models however usually assume relationships between observations depend on Euclidean distance between monitoring sites whose location is determined using two dimensional coordinates, and that relationships are not direction dependent. One example where these assumptions fail is when data are collected on river networks. In this situation, the location of monitoring sites along a river network relative to other sites is as important as the location in two dimensional space since it can be expected that spatial patterns will depend on the direction of water flow and distance between monitoring site measured along the river network. Euclidean distance therefore might no longer be the most appropriate distance metric to consider. This is further complicated where it might be necessary to consider both Euclidean distance and distance along the river network if the observed variable is influenced by the land in which the river network is embedded.

The Environment Agency (EA), established in 1996, is the government agency responsible for monitoring and improving the water quality in rivers situated in England (and Wales until 2013). A key responsibility of the EA is to ensure that efforts are made to improve and maintain water quality standards in compliance with EU regulations such as the Water Framework Directive (WFD, European Parliament (2000)) and Nitrates Directive (European Parliament, 1991). Environmental monitoring is costly and in many regions of the world funding for environmental monitoring is decreasing (Ferreyra et al., 2002). It is therefore important to develop statistical methods that can extract as much information as possible from existing or reduced monitoring networks. One way to do this is to identify common temporal patterns shared by many monitoring sites so that redundancy in the monitoring network could be reduced by removing non-informative sites exhibiting the same temporal patterns. In the case of river water quality, information about the shape of the river network, such as flow direction and connectivity of monitoring sites, could be incorporated into statistical techniques to improve statistical power and provide efficient inference without the increased cost of collecting more data. Reducing the volume of data required to estimate temporal trends would improve efficiency and provide cost savings to regulatory agencies.

The overall aim of this thesis is to investigate how information about the spatial structure of river networks can be used to augment and improve the specific trends obtained when using a variety of statistical techniques to estimate temporal trends in water quality data. Novel studies are designed to investigate the effect of accounting for river network structure within existing statistical techniques and, where necessary, statistical methodology is developed to show how this might be achieved. Chapter 1 provides an

introduction to water quality monitoring and a description of several statistical methods that might be used for this. A discussion of statistical problems commonly encountered when modelling spatiotemporal data is also included. Following this, Chapter 2 applies a dimension reduction technique to investigate temporal trends and seasonal patterns shared among catchment areas in England and Wales. A novel comparison method is also developed to identify differences in the shape of temporal trends and seasonal patterns estimated using several different statistical methods, each of which incorporate spatial information in different ways. None of the statistical methods compared in Chapter 2 specifically account for features of spatial structure found in river networks: direction of water flow, relative influence of upstream monitoring sites on downstream sites, and stream distance. Chapter 3 therefore provides a detailed investigation and comparison of spatial covariance models that can be used to model spatial relationships found in river networks to standard spatial covariance models. Further investigation of the spatial covariance function is presented in Chapter 4 where a simulation study is used to assess how predictions from statistical models based on river network spatial covariance functions are affected by reducing the size of the monitoring network. A study is also developed to compare the predictive performance of statistical models based on a river network spatial covariance function to models based on spatial covariate information, but assuming spatial independence of monitoring sites. Chapter 3 and 4 therefore address the aim of assessing the improvement in information extracted from statistical models after the inclusion of information about river network structure. Following this, Chapter 5 combines the ideas of Chapters 2, 3 and 4 and proposes a novel statistical method where estimated common temporal patterns are adjusted for known spatial structure, identified in Chapters 3 and 4. Adjusting for known structure in the data means that spatial and temporal patterns independent of the river network structure can be more clearly identified since they are no longer confounded with known structure.

The final chapter of this thesis provides a summary of the statistical methods investigated and developed within this thesis, identifies some limitations of the work carried out and suggests opportunities for future research. An Appendix provides details of many of the data processing steps required to obtain information about the river network structure in an appropriate form.

Acknowledgements

Firstly I would like to thank my supervisors Prof. Marian Scott and Dr Claire Miller for their limitless advice, support and guidance over the past four years. I am privileged to have had the opportunity to work with them. I would like to thank the Environment Agency for providing data and advice during the PhD, in particular Robert Willows, Linda Pope and John Douglass. I also gratefully acknowledge funding received from an Engineering Sciences and Physical Research Council Doctoral Training Account.

It has been a great pleasure and inspiration to work with the Statistics staff and post-graduates during my research, especially Ruth and Ally who offered much advice and lunchtime chat and to all of my office colleagues who made the PhD an enjoyable time. Thank you also to all of the support staff in the Maths building who answered many dumb questions and sorted out my many IT issues. Special thanks go to Lorraine who always has a friendly smile.

A huge thank you to all of my family. Your love, support and encouragement throughout my life to be the best I can be made me think it possible to return to university and attempt a PhD.

And finally to Shaw, without you I would never have finished this PhD. Your unfailing faith in me kept me going when everything seemed impossible and during the tough times you managed to convince me to keep going just a little longer. And yes, you were always right.

Contents

List of Figures

List of Tables

Abbreviations

.ssn	An R object required by the SSN package
AIC	Akaike's information criterion
AKSE	Average kriging standard error
ANCOVA	Analysis of covariance
DBF	Hypothesis test based on visual distance
DFA	Dynamic factor analysis
EA	Environment Agency for England and Wales
EOF	Empirical orthogonal functions
fANOVA	Functional analysis of variance
FDA	Functional data analysis
GAM	Generalized additive model
GCV	Generalized Cross Validation
ha	hectares
INLA	Integrated nested laplace approximations
km	Kilometer
LAD	Local authority district
LHA	Large hydrological area
LOOCV	Leave one out cross validation
ML	Maximum Likelihood
mg	milligrams
ml	millilitres
mm	millimetres
NIPALS	Nonlinear iterative partial least squares
netID	Network ID assigned during creation of .ssn object
PC	Principal component

PCA	Principal components analysis
RCA	Reach contributing area
REML	Restricted Maximum Likelihood
RMSPE	Root mean square prediction error
ROS	Regression on order statistics
SAM	Standard area measurements
smANCOVA	Analysis of covariance for non parametric curves
SMSE	Scaled mean square error
SPDE	Stochastic partial differential equations
SSN	Spatial stream network
SVD	Singular value decomposition
TON	Total oxidised nitrogen
USGS	United States Geological Survey
WFD	Water Framework Directive

Chapter 1

Introduction

'Freshwater is a finite resource, essential for agriculture, industry and even human existence.'

(Bartram and Ballance, 1996)

Statistical models for data collected over space are widely available and commonly used. These models however usually assume relationships between observations depend on Euclidean distance between monitoring sites whose location is determined using two dimensional coordinates, and that relationships are not direction dependent. One example where these assumptions fail is when data are collected on river networks. In this situation, the location of monitoring sites along a river network relative to other sites is as important as the location in two dimensional space since it can be expected that spatial patterns will depend on the direction of water flow and distance between monitoring site measured along the river network since Euclidean distance might no longer be the most appropriate distance metric to consider. This is further complicated where it might be necessary to consider both Euclidean distance and distance along the river network if the observed variable is influenced by the land in which the river network is embedded.

The Environment Agency (EA), established in 1996, is the government agency responsible for monitoring and improving the water quality in rivers situated in England (and Wales until 2013). A key responsibility of the EA is to ensure that efforts are made to improve and maintain water quality standards in compliance with EU regulations such as the Water Framework Directive (WFD, European Parliament (2000)) and Nitrates

Directive (European Parliament, 1991). The WFD is a piece of European legislation that "establishes an innovative approach for water management based on river basins, the natural geographical and hydrological units and sets specific deadlines for Member States to protect aquatic ecosystems" while the Nitrates directive "aims to protect water quality across Europe by preventing nitrates from agricultural sources polluting ground and surface waters and by promoting the use of good farming practices". These legislative aims are achieved in part by placing monitoring sites along the river networks in England (and Wales) and measuring several water quality variables. These variables are monitored over time and used to classify the status of water bodies. A water body is a section of river whose geology, pollution levels and other pressures make it distinguishable from other sections of river around it. The term 'water body' can also be applied to lakes, coastal waters and transitional waters connecting rivers to marine water. Member states are required to identify water bodies at the appropriate scale to manage the requirements of the directives. The status of water bodies is identified through statistical modelling and then classification of temporal patterns in water quality data and there are many different statistical techniques that could be used for this purpose. The WFD requires that all water bodies in Europe achieve "good ecological status" by 2015 but Burt et al. (2010) predicted that this was unlikely to be achieved, in large part due to the effect of diffuse pollution from agriculture. In fact, by 2015 only 53% of water bodies achieved 'good' ecological status (EEA, 2015). The main contributors to poor quality are intensive agricultural practices and sewage waste in areas of high urbanisation (EEA, 2015).

Environmental monitoring is costly and in many regions of the world funding for environmental monitoring is decreasing (Ferreyra et al., 2002). It is therefore important to develop statistical methods that can extract as much information as possible from existing or reduced monitoring networks. One way to do this is to identify common temporal patterns shared by many monitoring sites so that redundancy in the monitoring network could be reduced by removing non-informative sites exhibiting the same temporal patterns. In the case of river water quality, information about the shape of the river network, such as flow direction and connectivity of monitoring sites, could be incorporated into statistical techniques to provide additional information without the increased cost of collecting more data. Reducing the volume of data required to estimate temporal trends would improve efficiency and provide cost savings to regulatory

agencies. This thesis aims to investigate how information about the spatial structure of river networks can be used to improve the estimation of temporal trends.

1.1 The data

Data used in this thesis were provided by the EA for a single water quality variable - total oxidised nitrogen (TON) measured in mg/l, recorded on the main water quality monitoring network for England and Wales from 1990 to 2010. Water quality in Wales has been the responsibility of the Environment Agency Wales since 2013 but when data were provided for this thesis the EA held responsibility for both England and Wales. TON is the sum of nitrate (NO_3) and nitrite (NO_2) and these are usually combined due to nitrites often being present in such small amounts and the effect of nitrate and nitrite on water quality is very similar. Nitrates can enter the water supply from a variety of sources but these can generally be classified into two categories: point source and diffuse pollution. Point sources generally come from industrial outputs such as sewage treatment plants whereas diffuse pollution does not have a single point of origin and an example of this is fertilizers used in agriculture where nitrates enter the river supply as runoff from the surrounding land (European Parliament, 1991). Nitrates are nutrients, stimulating the growth of plankton and aquatic plants that are eaten by fish but large quantities can cause eutrophication, damage the ecosystem and deprive fish of oxygen, leading to poor ecological status. Under European legislation the upper safety limit of TON is 50mg/l (European Parliament, 2000) although in the United States of America this is reduced to 10mg/l (SDWA (1974), SDWA (1996)).

The EA water quality monitoring programme divides England and Wales into 59 large hydrological areas (LHA's) as shown in Figure 1.1, each of which contains a collection of river networks independent of those in neighbouring LHA's. The LHA's can be grouped into eight regions as shown in Table 1.1. Figure 1.2 shows the locations of the monitoring sites, coloured by region. Data are available for approximately monthly observations of TON at each monitoring site although some sites are sampled more or less frequently than this and data are not available for all sites for the full 21 year time period. The data are therefore a collection of time series, one for each monitoring site.

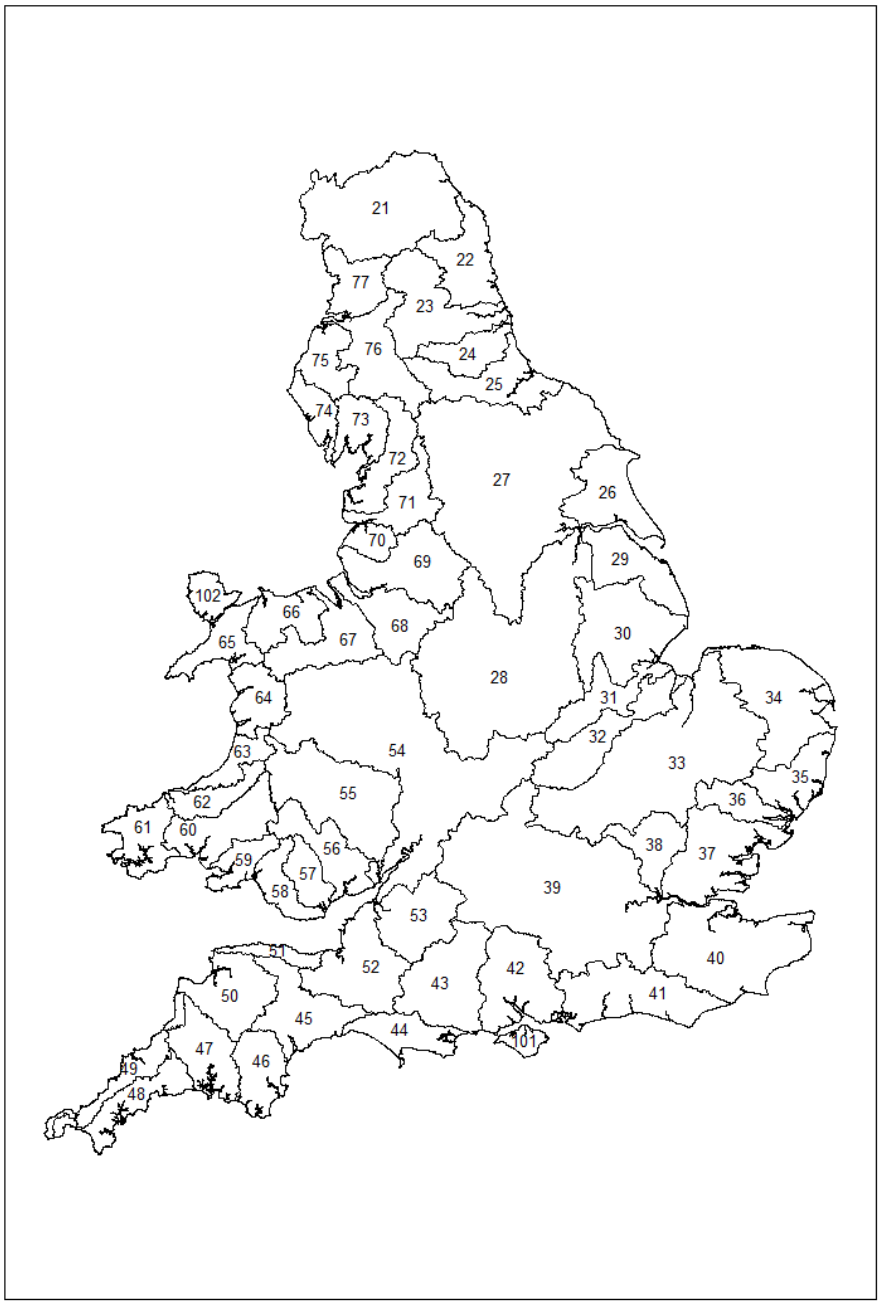


FIGURE 1.1: Large hydrological areas in England and Wales with numeric identifier. Black lines indicate borders of 59 LHA's.

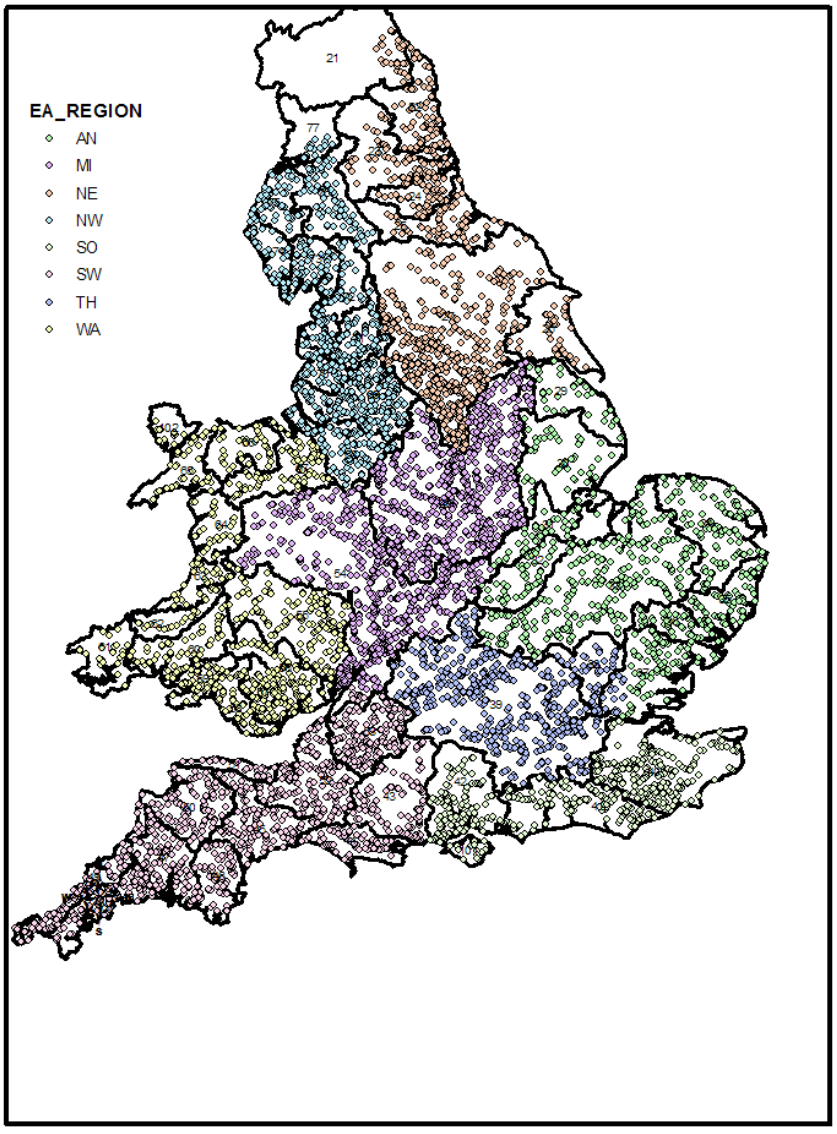


FIGURE 1.2: Large Hydrological Areas in England and Wales with monitoring stations (points). The 59 LHA's are grouped into 8 regions, represented by colour.

Region	Full Name	No. LHA's
AN	Anglia	9
MI	Midlands	2
NE	North East	7
NW	North West	10
SO	Southern	4
SW	South West	11
TH	Thames	2
WA	Wales	14

TABLE 1.1: Names of regions and number of LHA's within each region.

Withers and Nadarajah (2015) conclude that only 5 years of annual data are necessary to efficiently estimate a linear trend but assume a simple model with a linear trend, a constant sinusoidal seasonal pattern, and that the noise has constant variance and does not exhibit autocorrelation. The methods in this paper are demonstrated using air temperature data which might not exhibit the complexities of river water quality data and 5 annual means seems a surprisingly small amount of data when compared to the recommendations in Burt et al. (2008) who estimate that long time series of 30+ years of data are required to assess the effect of changes in land use on nitrate levels due to the slow response of water quality to changes in land management practices. In practice however, river water quality records in England (and Wales) are not available for such a long time period. Howden et al. (2011) also consider river water quality data (including nitrates) and find that at least 12 years of annual data are needed to separate short term hydrologic variability from long term changes in the system. This paper considers rivers from agricultural and non-agricultural catchments with the aim of creating general rules for length of water quality time series required to estimate long term trends. Monitoring sites with fewer than 120 observations (= 10 years of monthly observations) were removed from the data set so that temporal trends were not estimated for monitoring sites with data collected for fewer than 10 years. In this chapter long term trends are estimated from approximately monthly observations, after adjusting for seasonal patterns and so temporal trend, or long term changes, are estimated from at least 120 observations. Since data used in this Chapter are not aggregated to annual mean values, it was decided that monthly time series spanning at least 10 years were suitable to estimated temporal trends both at catchment level and for individual monitoring sites. In this thesis trends are estimated from annual means only in Chapter 5 and monitoring sites were selected to have 13 observations, meaning that 1 extra data point was included to be above the

minimum number of data points recommended in Howden et al. (2011).

2.3% of observations were recorded as below the limit of detection and monitoring sites with more than 50% of observations recorded as below the detection limit were removed from the data set since these monitoring sites were uninformative about trends in the data. Figure 1.3 shows the number of monitoring sites remaining after time series with fewer than 120 observations and series containing more than 50% censored values were removed.

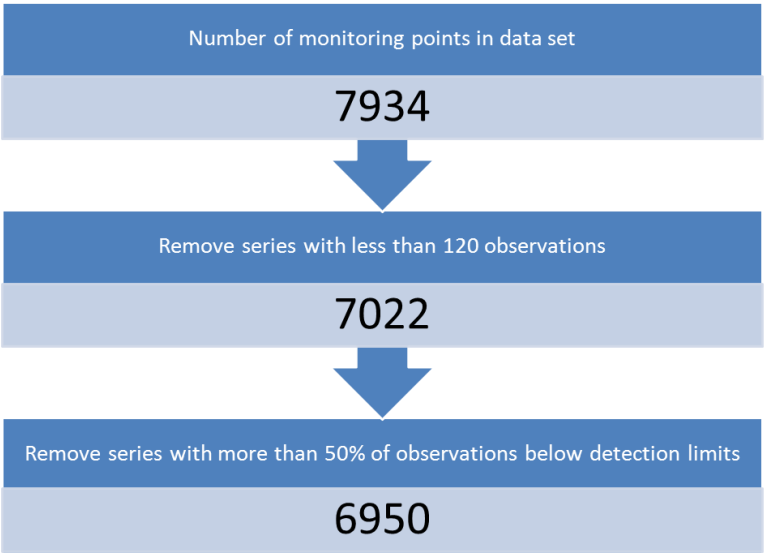


FIGURE 1.3: Number of monitoring stations remaining after various data management processes.

Henderson (2006) note that water quality data are often highly variable and skewed. One way to solve this is to take a natural log transformation of the data. A natural log transformation was taken of TON at all monitoring sites to stabilize the variability over time. All analyses in this thesis are therefore based on ln(TON), denoted log(TON). Singh et al. (1997) note that a transformation is preferable to modelling the median for skewed environmental data since highly skewed data will have a median lower than the mean, and confidence intervals will be unrealistically low. This is important as many environmental monitoring procedures are based on inferences on the mean.

One of the aims of this thesis is to incorporate spatial information about the river network into existing statistical techniques. Spatial information includes flow direction and distance between monitoring sites, where distance is measured along the river network between two monitoring sites ('as the fish swims') rather than using Euclidean

distance ('as the crow flies'). River network distance is called stream distance throughout this thesis. A geographic information system (GIS) can be used to estimate stream distance and flow direction in order to construct an asymmetric stream distance connectedness matrix. This requires two additional sets of data: a polyline (in GIS terminology) shapefile representing the shape of the river network and a digital elevation model (DEM) containing elevation information. The EA provided the polyline shapefile and the DEM was downloaded, in sections, from http://www.usgs.gov/. ArcGIS (v10.1) http://www.esri.com/ was used to calculate stream distances using the STARS toolkit (Peterson and Ver Hoef, 2014). Further information about the elevation data and steps required to calculate stream distances can be found in Appendix A.

Spatiotemporal environmental data such as water quality data collected by the EA, described above, can be modelled using a variety of statistical techniques to estimate spatial and temporal patterns, assess the effect of covariates on water quality and make predictions at unobserved locations. Sections 1.2, 1.3 and 1.4 describe statistical methods that can be used to model spatial and temporal patterns in environmental data while Section 1.5 discusses some of the issues commonly encountered when working with environmental data. Section 1.2 describes global and local methods commonly used to estimate the trend and seasonal pattern where data have been collected over time and Section 1.3 introduces statistical models that can be used to model the relationship between the response variable and covariates as well as trend and seasonal pattern. Section 1.4 introduces statistical techniques that can be used to investigate whether more than one trend or seasonal pattern is present in the situation where data are collected over time at multiple locations in space. It is intended that the remainder of this chapter is used to introduce notation and details of standard statistical methods and issues commonly encountered when these methods are applied to spatiotemporal data. Strengths, weaknesses, and adaptations of these methods when applied to data collected on river networks are considered more specifically in Chapters 2-5.

1.2 Representing trends and seasonal patterns

This section will describe statistical techniques that are commonly applied to data containing observations over time and where there is interest in estimating smooth trends or seasonal patterns. These techniques can be split into two categories: global and local

methods, and examples of each of these will be given. For local methods, some discussion is provided on choices that must be made when using these methods.

1.2.1 Global methods

When data have been recorded at several time points there will be interest in assessing how the response variable has changed over time (trend) and if the data contain periodicities such as seasonal patterns in environmental applications. A simple way to model a trend is to use a linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \text{for } i = 1,...,n \tag{1.1}$$

where Y_i is the response variable for observation i, x_i is the explanatory variable (in this case some measure of time since this is a model representing temporal trend) for observation i, ε_i are normally distributed errors for the i observations with a mean of zero and constant variance. $\boldsymbol{\beta} = (\beta_0, \beta_1)$ are coefficients that can be estimated using the least-squares estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

where $\mathbf{X}$ is the design matrix. A seasonal pattern could be simply modelled using a harmonic term such as

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 \cos(2\pi\omega x_i) + \beta_3 \sin(2\pi\omega x_i) + \varepsilon_i, \tag{1.2}$$

where $\omega = 1/M$ and M is the period over which the seasonal pattern is repeated e.g. $\omega = 365$ for daily data, 7 for weekly data, 12 for monthly data. The models in (1.1) and (1.2) are parametric and assume the relationship between the response and explanatory variable are of a fixed form that can be described with a few parameters. The models in (1.1) and (1.2) assume the trend is linear but this is often not the case in environmental applications. Models that allow trend to be non-linear and seasonal patterns to be more flexible than the harmonic representation can be fitted to allow for more flexible relationships between the response variable and time (or other continuous

covariates). An example of a model with a non-linear trend is polynomial regression where the simple linear model can be extended to a polynomial regression model of order m:

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i + ... + \beta_m x_i^m, \qquad \text{for } i = 1,...,n.$$

Polynomial regression is a simple example of basis expansion where a trend consists of a sum of functions of the data. Another simple basis expansion is the Fourier series where a non-linear trend $m(x)$ can be written as

$$m(x_i) \approx \frac{a_0}{2} + \sum_{j=1}^{p} a_j \cos\left(\frac{2\pi j x_i}{P}\right) + b_j \sin\left(\frac{2\pi j x_i}{P}\right),$$

where x_i are the data, a_0, a_j and b_j are parameters to be estimated, p are frequencies and P is the range of the covariate $\mathbf{x}$. Ramsay and Silverman (2006) state that the Fourier basis is useful when representing functions with no strong local features and where curvature is of the same order everywhere. Polynomial regression is useful if a polynomial of a low order is sufficient to model the data but along with the Fourier basis has some drawbacks. For example, it is reasonable to expect that the value of the response variable at x_i is more closely related to values of the response within a small range of x_i. The polynomial regression and Fourier expansion models do not have this local property and using a high order polynomial or including a greater range of frequencies can result in undesirable global effects. In addition to this as both of these models become more flexible the number of parameters to estimate increases requiring a correspondingly larger sample size. A further drawback is that these models can lead to high curvature at both ends of the range of $\mathbf{x}$ which is typically not supported by the data.

Instead, a more 'local' approach to estimating trends can be taken that will result in a non-linear trend but without the drawbacks of the global approach.

1.2.2 Local methods

1.2.2.1 Kernel smoothing

The local (linear) model aims to fit a regression model to a subsection of data around each data point x. Following the notation in Fan and Gijbels (1996), let the bandwidth h be the size of the neighbourhood around x. The data around x are then modelled using least squares, stressing the dependence of coefficients a and b on x as

$$Y_i = a(x) + b(x)X_i + \varepsilon_i \qquad \text{for } X_i \in x \pm h.$$

A weight scheme can be incorporated to give X_i closer to x more importance than those further away. The weight is a uni-modal, non-negative, symmetric function that decays sufficiently fast. The weight K, also called the kernel function $K\{(X_i - x)/h\}$ is assigned to each X_i and the following weighted least squares problem is minimized:

$$\sum_{i=1}^{n} \{Y_i - a(x) - b(x)X_i\}^2 K\left(\frac{X_i - x}{h}\right).$$

Bowman and Azzalini (1997) use a normal density for the kernel function. These local linear models are in fact a special case of local polynomial modelling where the polynomial is of degree 1. A more robust version of weighted local polynomial smoothing is LOWESS (LOcally WEighted Scatterplot Smoothing), introduced by Cleveland (1979) and Cleveland and Devlin (1988) where the following weighted local polynomial regression is fitted:

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j \right\}^2 K_h(X_i - x). \tag{1.3}$$

A polynomial of degree p is fitted locally with weighted least squares and the residuals obtained. Weights are then assigned to the residuals with large residuals receiving small weights and vice versa. The local polynomial fit is carried out again with the residual weights applied to their corresponding observations by multiplying the weight from the initial fit in (1.3) by the residual weight from the initial fit. This process is repeated a number of times. Cleveland (1979) recommend repeating the algorithm three times

and using a tricube kernel function where h is defined as the distance from a particular observation x and its r^{th} nearest neighbour and r is a proportion of the sample size.

The choice of h is an important issue in local linear modelling. A small h will lead to low bias but high variance in the local parameters that are based on few data points. On the other hand, a large h can create large bias. In fact if $h = 0$ this will lead to perfect interpolation of the data as opposed to $h = \infty$ which would lead to a global smoother and thus lose the benefits of local smoothing. h can be chosen by cross validation methods or the plug-in method described in Sheather and Jones (1991). Fan and Gijbels (1992) suggest using a variable bandwith to reflect variable amounts of smoothing around each data point. An interesting alternative to choosing h was proposed by Chaudhuri and Marron (1999) called *SiZer* (SIgnificant ZERo crossings of derivatives) where scale-space ideas are used to construct a 'SiZer map' - a graphical device that simultaneously displays the significance of features in the data for a range of bandwidths. In its original development (Chaudhuri and Marron, 1999) data were smoothed using local linear regression methods as described in the previous section, for several h values. Improvements were made to the method in Hannig and Marron (2006). SiZer was adapted for smoothing splines by Marron and Zhang (2005) and adapted for use in generalized additive models in Ganguli and Wand (2007) where significant features of surfaces was developed. SiZer was adapted by Rondonotti et al. (2007) to address the issue of significance of trends, taking into account correlation in the data. SiZer for the comparison of regression curves was developed by Park and Kang (2008) and for the comparison of time series by Park et al. (2009). Further extensions of SiZer include censored data (Marron and de Uña-Álvarez, 2004) and circular data (Oliveira et al., 2014).

Kernel smoothing, a local smoothing method, overcomes many of the drawbacks of the global smoothing methods discussed in the previous section but also has some disadvantages. For example, Peifer et al. (2003) note that computational complexity can be a problem if a kernel without finite support is used and also that boundary is often biased. Another form of local smoothing - spline smoothing - overcomes some of the drawbacks of kernel smoothing. For example, in the Discussion of Silverman (1985), spline smoothing is shown to be preferable to both fixed and variable bandwidth kernel smoothing in situations where the data have a long right hand tail. A second advantage of spline smoothing compared to kernel smoothing is that in spline smoothing constraints on the

estimated curve, such as being forced to go through the origin, can be satisfactorily accommodated whereas this is not easily accommodated within kernel smoothing. Spline smoothing is described in detail in the next section.

1.2.2.2 Spline smoothing

Rather than fit a smooth running line as in Section 1.2.2.1, flexible functions can be fitted using piecewise polynomials of order m joined together at points called *knots*. The spline function is continuous and $m-2$ times differentiable at each knot[1]. The range $[x_1, x_n]$ of observed values $\mathbf{x}$ is divided into subintervals and knots are placed at the boundaries of the sub intervals. A spline can then be fitted across the whole range $[x_1, x_n]$ by fitting a polynomial segment in each interval. The set of (internal) knots is $x_i^* : i = 1, ..., q-2$. The spline is therefore defined by the number of knots q, the locations of the knots x_i^* and the order m of the piecewise polynomial segments used to represent the smooth function. The number of basis functions k in a spline basis system is determined by the order of the polynomial to be fitted in each sub-interval and the number of interior knots with the relationship $k = m + q - 2$. In an unpenalized model the number of knots acts as a smoothing parameter since increasing the number of knots will increase the flexibility of the regression function. This reduces the bias of the estimate but has a higher variance than a function with fewer knots (Fan and Gijbels, 1996).

The piecewise segments of the polynomial spline can be of any order m but in practice it is rare to use a higher order than $m = 3$ corresponding to a cubic spline. Some useful variations of the cubic spline are the natural cubic spline (Green and Silverman, 1993) and the cyclic cubic spline. A natural cubic spline assumes that the second derivative at the first and last knot is zero i.e. $m''(a) = m''(b) = 0$ and the cyclic cubic spline can be used to represent periodic signals in the data such as the seasonal pattern in an environmental data set. In this case the second derivative at the first and last knot is equal $(m''(a) = m''(b))$.

In order to fit a polynomial spline it is necessary to choose the number and location of knots to use. These choices can be influential on the fitted model so care must be taken.

[1] For a spline of degree 0 $f(\cdot)$ does not need to be continuous; for a spline of degree 1 $f(\cdot)$ need not be continuous but must be differentiable.

Knots can be placed at each data point but this would lead to exactly interpolating the data as opposed to smoothing it. Instead of placing a knot at every data point a smaller number of knots can be placed across the range of $\mathbf{x}$. Equally spaced knots are computationally the simplest but knots might also be placed according to quantiles of $\mathbf{x}$ creating a spline that is more flexible in regions where there are more data (Ramsay and Silverman, 1997). Ideally, more knots should be placed over regions where the function being estimated is highly non-linear and fewer knots placed where the function only weakly deviates from linearity. Data-driven methods exist for determining the optimal location of knots, one of which is described in Friedman and Silverman (1989). The algorithm described begins with a dense set of knots and eliminates unnecessary points using an algorithm similar to variable selection techniques used in multiple regression. Others algorithms for knot placement can be found in de Boor (2001).

Two common choices of spline basis are the truncated power basis and the B-spline basis (de Boor, 1978). Given a set of q knots ($q-2$ interior knots plus one knot at each end) $\mathbf{x}^*$, the truncated power basis of degree $m-1$ is defined as

$$b_j = (1, x, ..., x^{m-2}, (x - x_1^*)_+^{m-1}, (x - x_2^*)_+^{m-1}, ..., (x - x_q^*)_+^{m-1}),$$

$$\text{where } (z)_+^{m-1} = \begin{cases} z^{m-1} & \text{for } z > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The truncated power basis has $m + q - 2$ basis functions. The smooth curve $f(\cdot)$ is a linear combination of the basis functions b_j and the optimal spline can therefore be found using ordinary least squares methods to find the corresponding β_j. The truncated power basis can be numerically unstable however due to high correlations between powers of x. The more stable B-spline basis is often used instead where the idea is to use local basis functions to represent $f(\cdot)$ i.e. the non-zero range of each basis function is a small proportion of the range of $\mathbf{x}$. B-splines have an appealing property called compact support meaning that a B-spline basis of order m is non-zero over a maximum of m adjacent intervals. This results in a sparse design matrix making B-splines computationally efficient.

To define a B-spline basis first define $k + m + 1$ knots, $x_1 < x_2 < ... < x_{k+m+1}$, and the interval over which the spline is to be evaluated lies within $[x_{m+2}, x_k]$. This means that the first and last $m + 1$ knot locations are essentially arbitrary (and take the value zero). Following the notation in Wood (2006), an $(m + 1)^{\text{th}}$ order spline can now be represented as

$$f(x) = \sum_{i=1}^{k} B_i^m(x)\beta_i, \tag{1.4}$$

where the B-spline basis functions are defined recursively as in (1.5) and (1.6).

$$B_i^m = \frac{x - x_i}{x_{i+m+1} - x_i} B_i^{m-1}(x) + \frac{x_{i+m+2} - x}{x_{i+m-2} - x_{i+1}} B_{i+1}^{m-1}(x) \qquad i = 1, ..., k \tag{1.5}$$

$$B_i^{-1}(x) = \begin{cases} 1 & x_i \leq x < x_{i+1} \\ 0 & \text{otherwise.} \end{cases} \tag{1.6}$$

The smoothness of the curve is controlled by the number and location of knots but this choice is essentially arbitrary. A more rigorous method for controlling smoothness is required. Before proceeding it should be noted that it is not possible to use standard hypothesis testing based on backward selection since a model based on $k - 1$ evenly spaced knots is not typically nested within a model based on k evenly spaced knots (Wood, 2006).

An alternative approach to controlling smoothing in the model by the number of knots is to fit a model using slightly too many basis functions and penalize excessive roughness in the least squares fitting procedure. There are two commonly used forms for this penalty, one of which involves a penalty based on the second derivative of $f(\cdot)$:

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \int_0^1 [f''(x)]^2 \mathrm{d}x. \tag{1.7}$$

The smoothing parameter λ controls the trade off between model fit and smoothness. As $\lambda \to \inf$, $f(\cdot)$ will become a straight line and for $\lambda = 0$, $f(\cdot)$ is an un-penalized regression spline estimate.

Eilers and Marx (1996) propose an alternative to this which penalizes B-splines based on the difference between coefficients of adjacent B-splines. Following the notation in Wood (2006), this penalty $\mathcal{P}$ can be written as

$$\mathcal{P} = \sum_{i=1}^{k-1} (\beta_{i+1} - \beta_i)^2 = \beta_1^2 - 2\beta_1\beta_2 + 2\beta_2^2 - 2\beta_2\beta_3 + ... + \beta_k^2,$$

and in vector matrix notation

$$\mathcal{P} = \boldsymbol{\beta}^\top \mathbf{D} \boldsymbol{\beta} = \boldsymbol{\beta}^\top \begin{bmatrix} 1 & -1 & 0 & \cdot & \cdot \\ -1 & 2 & -1 & \cdot & \cdot \\ 0 & -1 & 2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \boldsymbol{\beta}.$$

In this case, it is necessary to minimise

$$\sum_{i=1}^{m} \left\{ y_i - \sum_{j=1}^{n} b_j(x_i)\beta_j \right\}^2 + \lambda \sum_{j=k+1}^{n} (\Delta^k \beta_j)^2.$$

Choosing λ

The problem of estimating the degree of smoothness of the model is now related to choosing λ as opposed to choosing q. Ideally, it is desirable to choose λ so that $\hat{f(\cdot)}$ is close to $f(\cdot)$. A non-parametric approach to choosing the degree of smoothing in the model is to minimise the ordinary cross validation score

$$\nu_o = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_i^{[-i]} - y_i)^2, \tag{1.8}$$

where $\hat{f}_i^{[-i]}$ is the model fitted to all of the data apart from x_i. This calculation is computationally demanding but it can be shown (see Section 4.5.2 in Wood (2006) for details) that

$$\nu_o = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 / (1 - A_{ii})^2, \tag{1.9}$$

where $\hat{f}(x_i)$ is the estimate from fitting to all of the data and $\mathbf{A}$ is the corresponding hat matrix. In practice it is usually the generalized cross validation score ν_g that is minimized where the weights $1 - A_{ii}$ are replaced by the mean weight $\text{tr}(\mathbf{I} - \mathbf{A})/n$. By minimising ν_g a model is selected based on its power to predict an unknown observation as opposed to how well the model fits the data from which it was estimated. Wood (2011) proposes an alternative to minimising ν_g based on maximising the restricted maximum likelihood (REML) or marginal likelihood (ML) criteria. Wood (2011) suggests that this method of choosing λ offers some improvement in mean square error performance relative to GCV, at the same computational cost.

Another approach is to minimise Akaike's Information Criterion (AIC) (Akaike, 1973), which makes use of the number of parameters in the model and calculated as in (1.10) where n_{par} is the number of parameters in the model and L is the maximised likelihood function.

$$\text{AIC} = 2n_{par} - 2\log(L) \tag{1.10}$$

Other measures include AIC_c and BIC with notation as defined for (1.10) and n is the sample size:

$$\text{AIC}_\text{c} = \text{AIC} + \frac{2n(n_{par}+1)}{n - n_{par} - 1},$$

$$\text{BIC} = n_{par}\log n - 2\log(L).$$

In an unpenalized linear regression problem the number of parameters n_{par} is easy to calculate but in non-parametric problems such as the p-spline models described in Eilers and Marx (1996), counting the number of parameters is not meaningful since not all parameters are free due to the roughness penalty. In an unpenalized linear regression model, n_{par} is the trace of the hat matrix $\mathbf{H}$:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top$$

with design matrix $\mathbf{X}$. For the non-parametric model, Eilers and Marx (1996) follow Hastie and Tibshirani (1990) in using the trace of the smoother matrix $\mathbf{S}$ as the effective number of parameters. The fitted values $\hat{m}_i, i = 1, ..., n$ of a non-parametric model can be expressed as

$$\hat{m}_i = \mathbf{S}y_i,$$

where $\hat{m}_i$ are the fitted values, $\mathbf{S}$ is a smoothing matrix whose rows contain the weights required to estimate at each evaluation point and y_i are observed responses. An approximate value for the degrees of freedom of the non-parametric model can be calculated as

$$\mathrm{df} = \mathrm{tr}\left\{\mathbf{S}\right\} = \mathrm{tr}\left\{\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{X}^\top\right\}, \tag{1.11}$$

where $\mathbf{D}$ is a suitable difference matrix, the exact form of which depends on the basis being used. For example, in the p-spline model proposed in Eilers and Marx (1996) the penalty is the squared difference between adjacent parameter values and

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 0 & \cdot & \cdot \\ -1 & 2 & -1 & \cdot & \cdot \\ 0 & -1 & 2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

AIC, AIC_c and BIC are all possible criteria for automatic smoothing parameter selection methods, along with the non-parametric CV and GCV described earlier in the chapter. AIC_c is essentially AIC adapted for small sample sizes and BIC applies a stronger penalty for additional parameters than AIC. Francisco-Fernandez and Opsomer (2005) consider the effect of spatial correlation on automatic smoothing parameter selection for the local linear model when the smoothing parameter is selected using GCV.

Spline smoothing has computational advantages over other smoothing methods. For example, B-splines have the property of compact support, allowing for sparse design matrices making the model fitting procedure computationally efficient. Further advantages include the possibility of setting constraints on the estimated smooth curve such as

in cyclic splines where the endpoints are forced to be equal. This gives a good representation of cyclical patterns such as the seasonal patterns often observed in environmental data.

An alternative to polynomial basis functions are wavelets basis functions but since wavelets are not a focus of this thesis they will not be described in any detail here. See Ramsay and Silverman (1997) or Fan and Gijbels (1996) for a brief introduction and references therein for details.

This section has presented several methods for representing temporal trends and seasonal patterns within data collected over time. If appropriate, trends can be represented as simple linear functions of time, while seasonal patterns are most simply represented using harmonic terms. Some flexibility can be introduced into the trend by including polynomial terms or a Fourier expansion but these global approaches to non-linear trends have some drawbacks. These drawbacks can be overcome using local methods to estimate a flexible trend over time. Two local methods were discussed: kernel smoothing and spline smoothing. While kernel smoothing offers much improvement over global methods, there can be problems with estimating the curve well in the tails. Some discussion was provided to show that spline smoothing offered many advantages over kernel smoothing in terms of estimation in the tails, setting boundary conditions, and computational efficiency for large data sets. Spline smoothing requires selection of a smoothing parameter λ and methods for choosing λ were discussed. It is recommended that λ is chosen after considering multiple selection criterion to ensure the choice is robust against the selection criterion. In general, kernel smoothing and spline smoothing will yield similar results but the advantages of spline smoothing over kernel smoothing make spline smoothing the most appealing option when fitting flexible regression models.

1.3 Modelling trends and seasonal patterns

The previous section described methods for representing temporal trends and seasonal patterns as smooth functions of time but it is also possible to have smooth relationships between the response variable and one or more covariates. Smooth functions of covariates can be modelled within the additive model framework or using functional data analysis techniques. This section will introduce some notation for additive models

and functional data analysis. These are two possible approaches for modelling smooth, flexible relationships between response and covariate variables. A comparison of temporal trends and seasonal patterns estimated using the methods described in this section is given in Section 2.2.2 where there is interest in comparing the temporal trend and seasonal pattern estimated for a single LHA using a variety of statistical methods.

1.3.1 Additive Models

An additive model (GAM), (Hastie and Tibshirani, 1990) is a sum of functions of covariates where the relationship between the response and covariates is a smooth function, not constrained to be of a particular parametric form. Following the notation in Wood (2006), a univariate GAM with covariate x_1 can be represented as

$$g(\mu_i) = f_1(x_{1i}), \tag{1.12}$$

where $\mu_i \equiv \mathbb{E}(Y_i)$ and Y_i are assumed to follow an exponential family distribution. In an environmental context, x_1 might be *Year* and the model in (1.12) would represent the trend of the response Y over time. The smooth function $f(\cdot)$ can contain more than one covariate and bivariate smooth functions are used to represent interactions between two covariates. For example

$$g(\mu_i) = f_2(x_{2i}, x_{3i})$$

might represent spatial relationships in a data set where observations are made at monitoring stations, identified by some co-ordinate system and $x_2 =$ Easting, $x_3 =$ Northing. Trivariate terms can also be included, possibly representing the interaction between space and time. For example, (1.13) shows a model where the trend is allowed to vary over space.

$$g(\mu_i) = f_3(x_{1i}, x_{2i}, x_{3i}) \tag{1.13}$$

Sums of smooth functions of covariates are also allowed such as in (1.14) where x_4 might represent a seasonal pattern that does not vary over time.

$$g(\mu_i) = f_3(x_{1i}, x_{2i}, x_{3i}) + f_4(x_{4i}) \tag{1.14}$$

Parametric terms may be included along with the smooth functions. This is useful for example if it reasonable to assume that a covariate has a parametric relationship (linear, quadratic, harmonic for example) with the response variable. This model is represented in (1.15) where $\mathbf{X}_i^*$ is a row of the design matrix for parametric terms and $\boldsymbol{\theta}$ is the corresponding parameter vector.

$$g(\mu_i) = \mathbf{X}_i^*\boldsymbol{\theta} + f_3(x_{1i}, x_{2i}, x_{3i}) + f_4(x_{4i}) \tag{1.15}$$

The penalized regression spline representation of additive models is fitted using a penalized form of least squares where the model parameters are estimated as

$$\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{D}\right)^{-1}\mathbf{X}^\top\mathbf{y},$$

with design matrix $\mathbf{X}$, smoothing parameter λ, penalty matrix $\mathbf{D}$ and response vector $\mathbf{y}$.

1.3.1.1 Model comparison

As in the linear model, it is useful to compare models to test if simpler models are adequate compared to more complicated models. Hastie and Tibshirani (1990) recommend using the residual sum of squares and their associated effective degrees of freedom (1.11) to perform model comparisons. For example, the residual sum of squares can be defined as

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

A comparison of two models, where model_0 is nested within model_1, can be expressed as

$$F = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1/(\mathrm{df}_0 - \mathrm{df}_1)}{\mathrm{RSS}_1/\mathrm{df}_1}. \tag{1.16}$$

Hastie and Tibshirani (1990) suggest referring the observed non-parametric F statistic to an F distribution with $(df_0 - df_1)$ and df_1 degrees of freedom.

One of the reasons to compare GAM's is that it might not be necessary for all terms to be represented as smooth functions. A simple linear term might suffice in some cases. To determine if a linear term is more appropriate than a flexible smooth term two models should be fitted - one with a flexible term and one with a linear term - and the two models compared. A significant p-value from the approximate F-test would mean that the more complicated model, with the flexible term, is appropriate. McMullan et al. (2007) show how the F-test can be adapted to compare models when data are correlated and spatial and temporal correlation is described in Section 1.5.1.

Additive models have been used to model water quality recently in Orr et al. (2015), Miller et al. (2014), Carvalho et al. (2011) and Ferguson et al. (2008) among many others.

1.3.1.2 Additive models fitted using INLA

GAM's can also be fitted using Bayesian methods (Fahrmeir and Lang (2001) or Lang and Brezger (2004) for example). More recently, integrated nested laplace approximations (INLA) proposed by Rue et al. (2009) and Lindgren et al. (2011) can be used to estimate posterior distributions of additive model parameters. In the particular case of models with a spatial component the model is estimated using the SPDE (stochastic partial differential equation) approach (Lindgren et al., 2011) where a Gaussian field with a dense Matérn covariance matrix is approximated by a Gaussian Markov random field with a neighbourhood structure and sparse precision matrix. A description of a spatio-temporal additive model estimated using INLA is given below following the notation in Cameletti et al. (2011).

Let $y(\mathbf{s}_i, t)$ denote the spatio-temporal process $Y(.,.)$ that represents log(TON) concentrations recorded at monitoring station $\mathbf{s}_i$, $i = 1, ..., d$ and timepoint $t = 1, .., T$ so that

$$y(\mathbf{s}_i, t) = \mathbf{z}(\mathbf{s}_i, t)\boldsymbol{\beta} + \xi(\mathbf{s}_i, t) + \varepsilon(\mathbf{s}_i, t),$$

where $\mathbf{z}(\mathbf{s}_i, t) = (z_1(\mathbf{s}_i, t), ..., z_p(\mathbf{s}_i, t))$ denotes the vector of p covariates for site $\mathbf{s}_i$ at time t, and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)\prime$ is the vector of coefficients. $\varepsilon(\mathbf{s}_i, t) \sim N(0, \sigma_\varepsilon^2)$ is the measurement error defined by a Gaussian white-noise process, both temporally and spatially uncorrelated. The spatio-temporal structure is captured in $\xi(\mathbf{s}_i, t)$ where

$$\xi(\mathbf{s}_i, t) = \alpha\xi(\mathbf{s}_i, t-1) + \omega(\mathbf{s}_i, t)$$

for $t = 1, .., T$, $|\alpha| < 1$ and $\xi(\mathbf{s}_i, 1)$ derives from the stationary distribution $N(0, \sigma_\omega^2/(1-\alpha^2))$. $\omega(\mathbf{s}_i, t)$ is supposed to be temporally independent and is characterized by the spatio-temporal covariance function

$$\text{Cov}(\omega(\mathbf{s}_i, t), \omega(\mathbf{s}_j, t')) = \begin{cases} 0 & \text{if} \quad t \neq t' \\ \sigma_\omega^2 \mathcal{C}(h) & \text{if} \quad t = t' \end{cases}$$

for $i \neq j$. The purely spatial correlation function $\mathcal{C}(h)$ depends on the location $\mathbf{s}_i$ and $\mathbf{s}_j$ only through the Euclidean spatial distance $h = ||\mathbf{s}_i - \mathbf{s}_j|| \in \mathbb{R}$. $\mathcal{C}(h)$ is defined by the Matérn function given by

$$\mathcal{C}(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\kappa h)^\nu K_\nu(\kappa h),$$

with K_ν denoting the modified Bessel function of second kind and order $\nu > 0$ (Abramowitz and Stegun, 1964). ν is usually kept fixed and measures the degree of smoothness and $\kappa > 0$ is a scaling parameter related to the range ρ i.e. the distance at which the spatial correlation becomes almost null.

The spatial surface in this model is estimated using a triangulated mesh of the spatial domain. Care must be taken when constructing the mesh since this can affect the modelling results although Cameletti et al. (2011) state that an increasingly finer mesh does not improve model estimates. Examples of spatio-temporal modelling using INLA and the SPDE approach can be found in Blangiardo et al. (2013) who give a description of the method followed by several examples.

1.3.2 Functional data analysis

In functional data analysis the data units of interest are curves representing continuous functions observed at discrete points rather than individual data points. The smoothing techniques described in the previous section are used to smooth data and subsequent analyses are performed on the smoothed curves. Analyses might also be applied to derivatives of the curves and examples of this can be found in Ramsay and Silverman (2006) where an example of growth curve data is used to show that additional insight can be gained by analysing the second derivative (or acceleration) curve rather than the original data. Ramsay and Silverman (2006) provide a good introduction to functional data analysis and Ramsay et al. (2009) discuss the implementation of functional methods in `R` using the `fda` package (Ramsay et al., 2014).

Many standard statistical techniques have a functional data equivalent such as linear regression Faraway (1997), principal components analysis (Henderson, 2006) and clustering (Henderson (2006) discuss hierarchical clustering and James and Sugar (2003) focus on model based clustering). Summary functions can also be calculated analogous to summary statistics such as the functional mean and functional standard deviation. For functional data $x_i(t)$ where i=1,...,N functions of time (t) the functional mean is the pointwise average of the N functions:

$$\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t).$$

The functional variance can be similarly calculated as

$$\text{var}_X(t) = \frac{1}{N-1}\sum_{i=1}^{N}[x_i(t) - \bar{x}(t)]^2,$$

and the standard deviation is the pointwise square root of the variance function. Miller et al. (2014) use additive models to estimate smooth temporal trends and seasonal patterns for LHA's and calculate the functional mean and error bands of ± 2 standard deviations to highlight LHA's that behave differently from average.

This section has described the structure of additive models and introduced the idea of functional mean and functional variance. These methods are applied to data from a

single LHA in Section 2.2.2, where there is also some discussion around incorporating spatial information into each of these methods.

1.4 Modelling more than one trend

The previous section described statistical methods that could be used to model flexible relationships between response and covariate data. In the case of water quality data, the aim of fitting an additive model might be to identify the average temporal trend or seasonal pattern for several locations. A trivariate interaction term could be used in the model to assess whether the temporal trend or seasonal pattern changes across space but Miller et al. (2014) discuss the computational difficulties of this approach when working with a large data set. The functional data approach could be used to estimate the mean temporal trend or seasonal pattern from curves estimated at several locations but this does not provide any information about changes across space. Instead, methods such as dynamic factor analysis or principal components analysis could be used to investigate whether more than one dominant temporal trend or seasonal pattern is present in the data and maps of results could be used to investigate how temporal trends and seasonal patterns change across or are grouped within space. These two methods are described below along with some discussion of choices that must be made when applying these methods. As with the additive model and functional data approach, a discussion of strengths and drawbacks is given in Chapter 2 (dynamic factor analysis) and Chapter 5 (principal components analysis).

1.4.1 Dynamic factor analysis

Dynamic Factor Analysis (DFA) is a dimension reduction technique where a linear combination of m estimated common trends is used to describe n time series, where $m < n$. The DFA model aims to estimate underlying common trends or latent factors among several time series and can be described within the state-space model framework (Harvey, 1990). The common trends are assumed to be temporally correlated and groups of similar time series can be identified by looking at plots of the factor loadings. DFA has been applied in many areas including psychology (Molenaar, 1985), economics (Geweke,

1976), fisheries (Zuur et al. (2003), Zuur et al. (2003), Erzini (2005)), sea surface temperature (Friedland and Hare, 2007), and hydrologic applications (Muñoz-Carpena et al., 2005).

The DFA model can be written as in (1.17), following the notation of (Zuur et al., 2007):

$$\mathbf{Y}_t = \mathbf{c} + \mathbf{A}\mathbf{Z}_t + \boldsymbol{\beta}\mathbf{X}_t + \boldsymbol{\varepsilon}. \tag{1.17}$$

Here $\mathbf{Y}$ is a $n \times 1$ matrix of recorded values for each of the n time series at time t, $\mathbf{c}$ is a $n \times 1$ matrix of intercept values for each time series, $\mathbf{A}$ is a $n \times m$ matrix of factor loadings, $\mathbf{Z}_t$ is a $m \times 1$ matrix of values for each of the m common trends at time t, $\boldsymbol{\beta}$ is a $n \times p$ matrix of regression parameters, $\mathbf{X}_t$ is a $p \times 1$ matrix of covariate values at time t and $\boldsymbol{\varepsilon}$ is an $n \times 1$ matrix of errors at time t. It is assumed that $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \sigma^2\mathbf{V})$. $\mathbf{V}$ is positive definite and can be represented generally as in (1.18).

$$\mathbf{V} = \begin{bmatrix} 1 & \nu_{1,2} & \nu_{1,3} & \dots & \dots & \nu_{1,N-1} \\ \nu_{2,1} & 1 & \nu_{2,3} & \dots & \dots & \nu_{2,N-2} \\ \nu_{3,1} & \nu_{3,2} & 1 & & & \nu_{3,N-3} \\ \vdots & \nu_{4,2} & \nu_{4,3} & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \nu_{N-1,N} \\ \nu_{N-1,1} & \nu_{N-2,2} & \nu_{N-3,3} & \dots & \dots & 1 \end{bmatrix} \tag{1.18}$$

If $\mathbf{V}$ has a diagonal structure this implies the time series being modelled are independent and have the same variance if the diagonal elements are equal, and have different variances if the diagonal elements are different. A non-diagonal $\mathbf{V}$ implies there are relationships between the time series that are not captured by the common trends or covariates. Ideally, the explanatory variables would capture most of the variablity in the time series since this is the easiest model to interpret (Zuur et al., 2007). AIC is used to select the 'best' model in terms of the number of common trends, the subset of explanatory variables that are most relevant and the structure of the covariance matrix $\mathbf{V}$. The common trends z_t in $\mathbf{Z}$ in the (1.17) are random walks:

$$z_t = z_{t-1} + \eta_t \text{ and } \eta_t \sim \mathrm{N}(0, \sigma_\eta^2).$$

Zuur et al. (2003) describe in detail how the EM algorithm (Dempster et al., 1977) can be used to estimate model parameters. The EM algorithm is an iterative method used to estimate the maximum likelihood value of model parameters. In the expectation (E) step the expectation of the log likelihood function is calculated using the current estimates of model parameters and in the maximisation (M) step the expected log likelihood function of the E step is maximised. The iterations continue until convergence criteria are met. Full mathematical details of the DFA model fitting procedure can be found in Zuur et al. (2003). During model estimation the common trends are not estimated sequentially (as in principal components analysis, described later) so if two common trends are estimated, trend 1 is not necessarily the dominant pattern. The dominant pattern can be determined by fitting a model with a single common trend and a second model with two common trends and compare plots of the common trends for the two models. The dominant trend in the second model is the one which is most similar to the common trend estimated in the first model.

DFA models can be fitted using Brodgar software (`http://www.brodgar.com/`) or the `MARSS` package in `R` (Holmes et al., 2012). Harvey (1990) show that if the DFA model in (1.17) and (1.19) is not constrained then the model is unidentifiable. Zuur et al. (2003) found that with one of the contraints the EM algorithm is not robust and takes a long time to converge. They solve this by constraining the common trends to have a mean of zero for all time points. This means that the intercept values in $\mathbf{c}$ in (1.17) are the average level of $\mathbf{y}_t$ relative to $\mathbf{A}(\mathbf{x}_t - \bar{\mathbf{x}})$. Holmes et al. (2012) found that this can cause errors in the EM algorithm and instead estimate the DFA model on mean centered data and fix all of the elements in $\mathbf{c}$ to be zero. In Chapter 2 both Brodgar and `MARSS` are applied to water quality data. Brodgar is used to estimate common temporal patterns among fewer than 30 time series since this is the most time series that can be modelled with this software but was used in application of DFA due to its ease of implementation. `MARSS` was used to implement DFA for more than 30 time series.

DFA incorporates unexplained spatial correlation through the covariance matrix. This method does not explicitly include the river network structure but in its most general

non-diagonal form allows the covariance between two monitoring sites to reflect relationships not captured by the common trends or explanatory variables. DFA has been adapted within a Bayesian framework for spatial data by Lopes et al. (2008), Lopes et al. (2011) and Strickland et al. (2011).

Zuur et al. (2007) recommend that data are de-seasonalised before modelling common trends using DFA to avoid confounding between temporal trend and seasonal pattern. For example, monthly data recorded for several years can be de-seasonalised by subtracting the monthly average over all years from each monthly observation. Another option would be fit an additive model and extract the trend term or seasonal pattern along with corresponding model residuals and model these using DFA. Molenaar et al. (1992) and Alonso et al. (2011) have adapted DFA for time series with a seasonal component.

1.4.2 Principal components analysis

Principal components analysis (PCA) (Hotelling, 1933; Pearson, 1901) is a dimension reduction technique where the aim is to replace p correlated variables with $k < p$ uncorrelated variables. The k uncorrelated variables are known as *principal components* and are constructed as a linear combination of the p original variables. The weights used to calculate the linear combination are known as *loadings*. The k principal components are constructed to retain most of the variation present in the original p variables. Abdi and Williams (2010) provide a brief introduction to PCA while Jolliffe (2002) gives a detailed discussion of many of the issues to be considered when performing PCA and discusses many extensions of the method depending on the type of data being analysed. PCA is applied in many different research areas but Demšar et al. (2013) gives a detailed review of PCA applied to spatial data and Hannachi et al. (2007) review PCA methods for atmospheric science where data are spatiotemporal.

PCA is usually performed on multivariate data where rows of a data matrix correspond to observations and columns are values of different variables recorded for each observation. For example, observations could be samples of water taken from different locations on a stream network and variables could be chemicals in the water related to water quality. Some recent examples of this include Wilbers et al. (2014), Shrestha et al. (2008), Bengraïne and Marhaba (2003) and Petersen et al. (2001). Variables recorded at several time points for each monitoring site would mean that the data have three

dimensions: location, time, variable of interest (see Jolliffe (2002) for an introduction to PCA for this type of data). Richman (1986) introduces the idea of PCA modes which are the six possible combinations of any two of these three dimensions. Since the data provided by the Environment agency for this thesis contain values of a single variable (TON) recorded for 21 years at many monitoring sites this section will consider T- and S-mode PCA. Chapter 5 will describe modifications to T- and S-mode PCA that account for known spatial and temporal structure in the data.

The data matrix for T-mode PCA has p rows representing monitoring sites and n columns representing equally spaced time points and the aim is to identify time points with similar spatial patterns. Maps of scores show the spatial pattern represented by a given component and loadings indicate at what time points the spatial patterns occur. T-mode PCA is useful when the spatial pattern is expected to change over time and there is interest in identifying the different spatial patterns, represented by the scores. Some recent examples include Zhang et al. (2012) (sea level pressure), Hidalgo-Muñoz et al. (2011) (rainfall) and Barreira and Compagnucci (2011) (sea ice concentration anomalies).

In S-mode PCA the variables in the columns of the data matrix $\mathbf{X}$ are sites and the observations in the rows of $\mathbf{X}$ are time points. S-mode PCA aims to find groups of monitoring sites that have similar behaviour over time. Monitoring sites have similar behaviour if for component k a group of monitoring sites have loadings of the same sign and magnitude. Plotting the scores in time order gives the temporal pattern that is shared by the group of sites defined by each component. S-mode PCA is known as as empirical orthogonal functions (EOF's) in the climatology literature. S-mode PCA has been used to find regions with similar temporal patterns for precipitation (Ehrendorfer (1987), Neal and Phillips (2009)), surface wind (Jiménez et al., 2008) and streamflow (Kahya et al., 2008).

PCA is implemented as follows. Let $\mathbf{X}$ be an $n \times p$ data matrix where the n rows are observations and the columns are p different and usually correlated variables recorded for each of the n observations. $\mathbf{X}_i$ is the i^{th} row of $\mathbf{X}$. PCA aims to find loadings $\mathbf{v}_1$ to maximize the variance of the $i = 1, ..., n$ first principal component (z_{i1}). The variance is calculated as

$$\frac{1}{n-1}\Sigma_{i=1}^{n}\left(z_{i1}-\bar{z}_{.1}\right)^2, \tag{1.19}$$

where $z_{i1} = \mathbf{X}_i\mathbf{v}_1$, subject to the normalization constraint $\mathbf{v}_1^\top\mathbf{v}_1 = 1$ and $\bar{z}_{.1}$ is the mean of the first principal component over n observations. Next let $z_{i2} = \mathbf{X}_i\mathbf{v}_2$ and choose $\mathbf{v}_2^\top$ to maximize the sample variance of z_{i2} subject to the normalization constraints $\mathbf{v}_2^\top\mathbf{v}_2 = 1$ and $\text{cov}(z_{i1}, z_{i2}) = 0$ so that the two components are orthogonal. In this way k principal components can be defined and the loadings for k components stored in $p \times k$ matrix $\mathbf{V}$. The scores for k components can be represented in the $n \times k$ matrix $\mathbf{Z} = \mathbf{XV}$. Scores are therefore a weighted linear combination of the variables and these scores are often used in place of the original data matrix in other applications such as regression or clustering (Jolliffe (2002), Demšar et al. (2013)). The dimensionality of the original $n \times p$ data matrix has been reduced to $n \times k$ while retaining much of the variance in $\mathbf{X}$.

PCA can be performed using an eigen decomposition of the sample covariance (or correlation) matrix, singular value decomposition (SVD) of the data matrix or by iterative methods.

Eigen decomposition

PCA can be performed by an eigen decomposition of the sample covariance (or correlation) matrix $\mathbf{S}$ where

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^\top\mathbf{X},$$

and eigenvectors contained in the matrix $\mathbf{V}$ are found such that

$$\mathbf{SV} = \mathbf{\Lambda V}.$$

Here, $\mathbf{\Lambda}$ is a $p \times p$ diagonal matrix and each element is an eigenvalue of $\mathbf{S}$, sorted in descending order of magnitude. The eigenvectors that form the p columns of $\mathbf{V}$ are the loadings, and the eigenvalues in $\mathbf{\Lambda}$ are the variances of the scores in (1.19). Since

the eigenvalues represent variances they are non-negative and therefore $\mathbf{S}$ is positive semi-definite.

Singular value decomposition

Singular value decomposition (SVD) is a decomposition of the rectangular data matrix $\mathbf{X}$ such that

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \tag{1.20}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors of $\mathbf{X}$ respectively. $\mathbf{D}$ is a diagonal matrix containing the singular values and $\mathbf{D} = \mathbf{\Lambda}^{\frac{1}{2}}$. In SVD the loadings are the columns of $\mathbf{V}$ and the scores are $\mathbf{UD}$.

Iterative methods

A principal components analysis simultaneously calculates as many principal components as there are variables in the data. For data sets with a large number of variables the need to compute the covariance matrix for eigen decomposition is computationally prohibitive and although SVD can be used directly on the data matrix this is still computationally demanding for large data sets (Jolliffe, 2002). Von Storch and Zwiers (2001) suggest swapping the space and time dimensions of the data matrix if the number of observations is much smaller than the number of variables. Alternatively, since PCA usually involves discarding several components that explain little of the variation contained within the data it might be more practical to calculate only the first few principal components. Iterative methods begin by calculating the first principal component with loadings that maximise the variance of the scores. This component is then removed from the data, a process called *deflation* (Demšar et al., 2013) and the next component is found on the deflated data. Algorithms for this process are presented by Vines (2000) and Partridge and Calvo (1998). Van den Dool et al. (2000) and Baldwin et al. (2009) use iterative techniques on climate data while Pinto da Costa et al. (2011) and Abraham and Inouye (2014) present algorithms to perform PCA on genome data that extracts only the first few principal components.

PCA is usually performed on centered data where the p^{th} column of $\mathbf{X}$ is centered by subtracting the mean of each column from the n observations in the column. The data might also be scaled by dividing observations in the p^{th} column by the column standard deviation σ_p and is useful when variables are recorded on different scales. Without this scaling the first component would be dominated by the variable with the largest variance and similary with subsequent components. If data are recorded on the same scale such as in S- or T- mode PCA then it is not necessary to scale the data unless the analysis is likely to be dominated by a few highly variable monitoring sites. SVD of centered (and scaled) data is equivalent to eigen decomposition of the covariance (correlation) matrix.

1.4.2.1 Choosing how many components to retain

Once PCA has been carried out it is necesary to choose how many components to retain for interpretation or reconstruction of the data. A complete decomposition of the data with p variables will yield p components but since PCA is often used for dimension reduction it is necessary to choose $k < p$ components to retain so that the $n \times k$ scores matrix is of smaller dimension than the original data matrix. The literature contains many suggestions for choosing k but there is no single method that can be declared as best. Some of these methods are briefly described below.

The scree plot (Cattell, 1966) is produced by plotting the eigenvalues from the decomposition against component number and looking for an 'elbow' in the plot where the slope changes from steep to flat. In this case k is chosen as the number of components before the elbow. This is a simple way of choosing the number of components to retain but is subjective and difficult if the magnitude of the eigenvalues decreases smoothly. Instead, the proportion of variance in the original data explained by component k is $\frac{\Lambda_k}{\Sigma_{k=1}^{p}\Lambda_k}$ and k is chosen so that the cumulative proportion of variance explained by the first k components reaches a specified threshold. Similar to the scree plot the threshold is arbitrarily chosen and this method can result in many components being retained that explain only a very small proportion of the variance in the data so as to meet the threshold. Another option is to retain k components such that the eigenvalues of the first k components are all greater than the average eigenvalue i.e.

$$\Lambda_k > \frac{1}{p}\Sigma_{k=1}^{p}\Lambda_k. \tag{1.21}$$

If PCA has been performed on the correlation matrix then this amounts to retaining components with eigenvalues greater than 1. More complex methods exist for choosing the number of components to retain such as the broken-stick model based on the expected distribution of eigenvalues (Frontier, 1976). Descriptions of this and comparison with several methods for choosing k can be found in Peres-Neto et al. (2005) while Valle et al. (1999) develop a selection criterion based on the variance of reconstruction error.

1.4.2.2 Interpreting the results

After choosing the number of components to retain the loadings and scores are usually inspected with the aim of providing meaningful interpretation of the results and plots are often used for this. The loadings in a single component indicate which variables contribute most to that component as these variables will have the highest loadings. Loadings can be positive or negative and so, for example, two variables whose loadings are of the same sign and similar magnitude are positively correlated but if they are of similar magnitude and have opposite signs then they are negatively correlated. Variables whose loadings are close to zero contribute little to that component. Ideally, loadings will reveal groups of variables representing different aspects of the data that can be thought of as latent variables. Inspection of the scores reveals groups of observations that are similar in terms of the latent variables. If more than two components are retained then it can be difficult to inspect several maps of loadings or scores (depending on the PCA mode used). Harris et al. (2011) and Harris et al. (2015) use glyph plots where, for each monitoring site a glyph is plotted on a map. The glyph consists of lines radiating from a central point, one for each component retained, and the length reflects the relative magnitude of the loadings or scores. The glyph plot aids visual inspection of several components at once. The biplot (Gabriel, 1971) is a commonly used and useful tool to interpret pairs of principal components. Scores are plotted as points and loadings as arrows and groups of scores indicate similar observations. Scores in the same quadrant as arrow heads indicate observations with high values for a particular variable. Detailed information about biplots can be found in Gower et al. (2011) and Jolliffe (2002).

Many attempts have been made to aid interpretation of principal components. Rotation of the loadings, originating in factor analysis, is commonly used to force some simplified structure on the loadings by forcing absolute values of loadings to be high on a single axis rather than medium values spread over several axes. Richman (1986) discusses rotation of loadings in great detail. Rotation can be orthogonal or oblique (relaxing the orthogonality constraint). A commonly used rotation is varimax (Kaiser, 1958) which aims to maximise the variance of squared loadings. Vines (2000) notes that varimax rotation tends to "share out variance equally among components, losing the ordering of the components based on the variance explained". Huth (1996) discusses some methodological issues when using PCA specifically for the classification of circulation patterns and show that oblique rotation of components is more suitable than orthogonal rotation or no rotation.

One way to simplify the loadings is to restrict the values to a small number of values. An extreme version of this is found in Hausman (1982) who constrains loadings to take values of ± 1 or 0. Similarly, Vines (2000) present an algorithm that constrains loadings to take integer values but does not restrict the magnitude. Jolliffe and Uddin (2000) introduces the Simplified Component Technique (SCoT) that combines variance maximisation of the scores with simplification into a single step rather than PCA followed by rotation. If the simplicity parameter is zero then the resulting principal components (loadings) are identical to the PCA solution and increasing the value of the parameter simplifies the structure of the components until a value of one means that each simplified component is identical to one of the original variables, with zero loadings for all other variables in that component. Jolliffe (2002) states that the choice of simplicity criterion is usually less important than the choice of k, the number of components to retain. Jolliffe et al. (2003) develop a simplified PCA technique SCoTLASS (Simplified Component Technique - LASSO) that does not involve rotation of loadings and is based on the LASSO (least absolute shrinkage and selection operator, (Tibshirani, 1996)). The idea is to find loadings that successively maximise the variance of the scores while imposing the constraint that the sum of p loadings for a particular component is less than or equal to some tuning parameter t. For $t \geq \sqrt{p}$ this is the PCA solution while for $t = 1$ there is exactly one non-zero loading in each component. There is no solution for $t < 1$. Jolliffe and Uddin (2000) explain that none of these simplifications are able to produce orthogonal loadings as well as uncorrelated scores, a key feature of PCA.

Demšar et al. (2013) note that SCoTLASS has high computational cost and that more efficient algorithms for sparse principal components have been developed such as Zou et al. (2006) and Shen and Huang (2008). Guo et al. (2015) use a weighted version of the LASSO method for sparse principal components analysis in Zou et al. (2006) to perform PCA on imaging data which is high dimensional (large p) but with a small number of observations (small N).

This section has described the technical detail and introduced notation for two statistical approaches that can be used to identify common temporal patterns in spatiotemporal data. Dynamic factor analysis is applied to data provided by the EA in Section 2.2.1 and Section 2.2.2. Principal components analysis, specifically when applied to spatiotemporal data, is considered in Chapter 5. The strengths and weaknesses of these two approaches are discussed in the relevant sections of this thesis.

1.5 Statistical issues with environmental data

Henderson (2006) states that:

> Water quality trend analysis is often required to confront data that are: highly variable; irregularly collected and subject to missing values; positively skewed; censored at detection limits; subject to seasonal variation; temporally correlated and potentially confounded with other covariates.

and these issues are found in many environmental examples. This section will address the problems of data that are irregularly collected, subject to missing values, censored at detection limits, subject to seasonal variation and temporally correlated as well as data that are spatially correlated. Highly variable data has already been addressed by taking a natural log transformation of TON observations and covariates are discussed in Chapter 4 within the context of spatial rather than temporal trends.

1.5.1 Autocorrelation

Many statistical methods assume observations are independent but this is often not valid for spatio-temporal data. Data recorded over time can be correlated as can data recorded in space where it is assumed that observations close together in time or space

are more similar than observations far apart. When analysing spatio-temporal data it is necessary to take into account correlation between observations as not doing so will lead to underestimation of standard errors of model parameters (Cressie, 1993). In this situation autocorrelation is viewed as a nuisance parameter since it affects the estimation of model parameters and must be accounted for, even if there is no interest in using the autocorrelation structure for predictive purposes. For example, Cressie (1993) shows that correlated data have fewer effective degrees of freedom than independent data which can cause spuriously significant parameters while estimating trends and seasonal patterns meaning that statistical models might be developed with more statistically significant parameters than are actually required. Giannitrapani et al. (2006) note that the main effect of autocorrelation is on standard errors and model comparisons rather than affecting estimation of the trend and seasonal pattern and show how the residual sum of squares, degrees of freedom and standard errors in additive models can be adjusted for autocorrelation. The particular form of autocorrelation is therefore only of interest in that it is used to make adjustments to model selection procedures.

It might also be that there is interest in identifying and modelling the autocorrelation structure explicitly. Examples of this can be found in geostatistical examples where the spatial autocorrelation function is first identified and then used to make predictions at unsampled locations. Spatial autocorrelation models for river networks are discussed in detail in this thesis. Methods used to estimate temporal and spatial autocorrelation are described in this section and estimation of spatial autocorrelation specifically for river networks are discussed in Chapter 3.

First, statistical approaches to model temporal autocorrelation are described in Section 1.5.1.1 and this is followed by a description of statistical methods used to model spatial autocorrelation in Section 1.5.1.2.

1.5.1.1 Temporal

Data recorded over time (Y_t) are known as time series. A good introduction to time series methods can be found in Brockwell and Davis (1996) which is the main reference for this section. A single time series might be represented as the sum of a trend component (m_t), a seasonal component (s_t) and a stationary white noise process (ε_t) as in (1.22).

$$Y_t = m_t + s_t + \varepsilon_t, \qquad \text{where} \qquad t = 1, ..., n. \tag{1.22}$$

A time series is usually modelled by removing the trend and seasonal components and applying a stationary model to the residuals $\boldsymbol{\varepsilon}$. A time series process $\{X_t : t = 1, ..., n\}$ is *strictly stationary* if the joint distribution $f(X_1, ..., X_m)$ for $m < n$ is identical to the joint distribution $f(X_{1+\tau}, ..., X_{m+\tau})$ where τ is a time lag. In other words, shifting the time series by τ has no effect on its joint distribution. This assumption is rather restrictive and so in practice it is usually assumed that a time series process is *weakly stationary*. This means that the mean and variance are constant over time and that the autocovariance and autocorrelation functions depend only on the lag τ (these will be defined later). For weak stationarity it is assumed that the first two moments of the distribution are constant over time whereas strong stationarity also assumes that higher moments are constant over time.

One commonly used stationary process model is the AutoRegressive or AR(p) model written for response of interest y_t as

$$y_t = \sum_{i=1}^{p} \delta_i y_{t-i} + e_t, \tag{1.23}$$

where δ_i are coefficients and e_t is a white noise process with mean of zero and constant variance. Methods for estimating p will be discussed later. The coefficients δ_i can be estimated using methods such as the Yule-Walker equations, ordinary least squares (OLS) or maximum likelihood. It is worth noting that if the data are normally distributed then the OLS and maximum likelihood estimates will be the same. In the AR(p) model correlation is accounted for by expressing the current value y_t as a finite linear combination of p earlier values. Alternatively, y_t can be expressed in a Moving Average or MA(q) model in terms of previous and current process innovations as

$$y_t = e_t + \sum_{j=1}^{q} \theta_j e_{t-j}, \tag{1.24}$$

where θ_j are coefficients and e_t as before. The coefficients θ_j are estimated using conditional least squares since OLS would required maximising

$$\sum_{t=1}^{n} \left[y_t - \mu - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q} \right]^2 ,$$

and the parameters θ_j depend on the unknown random variables e_{t-j}. Combining (1.23) and (1.24) gives an AutoRegressive Moving Average or ARMA(p,q) model:

$$y_t = \sum_{i=1}^{p} \delta_i y_{t-i} + e_t + \sum_{j=1}^{q} \theta_j e_{t-j}. \tag{1.25}$$

Parameters of the ARMA(p,q) model are estimated using conditional least squares. Fitting AR(p), MA(q) and ARMA(p,q) models involves first removing any trend or seasonality before modelling the residuals with a short term correlation model and allows the shape of the trend and/or seasonal pattern to be modelled explicitly. Another approach involves modelling the trend, seasonality and correlation simultaneously using an ARIMA(p,d,q) model where the trend is removed by d-order differencing but not explicitly modelled. If Y_t denotes the value of the time series at time t then the first difference is $\Delta Y_t = Y_t - Y_{t-1}$ and this will remove a linear trend. Period differencing can be applied for a given period d as $\Delta^d Y_t = Y_t - Y_{t-d}$ where d is the frequency of the data. For example, monthly recorded data would have $d = 12$.

A further extension of this is the SARIMA$(p, d, q) \times (P, D, Q)$ model which can be used to remove trend and seasonality. This model is an ARIMA(P, D, Q) model with ARIMA(p, d, q) residuals. In the SARIMA model differencing is used to obtain the detrended series ΔY_t and then period differencing is used to remove seasonality. The residuals can then be modelled using an ARMA(p, q) process. If more than one time series is to be modelled then Vector ARMA or VARMA models can be used.

Once trend and seasonality have been removed either by explicit estimation or differencing it is necessary to choose which order of AR(p) or MA(q) model to fit. The sample autocorrelation function (ACF) can be used to identify appropriate values of p or q. The sample autocorrelation function for time series y_t with sample mean $\bar{y}$ at lag τ can be written as

$$\hat{\rho}_\tau = \frac{\sum_{t=1}^{n-1} (y_t - \bar{y})(y_{t+\tau} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2},$$

which has the sample autocovariance function at lag τ as the numerator and the autocovariance function at lag 0 as the denominator. The ACF is the correlation between the series y_t and itself at several different lags. The ACF is plotted against several values of τ and this is known as a correlogram. If the ACF shows significantly non-zero values up to and including lag q this suggests an MA(q) model might be appropriate. If 95% confidence intervals are used then it can be expected that around 5% of the q values will give a significantly non-zero ACF value by chance and careful thought should be given to significant results for large lags unless knowledge of the data suggests this might be sensible. The ACF is always 1 at lag 0. A smooth decline in the ACF or tapering (smooth decline that alternates between positive and negative values) indicates an AR model might be more appropriate. To identify the order of an AR(p) model the partial autocorrelation function (PACF) can be used. The PACF at lag d shows the autocorrelation between observations y_t and y_{t+d} not accounted for by lags 1,...,d-1 so no PACF value is produced at lag 0. Significantly non-zero PACF values at lag d indicate an AR(d) model is appropriate while a smooth decline in the PACF suggests an MA model should be considered. MA and AR models of various orders can be compared using the AIC value, described earlier in this chapter.

Often an AR(1) model is sufficient to capture autocorrelation in water quality time series. For example Clement et al. (2006) use an AR(1) process as the temporal component in a spatio-temporal model of dissolved oxygen in Belgian rivers and Andrés Houseman (2005) use a first order autoregressive model for depth data in Boston Harbour for unequally spaced observations.

1.5.1.2 Spatial

It is common for environmental data to have a spatial component and statistical modelling in this case aims to capture trends in space rather than time. There are three broad categories of spatial data: geostatistical, areal and point process. Key texts for spatial statistics include Cressie (1993) (spatial data), Banerjee et al. (2004) (bayesian spatial statistics), Cressie and Wikle (2011) (spatio-temporal data), Diggle et al. (1983) (point processes) and Chilès and Delfiner (1999) (geostatistical). This thesis focuses on geostatistical data and so the following section will describe models for geostatistical data; areal and point process models will not be discussed any further. References

have been provided and should be consulted for a description of areal and point process models.

Geostatistical data models

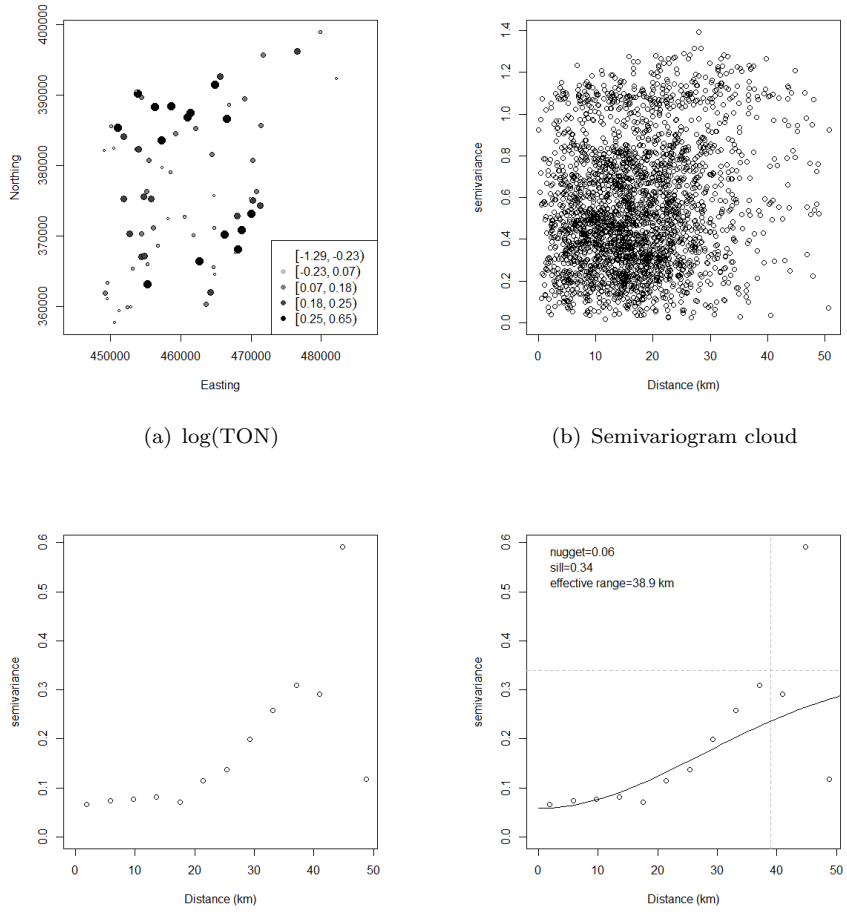
Suppose a stochastic process $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ where D is a subset of r-dimensional Euclidean space. In the case of spatial data, r is 2, representing a co-ordinate system. $Y(\mathbf{s})$ often represents the level of some response variable recorded at site $\mathbf{s}$. It is often of interest to make inferences about the spatial process $Y(\mathbf{s})$ at unobserved locations. It is usually assumed that spatial autocorrelation between sites depends only on the distance between them.

Spatial autocorrelation is usually quantified using the variogram. First, the variogram is plotted and an appropriate covariance model is fitted to estimate the parameters of the spatial covariance function. In practice the empirical semi-variogram is used to estimate population level covariance parameters from a sample of data, where the semi-variogram values are the variogram values/2. Since the variogram represents variability between pairs of spatially located points, the semi-variogram can be thought of as representing variability at a single point. Following the notation in Banerjee et al. (2004), the empirical semi-variogram estimator $\hat{\gamma}(\mathrm{h})$ for a given distance h is defined to be

$$\hat{\gamma}(h) = \frac{1}{2\mathrm{N(h)}} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(h)} [Y(\mathbf{s}_i) - Y(\mathbf{s}_j)]^2,$$

where $N(h)$ is the set of pairs of points $\mathrm{s}_i, \mathrm{s}_j$ (such that $i = j = 1, ..., n$ where n is the number of monitoring sites), separated by distance h and $|N(h)|$ is the number of pairs of points in the set. $Y(\mathrm{s}_i)$ are realisations of stochastic process $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ where D is a fixed subset of r-dimensional Euclidean space. In the spatial context, r is usually 2 (Eastings and Northings) or 3 (Eastings, Northings and altitude above sea level for example). For example, $Y(\mathbf{s})$ might represent total oxidized nitrogen measured at a number of monitoring sites (Figure 1.4(a)).

$\hat{\gamma}(\mathbf{h})$ plotted against $\mathbf{h}$ gives the semivariogram cloud (Figure 1.4(b)). It can be computationally demanding to produce this plot and difficult to observe the relationship between $\hat{\gamma}(\mathbf{h})$ and $\mathbf{h}$ so instead the data are 'binned' as in Figure 1.4(c), with selected



(a) log(TON)

(b) Semivariogram cloud

(c) Binned semivariogram

(d) Fitted semivariogram model: Gaussian

FIGURE 1.4: Semivariogram example

values of **h** as midpoints for the bins. The number of bins must be carefully chosen - too few bins can obscure the relationship between **h** and $\hat{\gamma}(\mathbf{h})$ and too many bins can result in bins representing small numbers of pairs of points. It is often the case that $N(\mathrm{h})$ is small at large values of h and the binned semivariogram will exhibit erratic behaviour. It is recommended (Journel and Huijbregts, 1978) that for each bin $N(\mathrm{h}) > 30$.

The next step is to choose a semivariogram model to fit to the 'data', where the data are now the binned values of $\hat{\gamma}(\mathbf{h})$. Figure 1.4(d) shows the fitted line for a Gaussian semivariogram model for the binned semivariogram in Figure 1.4(c). Common choices for the semivariogram model are Exponential, Gaussian and Spherical (see Table 1.2), with Exponential and Gaussian both being special cases of the Matérn function. The semivariogram model contains three parameters: *nugget, sill* and *range*. The nugget τ^2

represents variability at smaller distances than the smallest binned distance or measurement error (i.e. $h = 0$) while the sill $\tau^2 + \sigma^2$ (where σ^2 is known as the *partial sill*) is the asymptotic value of the semivariogram representing variability at large distances. The range parameter $1/\phi$ is the distance h at which $\gamma(h)$ first reaches the sill, beyond which observations are assumed to be independent. For some semivariogram models such as those belonging to the Matérn class, the range is referred to as the *effective range* which is the distance h beyond which spatial correlation is less than 0.05. In spatial statistics distance is usually Euclidean (straight line) distance between two points. Chapter 3 discusses spatial statistics for non-Euclidean distance.

Model	Semivariogram		
Spherical	$\gamma(h) =$	$\tau^2 + \sigma^2$	if $h \geq 1/\phi$
		$\tau^2 + \sigma^2\{\frac{3\phi h}{2} - \frac{1}{2}(\phi h)^3\}$	if $0 < h \leq 1/\phi$
		0	otherwise
Exponential	$\gamma(h) =$	$\tau^2 + \sigma^2(1 - \exp(-\phi h))$	if $h > 0$
		0	otherwise
Gaussian	$\gamma(h) =$	$\tau^2 + \sigma^2(1 - \exp(-\phi^2 h^2))$	if $h > 0$
		0	otherwise

TABLE 1.2: Some commonly used variogram models.

Banerjee et al. (2004) state that fitting the semivariogram model 'has traditionally been as much art as science' and note that historically the model has been fitted 'by eye' however more formal model selection procedures can be used such as weighted least squares (WLS) (Cressie, 1985), maximum likelihood (ML) or restricted maximum likelihood (REML) (see Cressie (1993) for further details). Rathbun (1998) discusses the use of WLS and REML for spatial modelling.

As with all statistical models, certain assumptions must be satisfied in order that the model is valid. The assumptions for spatial modelling are:

- $Y(\mathbf{s})$ is stationary
- $Y(\mathbf{s})$ is ergodic
- $Y(\mathbf{s})$ follows a multivariate normal distribution
- $Y(\mathbf{s})$ is isotropic

It is assumed (following the notation in Banerjee et al. (2004)) that the spatial process has a mean $\mu(s) = E(Y(s))$ and that the variance of $Y(s)$ exists for all $\mathbf{s} \in D$. $Y(\mathbf{s})$ is said to be Gaussian if $\mathbf{Y} = (Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n))^\top$ follows a multivariate normal distribution. The process is said to be *strictly stationary* if for any set of $n \geq 1$ sites $\{\mathbf{s}_1, ..., \mathbf{s}_n\}$ and any $\mathbf{h} \in \mathcal{R}^r$, the distribution of $(Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n))$ is the same as the lagged distribution $(Y(\mathbf{s}_1 + \mathbf{h}), ..., Y(\mathbf{s}_n + \mathbf{h}))$.

The process is said to be *weakly stationary* or *second order stationary* if the process has a constant mean i.e. $\mu(\mathbf{s}) = \mu$ and the covariance between two sites can be described by a covariance function $C(\mathbf{h})$ that depends only the separation vector $\mathbf{h}$:

$$\text{Cov}(Y(\mathbf{s}), Y(\mathbf{s} + \mathbf{h}) = C(\mathbf{h}).$$

Intrinsic stationarity is where the process has a constant mean and the variance of the increments $Y(\mathbf{s}_i) - Y(\mathbf{s}_j)$ is a function of vector of distances $\mathbf{h}$. This function is denoted by the variogram $2\gamma(\mathbf{h})$ and semivariogram $\gamma(\mathbf{h})$.

The covariance function $C(\mathbf{h}) = \text{Cov}(Y(\mathbf{s}), Y(\mathbf{s} + \mathbf{h}))$, when plotted, is known as the *covariogram* and the relationship between the variogram and covariogram can be shown as

$$\begin{aligned} 2\gamma(\mathbf{h}) &= \text{var}(Y(\mathbf{s} + \mathbf{h}) - Y(\mathbf{s})) \\ &= \text{var}(Y(\mathbf{s} + \mathbf{h})) + \text{var}(Y(\mathbf{s})) - 2\text{cov}(Y(\mathbf{s} + \mathbf{h}), Y(\mathbf{s})) \\ &= C(\mathbf{0}) + C(\mathbf{0}) - 2C(\mathbf{h}) \\ &= 2[C(\mathbf{0}) - C(\mathbf{h})] \end{aligned}$$

and so $\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$.

The process is *ergodic* if $C(\mathbf{h}) \to 0$ as $||\mathbf{h}|| \to \infty$, where $||\mathbf{h}||$ is the length of the $\mathbf{h}$ vector. This is analogous to the idea that the standard error of a sample mean $\to 0$ as the sample size $n \to \infty$ only in the spatial average case n is the number of sampling locations rather than repetitions at a single location.

The assumption of *isotropy* means that $\gamma(\mathbf{h})$ is the same in all directions. Typically this assumption can be verified using a multi-directional variogram to assess if spatial

autocorrelation is of the same degree in all directions. This assumption is not valid when considering a response such as point source pollution in a river network where the relationship between pairs of monitoring stations is likely to be influenced by the direction in which water flows between the two points. Valid covariance models for this situation have recently been developed and are discussed in detail in Chapter 3.

Once a valid spatial covariance model has been selected, it can be used to predict the value of the variable Y at an unobserved location $Y(s_0)$, using the data in $\mathbf{Y} = (Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n))^\top$. This is commonly referred to as *kriging* (Matheron, 1963). Kriging allows prediction maps to be calculated across a whole geographical area where predictions are based on the spatial relationships defined by the covariance function. *Ordinary kriging* is when predictions are based on no covariates and *Universal kriging* is when covariates are available. *Block kriging* is a kriging method where the average expected value in an area around an unobserved location is predicted rather than the estimated exact value of a single prediction point. Block kriging provides smooth interpolated predictions. Kriging is essentially a weighted spatial average of the observed sites where greater weight is given to observations nearer the prediction site than those further away. Murphy et al. (2010) compare several kriging methods for water quality data.

Modelling spatial correlation among data collected on river networks provides particular challenges since the assumption that monitoring sites close to each other in space behave in a similar way might not hold. For example, a monitoring site located on a small tributary river only a short distance from a monitoring site on the main stem of the river might record very different values for water quality parameters. Spatial covariance functions are based on Euclidean distance between monitoring sites which might not be the best way to describe distance between monitoring sites located on river networks. The assumption of isotropy is likely to be violated since water flows in one direction and so it is not reasonable to expect observations recorded at upstream sites to be influenced by observations at downstream sites. It might also be that observations from monitoring sites not connected by water flow are correlated due to sharing runoff from land surrounding the monitoring sites meaning that a combination of covariance functions based on river network structure and Euclidean distance might be necessary to describe spatial relationships across a whole river network. These problems have been addressed by the work in ver Hoef et al. (2006), Peterson and ver Hoef (2010) and the covariance functions developed are described in detail in Chapter 3.

1.5.2 Missing data

This section will describe the mechanisms by which data can be 'missing', followed by a discussion of the impace of missing values throughout this thesis and why missingness may or may not be a problem. Next, some approaches that can be used to replace missing values with a sensible numeric value (imputation) are described. Finally, statistical methods that can accommodate missing values, specifically in the context of principal components analysis, are discussed.

Missing data occur when a response is not recorded for a subject on a variable (or variables in the case of multivariate data). When considering environmental data, monitoring equipment can break down causing gaps in the data and monitoring networks are altered over time by adding or removing monitoring sites. Standard statistical methods are often designed to be applied to complete data sets such as time series methods where it is assumed that data are recorded at regularly spaced intervals. If some observations are missing then the data must be completed in some way or the statistical methods modified to account for missing data. Rubin and Little (2002) provide a detailed discussion of the implications of missing data and how to adapt several statistical methods to account for missingness. The mechanism that leads to missing data should be considered before analysing incomplete data and Rubin (1976) define these mechanisms based on the relationship between missingness and the underlying values in the dataset. Specifically data can be:

- Missing completely at random (MCAR): the probability of a value being missing is not related to the observed or missing values.
- Missing at random (MAR): the probability of a value being missing is only related to the observed values but not the unobserved values.
- Missing not at random (MNAR): there are systematic and informative reasons why data are missing and these should be carefully considered and investigated.

Missing data are encountered throughout this thesis as values missing in time. The statistical methods applied in Chatper 2 (dynamic factor analysis, additive models and functional data analysis) can all accommodate data with missing values. The work in Chapters 3 and 4 considers nitrate levels at a single time point so missing data are not a problem here. Missing data are considered explicitly in this thesis in Chapter 5 where

principal components analysis (PCA), introduced earlier in this chapter, is applied to spatiotemporal data. Some attention has been given in the literature to missing data specifically within the context of PCA. PCA requires a data set or covariance matrix with no missing values. If missing values are present then complete case analysis can be used so that only observations with values recorded for all variables are analysed. Although easy to implement, this can result in working with a greatly reduced data set and produce biased results depending on the mechanism by which missingness is caused. Ilin and Raiko (2010) describes various approaches for applying PCA in the presence of missing values and Nelson et al. (1996) describe methods for calculating principal component scores when data are missing. Three broad categories of methods to deal with missing data are discussed here: imputation, robust estimation of the covariance matrix and the EM algorithm.

Imputation involves replacing missing values with a sensible value. This can include replacing missing values with observed values elsewhere in the dataset (hot deck imputation) or substituting the mean of the observed values for missing values (mean imputation). Regression imputation involves replacing missing values by values estimated from regressing observed values. Rubin and Little (2002) note that statistical methods must be modified to account for missing data if valid inferences are to be made. Plaia and Bondì (2006) discuss single imputation for environmental data while Taylor et al. (2013) compare principal components estimated from incomplete data where the missing observations have been replaced using imputation methods to principal components estimated from a covariance matrix based on incomplete data. They find that principal components from imputed data are preferable in terms of reconstruction error. Josse and Husson (2012) introduce an iterative multiple imputation method based on PCA that aims to minimise reconstruction error. Iterations move between estimating principal components and reconstructing the data (including missing values) and the aim is to minimise reconstruction error. The number of components to be estimated must be specified *a priori* since this number is used during the iterative algorithm. A function based on cross validation is provided for this and graphical tools to assess the variability of loadings as a result of imputation are provided in the `missMDA` (Husson and Josse, 2015) package in `R`. The algorithm is regularized to prevent overfitting. Josse and Husson (2012) claims their approach is superior to the NIPALS (nonlinear iterative partial least squares) algorithm (Wold and Lyttkens, 1969), an iterative method for PCA based

on least squares regression.

The covariance matrix could be constructed using pairwise methods where the covariance matrix is calculated using available values but there is no guarantee the covariance matrix will be positive definite. Higham (2002) presents methods to approximate the nearest positive definite covariance matrix and PCA could then be carried out on the approximated covariance matrix. Schneider (2001) deals with missing data specifically in the context of climate data where the number of variables (monitoring sites) greatly exceeds the number of observations (time points), and deals with estimating mean values and covariance matrices in the presence of missing data. The EM algorithm is employed and regularized in a manner similar to ridge regression to deal with the fact that the regression model parameters are being estimated from rank-deficient data. It is shown that this method is more accurate than a non-iterative imputation technique based on truncated principal component analysis in Smith et al. (1996) where a single smoothing parameter is chosen for the whole data set rather than adaptive selection from GCV used in Schneider (2001). It is also shown in Schneider (2001) that a spatial or temporal covariance structure can be used to improve estimates of missing values using the regularised EM algorithm. Lounici et al. (2014) provide an algorithm for estimation of a covariance matrix in the presence of missing data that does not require imputation of the raw data values.

Rather than imputing data or constructing a covariance matrix from incomplete data, Chen (2002) discusses using the EM algorithm to perform PCA directly on the incomplete data. Chen (2002) notes that for PCA it is generally assumed that the variables have been mean centered but if a significant portion of the data are not available then the mean might not be well estimated. Instead, Chen (2002) recommends Wiberg's method (Wiberg, 1976) so as to be robust against a poorly estimated mean. Rubin and Little (2002) calls this approach a 'model-based' procedure. Skočaj et al. (2007) modify the EM algorithm for spatio-temporal image data and assign weights of 0 to missing values combined with a temporal smoothing parameter to improve data reconstruction in the presence of missing values.

In this thesis, missing data are encountered in Chapter 5, where the statistical technique discussed (PCA) requires a data set with no missing values. The method proposed in Josse and Husson (2012) is used to estimate missing values since PCA is implemented

using singular value decomposition performed on the data matrix, rather than eigen decomposition performed on the covariance matrix, meaning that the missing data methods relating to constructing a covariance matrix in the presence of missing values are not suitable here. Although PCA can be performed in the presence of missing values using approaches based on the EM algorithm, it seemed more appropriate to create a data matrix with no missing values that could be used as data to investigate the effect of adjusting PCA for spatial and/or temporal structure in the data. Adjusting PCA for structure in the presence of missing data might lead to confounding between identifying the effect of adjusting for known structure and the effect of missing values. The method proposed in Josse and Husson (2012) also considers uncertainty due to missing values and is regularised to prevent overfitting so was used in this thesis as it seemed the most appropriate approach, given the aims of this thesis.

1.5.3 Limits of detection

This section describes statistical methods that can be applied to data where some observations are recorded as below (or above) the limit of detection, often described as 'censored data'. First, different types of censoring are described followed by a description of approaches commonly used to replace censored observations with a numeric value. The strengths and weaknesses of these approaches are discussed, followed by a justification of the method selected used to handle missing data throughout this thesis.

Data can be recorded as below (or above) the limit of detection, where it is known only that the value of the observation is between zero and the detection limit of the laboratory equipment (or above the maximum detection level of the equipment). This is known as left (right) censored data when the data are recorded or reported as being below (above) a specified threshold. Data might also be 'interval censored' where it is only known that the observation lies between two values. Environmental data sets will often contain data that are left censored and throughout this thesis it is assumed that any reference to censored data means that the data in question are recorded as below the limit of detection. Eastoe et al. (2006) indicate it is better to incorporate censored observations into statistical analyses in some way rather than not at all and this is in agreement with Helsel (2012) who says that deletion of the censored observations is the worst way of dealing with them as it creates bias in the data. It is necessary therefore to incorporate

the censored value into any summary statistics or analyses. This is complicated further when a data set contains several different detection limits, possibly as a result of changes or improvements to monitoring equipment over time and is known as 'multiply censored' data. Helsel (2010) gives a concise summary of the development of statistical methods for left censored data and a more exhaustive description can be found in Helsel (2012), on which the following section is based.

The simplest way of handling censored data is substitution where censored values are replaced with an (arbitrary) alternative value such as the detection limit or a multiple of the detection limit. Substitution creates bias in the data, the extent of which will depend on the proportion of observations that are recorded as below the detection limit. Helsel (2012) refers to substitution as 'fabrication' and strongly recommends this approach should not be used. Instead, imputation methods are recommended where the idea is to estimate summary statistics for the distribution of the data, taking into account the censored observations. Values are then simulated based on this distribution and subject to the constraint that the simulated value falls below the detection limit. The simulated values are then imputed into the data set in place of the censored values. A brief description of three common approaches to imputation is given below.

Maximum Likelihood Estimator

The maximum likelihood estimator (MLE) is a parametric approach suitable for data with multiple detection limits and involves fitting an assumed distribution to the observed data values and the proportion of censored values. The parameter estimates obtained from the fitting procedure describe the distribution with the maximum likelihood of producing the data (censored and uncensored). This method performs well if the data set is large enough (n>30) but problems can occur if the distribution is misspecified, resulting in poor estimates of the mean and variance. Shumway et al. (2002) show that if the data are thought to arise from a log-normal distribution, then back transformation of the estimated parameters can produce biased estimates due to the non-linear relationship between the log (transformed) and linear (original) scales. Bias corrections can be carried out but Shumway et al. (2002) state that under severe censoring conditions many bias corrections correct in the wrong direction. They propose a bias correction based on the Quenouille-Tukey jacknife to avoid this problem. Helsel (2012)

states that if the data are log-normally distributed and there is enough uncensored data to adequately estimate the parameters of the distribution then MLE provides the best estimate. It is also noted that 'large' increases with increased skewness.

Kaplan-Meier

The Kaplan-Meier (KM) approach is a non-parametric technique developed for use in survival analysis where the observations are often right censored. Helsel (1990) suggests this technique could be used for left censored data by 'flipping' the data to construct a right censored data set to which standard survival analysis techniques can be applied to calculate summary statistics of the data set. A fixed constant is chosen (a value greater than the maximum observed value say) and each (censored and uncensored) value is subtracted from this constant.

The KM approach in the context of survival analysis estimates the probability that observations will survive beyond a certain timepoint. For left censored data, this means that the probability that observations will fall below a limit of detection is estimated. This method is suitable for data with multiple detection limits when less than 50% of observations are censored (Helsel, 2012). An advantage of this technique is that it does not require specification of a probability distribution thus removing the risk of misspecification suffered by MLE. Helsel (2010) notes however that KM should only be used when there are multiple detection limits.

Regression on Order Statistics

Regression on Order Statistics (ROS), first proposed by Helsel and Cohn (1988), is a semi-parametric method that models censored distributions using a linear regression model of observed values against their normal quantiles. It is assumed that the response (uncensored observations) is a linear function of the explanatory variable (normal quantiles) and that the errors have constant variance. The ROS method is described below following the notation in Shumway et al. (2002).

Suppose there are n_o transformed (log transformed or otherwise) observations $y_i, i = 1, ..., n_0$ below a common transformed detection limit U and n_1 observations $y_i : i =$

$n_0+1, ..., n_0+n_1$ that are observed and greater than U. Assuming that $y_i \overset{iid}{\sim} N(\mu_y, \sigma_y^2)$, the mean and variance will satisfy the equation

$$y_i = \mu_y + \sigma_y \Phi^{-1}(P_i),$$

where $P_i = \text{prob}(Y_i) \leq y_i$ and $\Phi^{-1}(\cdot)$ denotes the inverse of the cumulative normal distribution function. Under this setup, a regression on the normal scores should yield an intercept and slope parameter that are the mean and variance of the transformed observations. It is common (Shumway et al., 2002) to replace the probabilities by the adjusted ranks so that the regression becomes

$$y_i = \mu_y + \sigma_y \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right) + \varepsilon_i, \tag{1.26}$$

where $i = n_0 + 1, n_0 + 2, ..., n_0 + n_1$ and ε_i are uncorrelated and have equal variance. Predicted values for the censored values can be obtained using Equation (1.26) and then back-transformed (if necessary) along with the uncensored observations. The mean and variance can then be calculated on the original scale.

Shumway et al. (2002) found that ROS produces estimates of the same quality as MLE for moderate ($n = 50$) samples, and of better quality than MLE for small ($n = 20$) samples. Helsel (2012) states that ROS is an alternative to Kaplan-Meier (the most generally applicable method) when more than 50% of the data are censored and an estimate of the median is desired. Helsel (2012) also states that ROS outperforms MLE when $n < 30$ and is more effective when 50% to 80% of observations are censored.

The ROS method is used in this thesis to impute suitable values for observations recorded as below the limit of detection. This method was chosen since the time series analysed in this thesis have up to 50% of observations recorded as 'below the detection limit' and there are multiple detection limits. ROS is a suitable imputation method for both of these situations. ROS was applied using the `NADA` (Lee, 2012) package in `R`.

Section 1.5 has addressed many of the issues commonly encountered when working with spatiotemporal/environmental data mentioned in Henderson (2006). Temporal correlation is accounted for in Chapter 2 using dynamic factor analysis and principal components analysis is adjusted for temporal correlation in Chapter 5. Spatial correlation

is revisited in Chapters 3-5 where spatial correlation in the context of river networks is specifically discussed. Missing values in the data are replaced with values estimated using the methods proposed in Josse and Husson (2012) when it is necessary to work with a complete data set in Chapter 5. Limit of detection issues were addressed before any analysis was carried out in Chapters 2-5 and the ROS method was used, for the reasons discussed in Section 1.5.3.

1.6 Aims and objectives

This chapter has introduced water quality monitoring and described several statistical methods that can be used to estimate temporal and spatial patterns of water quality parameters. The motivation for considering such approaches is that regulatory agencies are coming under increasing pressure to reduce costs and one way to do this would be to enhance statistical analyses by incorporating spatial information into statistical techniques. This would increase the information on which statistical models are based without a corresponding increase from collecting additional data. The remainder of this thesis concentrates on identifying and comparing temporal trends and adapting existing statistical methods to account for known spatial and temporal structure. The overall aim of this thesis is to develop statistical methods to identify common temporal patterns in water quality data which take into account the spatial structure of river networks. Specifically this thesis aims to

1. Develop statistical methods to compare curves and identify the nature of any differences.

2. Determine an appropriate spatial covariance structure for data recorded on river networks by considering spatial relationships at multiple scales.

3. Develop a study to compare statistical models based on spatial covariance functions with models based on spatially observed covariates.

4. Design and implement a study to assess the effect of reducing the size of a water quality monitoring network.

5. Develop a novel adaptation of a dimension reduction technique to identify dominant temporal patterns, taking into account known spatial and temporal structure in the data.

Chapter 2

Comparing temporal trends and seasonal patterns

This chapter presents exploratory analysis of the data provided by the Environment Agency, and proposes methods that can be used to compare temporal trends and seasonal patterns estimated for different spatial locations and estimated using a variety of statistical methods.

Chapter 1 introduced several statistical methods commonly used to estimate temporal trends and seasonal patterns in water quality data. These temporal patterns could be estimated at individual monitoring sites or for large hydrological areas (LHA's, defined in Section 1.1) where data for several monitoring sites are combined. Once estimated, these curves can be compared to find similarities or differences between sites or LHA's, and to try to understand the nature of any differences. Differences might be a result of geographic location, indicated by changes in curves moving from North to South or East to West, for example. Differences in shape of curves might also be a result of biological drivers identified by regulatory agencies and so identification of differences might improve understanding of the processes driving seasonal patterns in the data.

There are many ways in which curves might be compared and the nature of differences and similarities assessed, some of these are described below. Next, a dimension reduction technique - dynamic factor analysis (DFA) - is used to identify dominant patterns in the temporal and seasonal evolution of nitrates at LHA level. DFA is used to provide an initial exploration of a large complex dataset with the aim of identifying temporal

patterns common to several LHA's. Following this, the focus of the chapter moves to a single LHA where the temporal and seasonal pattern is estimated using other statistical methods. Many statisticians will use their own preferred statistical method for estimating temporal trends and seasonal patterns without any guidance as to which method is most appropriate. A novel approach is developed to compare the shapes of curves estimated from commonly used statistical methods, described in Section 1.3 and Section 1.4, to understand how these methods might lead to different conclusions about the data.

Finally, the Comments section includes some discussion on improvements that could be made to the proposed comparison methods. A discussion on how each of the statistical methods compared in Section 2.2.2 incorporates spatial information is also given, along with guidelines to help analysts choose which statistical method to use when estimating temporal trend and seasonal pattern for data collected from several monitoring sites.

2.1 Comparing curves

Curves can be compared in terms of differences or similarities between curves (or groups of curves). This section describes several methods discussed in the literature by which these might be assessed.

Differences

Many methods in the literature compare two or more non-parametric regression curves where it is assumed that each curve represents a different group. Hall and Hart (1990) test the equality of two non-parametric mean curves using a bootstrap test with an application to acid rain data. Härdle and Marron (1990) compare non-parametric curves in a parametric framework. Delgado (1993) propose a method to test the equality of several nonparametric smooth curves that does not depend on the choice of smoothing parameter. Young and Bowman (1995) discuss an adaptation of analysis of variance for non-parametric curves where curves are estimated using the Gasser-Müller approach to reduced bias in the estimation of the smooth function and apply this to Spanish onion data. Dette and Neumeyer (2001) also describes methods for non-parametric analysis

of covariance and note that their method is applicable in cases of different design points and heteroscedastic errors. Neumeyer et al. (2003) propose a test to compare two curves that allows for different design points and heteroscedastic errors and resolve problems in the proofs of Kulasekera (1995) and Kulasekera and Wang (1997) who compare regression curves with different design points but under the assumption of homoscedastic errors. Pardo-Fernández et al. (2007) develop a method for testing the equality of two or more non-parametric regression functions using Kolmogorov-Smirnov and Cramér-von Mises type statistics and apply this to Dutch household expenditure data. Park and Kang (2008) modify the SiZer method (introduced in Chapter 1) to compare two or more smooth regression curves estimated using local linear regression fitted with Gaussian kernel weighted least squares for a variety of bandwidths. Confidence intervals (adjusted for multiple comparisons) are calculated for the difference between two curves at a specific time point and bandwidth and these are used to assess the significance of differences between curves.

Statistical methods also exist to compare groups of non-parametric curves to investigate differences in group means. Fan and Lin (1998) compare two or more groups of non-parametric curves using an adaptive Neyman test and wavelet thresholding. Walker and Wright (2002) compare groups of non-parametric curves fitted using generalized additive models with an adapted anlysis of variance approach. This method is suitable in cases where each curve is based on the same design points and has the same level of smoothing and can also be used to compare several smooth curves as well as groups of curves. Zhang et al. (2010) consider a hypothesis test that two groups of curves observed without noise have the same mean function. There are many examples in the functional data analysis literature where groups of curves are compared using a functional version of analysis of variance (fANOVA). For example, Sain et al. (2011) compare two dynamic downscaling methods with an application to summer temperature and precipitation over North America, Saeys et al. (2008) show how fANOVA can be used to analyse spectroscopy data, Cortina-Borja et al. (2011) apply fANOVA to physical activity data to identify times of the day and year when children are most active and Andrade et al. (2014) use fANOVA to investigate the effect of walking sticks on the gait of stroke patients. Some adaptations of the fANOVA methodolgoy include Kaufman et al. (2010) who implement fANOVA within a Bayesian framework and Girimurugan and Chicken (2013) who develop wavelet ANOVA for functional data.

If any of the tests described above show that curves, or groups of curves, are significantly different then the next stage of the analysis will be to investigate the nature of the differences. For example, in functional data analysis, Vsevolozhskaya et al. (2014) compare treatment levels and develop follow up pairwise comparison tests to investigate significant differences between group means over small ranges of time to find out at what time points the treatments differ. Cox and Lee (2008) also develop a multiple comparison method but for pointwise comparisons and show that their adjustment for multiple testing is an improvement on the Bonferroni correction which does not perform well due to the presence of temporal correlation. Functional PCA can help identify sections of curves that are the main sources of variability (Ramsay and Silverman, 2006).

Rather than comparing the mean or variance of curves, it might be of interest to compare curves based on shape. For example, Minas et al. (2011) uses functional data analysis techniques to produce gene expression curves and note that most tests of differences between groups are based on area between curves or the sum of squared differences between curves at a number of discrete points on the x-axis. Distance measures reflecting differences in shape are described and a test statistic is developed to test the hypothesis that the shape of the mean curve for two or more groups is the same.

Differences in shape of curves might be investigated by considering derivatives of the curves rather than the original curves. Djurdjevic et al. (2012) use the first derivative to investigate the dendrite coherency point of alloys, Amna et al. (2012) look at the first derivative of bacterial growth curves, McIntyre et al. (2011) use the first derivative of spectra curves to detect counterfeit whisky samples. The second derivative of near infra red spectral data is used in Sinelli et al. (2010) to monitor the freshness of minced beef and Shen et al. (2012) use the second derivative of spectra curves to discriminate between Chinese rice wines. Müller (2012) discusses the use of derivatives of growth curves to investigate child growth. Gorrostieta et al. (2014) use the first and second derivatives to characterise ocean wave profiles. Chaudhuri and Marron (2002) discuss the benefits of investigating the second derivative, or curvature, of a curve as well as the first derivative, or slope. Silverman and Ramsay (2005) notes that when finding landmarks on a smooth curve it is often easier to identify landmark locations at some derivative level. Within the SiZer framework described in Chapter 1 Chaudhuri and Marron (1999) and Hannig and Marron (2006) show how to calculate confidence intervals for derivative curves with adjustments made for multiple comparisons since confidence intervals are calculated for

several time points and a range of bandwidths. Orr et al. (2015) simulate curves using the output from an additive model fitted using the `mgcv` package (Wood, 2006) in `R` and this method can be adapted to simulate derivative curves from which the 2.5th and 97.5th quantiles can be calculated as the lower and upper limits of a confidence interval (Simpson, 2014).

In the functional data literature differences in shape of curves might be investigated by considering registered curves where curves are aligned based on landmarks of interest (Ramsay and Silverman, 1997). Rather than simply analysing the aligned curves it might be of interest to analyse properties of the alignment procedure such as the horizontal variance of the curves as in Lu and Marron (2013).

This section has provided many examples found in the literature to compare curves, compare groups of curves and identify the nature/location of any significant differences. In Section 2.2.2 the approach developed in Minas et al. (2011) is used to compare curves estimated using several statistical methods, to investigate differences between these estimated temporal trends and seasonal patterns. This approach was chosen since comparisons are based on shape rather than estimated values and this is useful since not all curves being compared are estimated on the same scale.

Similarities

The previous section discussed some statistical methods for finding significant differences between curves but another option is to look for similarity between curves. Magnuson et al. (1990) state that temporal coherence is the "degree to which different locations within a region behave similarly through time" and there are many different examples in the literature of how "similarly" might be defined, some of which are described here.

Correlation has been used as a measure of coherency between time series in many studies. For example, Magnuson et al. (1990) estimate the correlation between data recorded at seven lakes in Wisconsin, USA, over a seven year period for 37 limnological variables. Correlation is calculated between pairs of lakes for each of the variables and mean correlation and percentage of strong correlations were calculated for each pair of lakes across all variables and for each pair of variables across all lakes. George et al. (2000)

use a similar approach to investigate coherence between six lakes in the Lake District, UK, for seven variables recorded over a 30-40 year period. Average coherence was calculated for each variable between lake pairs for each season to investigate differences in coherence due to seasonality. Other examples of using correlation between time series as a measure of temporal coherence include Pace and Cole (2002), Fölster et al. (2005), Magnuson et al. (2006), Patoine and Leavitt (2006), Lansac-Tôha et al. (2008), Caliman et al. (2010) and Sánchez-López et al. (2015). Ghanbari and Bravo (2011) use squared coherence, the frequency domain analogue of correlation to investigate the strength of linear relationships between lake ice, local climate and teleconnections (see also Hanson et al. (2004) and Kao et al. (2008) for further examples.).

Time series can be represented in the frequency domain as wavelets that capture non-regular periodicities in the data. Regular periodicity is usually represented using Fourier analysis. Grinsted et al. (2004) develop a measure of coherency that can be used to compare the wavelet representation of two time series and apply this to a comparison of the Arctic Oscillation and Baltic Sea ice extent. Maraun and Kurths (2004) introduce a statistical test based on Monte Carlo simulations to assess the significance of wavelet coherency. Hassan et al. (2010) consider wavelet coherence for uterine electrical activity and Sanderson et al. (2010) develop a novel measure of wavelet coherency with applications to neuroscience data. Other examples of coherency measured in the wavelet frequency domain include Lachaux et al. (2002) (brain signals) and Polansky et al. (2010) (animal location time series). Park et al. (2014) and references therein provide detailed results for wavelet coherence and wavelet partial coherence.

Hari et al. (2006) define coherence as the proportion of variance shared (or R^2 in standard linear modelling terminology) between a linearly detrended time series of river water temperature and the linearly detrended mean of the other 24 time series. This approach is also used by Baines et al. (2000) who note that R^2 might have an advantage over the Pearson correlation coefficient since it does not differentiate between positive and negative relationships.

Blenckner et al. (2007) use a meta-analysis approach, usually found in biostatistical applications, to investigate the common patterns in physical, chemical and biological traits in 18 lakes in Europe and coherency between these traits and the North Atlantic Oscillation. This is not a meta-analysis in the usual sense however since Blenckner

et al. (2007) apply all analyses to a single data set rather than combining multiple data studies to provide an overall estimate of trend. By using a meta-analysis approach however Blenckner et al. (2007) are able to focus on common trends shared by the 18 lakes rather than looking at each lake individually.

Coherency of time series might also be investigated using principal components analysis (PCA), introduced in Chapter 1, where PCA is used to find locations that behave in a similar way over time, called "regionalization" (Ehrendorfer, 1987). PCA for regionalisation has been applied in areas including rainfall (Ehrendorfer (1987), Neal et al. (2004)), hydrology (Monk et al. (2007), Carey et al. (2010)), surface wind (Jiménez et al., 2008) and streamflow (Kahya et al., 2008).

Dynamic factor analysis (DFA) described in Chapter 1 can be used to model common temporal patterns among many time series. Time series with loadings of similar magnitude for a particular common trend can be said to behave in a similar way over time. An appealing feature of DFA is that the estimated common trends are constructed to have temporal dependence.

In functional data analysis, functional clustering can be used to find groups of curves with similar behaviour over time as in James and Sugar (2003), Henderson (2006), Ignaccolo et al. (2008) and Haggarty (2012).

This section has described many examples found in the literature to identify similarities between curves. In this thesis dynamic factor analysis (Chapter 2) and principal components analysis (Chapter 5) are used to identify LHA's or monitoring sites exhibiting similar temporal trends and seasonal patterns. In particular, a novel adaptation of PCA is developed to account for river network structure. The results for DFA and PCA are easily plotted as time series or maps and thus are suitable when working with spatiotemporal data.

2.2 Application to EA data

The previous section described several methods that could be used to investigate differences and similarities between (groups of) curves. This thesis aims to adapt statistical methods that can be used to identify common temporal patterns to take into account

spatial relationships in the data. As a first step towards this, dynamic factor analysis (DFA, described in Chapter 1) is applied to the 59 temporal trends and seasonal patterns estimated at LHA level. This provides some exploratory analysis of the data and improves understanding of a statistical technique suitable for identifying common temporal patterns. DFA is applied in this chapter assuming spatial independence, a reasonable assumption when the spatial units of interest are LHA's, constructed to be independent of surrounding areas. This assumption is not suitable however when focussing within a single LHA since spatial dependence is likely to be present in river networks. Spatial dependence within a river network is investigated in detail in Chapters 3 and 4. The temporal trend and seasonal pattern are estimated for a single LHA using four different statistical methods. Analysts will often favour a particular statistical technique rather than explore several possible analysis methods. Section 2.2.2 develops a novel approach to compare the temporal trends and seasonal patterns estimated from four statistical methods, with the aim of identifying differences in shape and to assess if different statistical methods lead to different conclusions. The curves will be compared using the DBF statistic (Minas et al., 2011), to test if the shapes of curves are significantly different, and curvature will be used to determine the nature of any significant differences. The DBF statistic and curvature calculations are described in detail below followed by application of these methods to the data from an LHA in Wales.

DBF statistic

Minas et al. (2011) propose the distance based test statistic

$$\mathrm{DBF}_{\Delta_{d_V}} = \mathrm{tr}(\mathbf{B}_{\Delta_{d_V}})/\mathrm{tr}(\mathbf{W}_{\Delta_{d_V}}),$$

where $\mathrm{tr}(\mathbf{B}_{\Delta_{d_V}})$ and $\mathrm{tr}(\mathbf{W}_{\Delta_{d_V}})$ are the trace of distance matrices representing between and within group variability, respectively. $\mathrm{DBF}_{\Delta_{d_V}}$ is used to test the hypotheses

$$H_0 : d_V(\mu^{(i)}, \mu^{(j)}) = 0 \text{ vs. } H_1 : d_V(\mu^{(i)}, \mu^{(j)}) \neq 0.$$

Marron and Tsybakov (1995) show that d_V between the mean functions of two groups of curves $\mu^{(i)}$ and $\mu^{(j)}$ can be calculated as

$$d_V(\mu^{(i)}, \mu^{(j)}) \equiv \left(\int_0^1 \delta(i,j)^2 \mathrm{d}t + \int_0^1 \delta(j,i)^2 \mathrm{d}t \right)^{\frac{1}{2}}, \tag{2.1}$$

where for a given time-point t, $\delta(i,j)$ is the minimum Euclidean distance between point $\mu^{(i)}(t)$ and all points on $\mu^{(j)}(i \neq j)$ and $\delta(j,i)$ is the minimum Euclidean distance between point $\mu^{(j)}(t)$ and all points on $\mu^{(i)}$. This measure of distance aims to capture how differences between curves are judged 'by eye'.

The test is accomplished by calculating visual distance d_V for all pairs of curves and creating a distance matrix whose rows and columns are labelled by group number. The DBF statistic is calculated and the rows and columns are permuted P times with the DBF statistic calculated for each permutation. The p-value of the test is the proportion of times the DBF statistic from the permuted distance matrix is greater than the original DBF statistic. The test can also be used to compare more than two functional mean curves. To compute the DBF statistic in Minas et al. (2011) the curves must first be normalised thus ensuring similarity relates to the shape of the curves being compared rather than scale (Minas et al., 2011). Normalisation can be carried out as in (2.2) where y_{Lt} are log(TON) for LHA $L : 1, ..., 59$ at time point $t : 1, ..., T$.

$$y_{Lt_{\text{norm}}} = \frac{y_{Lt} - \min(y_{Lt})}{\max(y_{Lt}) - \min(y_{Lt})} \tag{2.2}$$

The DBF test is a global test of differences and does not provide information on the nature of any differences. A visual aid would be useful to help identify the nature of differences between curves. The next section describes how curvature of a graph is calculated and later in the chapter plots of curvature are used to highlight some of the differences in shape between curves once the global test of differences indicates the presence of significant differences.

Curvature

Curvature (or $\kappa(t)$), calculated using (2.3) where $f'(t)$ and $f''(t)$ are the first and second derivatives respectively of a smooth curve at time t, is a measure of how 'bendy' a curve is.

$$\kappa(t) = \frac{f''(t)}{[1 + (f'(t))^2]^{\frac{3}{2}}} \tag{2.3}$$

Large positive or negative values of $\kappa(t)$ indicate turning points and a value of zero means there is no curvature i.e. a straight line. If it can be assumed that the slope of the curve (the first derivative) is small compared to unity then the second derivative can be used as an approximation to (2.3). First and second derivative curves are calculated in this thesis using the `SiZer` (Sonderegger, 2012) package in `R`.

Since the smooth curve is the smoothed version of an estimate of temporal trend or seasonal pattern rather than true value it would be useful to have a confidence interval for $\kappa(t)$ and this requires knowledge of the distribution of $\kappa(t)$. Meyer (1970) shows how a Taylor expansion of $\kappa(t)$ can be used to approximate the expectation and variance of $\kappa(t)$ which can be used to calculate a confidence band for the curvature curve. $f''(t)$ and $f'(t)$ can be thought of as random variables with a mean$_t$ and variance$_t$ (obtained from the `SiZer` output). To calculate a 95% confidence interval for $\kappa(t)$, start by writing (2.3) as

$$Z = \frac{X}{Y} \text{ i.e. } Z = H(X, Y) \tag{2.4}$$

where $Z = \kappa(t)$, $X = f''(t)$ and $Y = [1 + f'(t)^2]^{\frac{3}{2}}$. The problem now is to estimate $E(Z)$ and $V(Z)$, the mean and variance of a ratio of random variables. According to Meyer (1970),

$$\begin{aligned} E(Z) &\simeq H(\mu_X, \mu_Y) + \tfrac{1}{2}\left[\tfrac{\delta^2 H}{\delta X^2}\sigma_X^2 + \tfrac{\delta^2 H}{\delta Y^2}\sigma_Y^2\right], \\ V(Z) &\simeq \left[\tfrac{\delta H}{\delta X}\right]^2 \sigma_X^2 + \left[\tfrac{\delta H}{\delta Y}\right]^2 \sigma_Y^2, \end{aligned} \tag{2.5}$$

where $H = H(\mu_X, \mu_Y)$ is H (see Equation (2.4)) evaluated at μ_X, μ_Y. To calculate this μ_X and σ_X^2 are obtained from the `SiZer` output and μ_Y and σ_Y^2 can be calculated as a function of a single random variable as in Meyer (1970) as follows:

$$\text{Let } S = H(T) = (1 + T^2)^{3/2}, \tag{2.6}$$

where $T = f'(t)$ is a function of a single random variable. Meyer (1970) show that the mean and variance of S (E(S) and V(S) respectively) can be calculated as

$$\begin{aligned} E(S) \simeq & \quad H(\mu_T) + \tfrac{H''(\mu_T)}{2}\sigma_T^2 \\ = & \quad \left[1+\mu_T^2\right]^{3/2} + \tfrac{3(1+\mu_T^2)^{1/2}+3\mu_T^2(1+\mu_T^2)^{-1/2}}{2}\sigma_T^2 \\ = & \quad \mu_Y, \end{aligned}$$

$$\begin{aligned} V(S) \simeq & \quad \left[H'(\mu_T)\right]^2 \sigma_T^2 \\ = & \quad \left[3T\left(1+\mu_T^2\right)^{1/2}\right]^2 \sigma_T^2 \\ = & \quad \sigma_Y^2. \end{aligned}$$

Remembering that $Z = \frac{X}{Y}$, this gives

$$\begin{aligned} \tfrac{\delta H}{\delta X} &= \tfrac{1}{Y}, \\ \tfrac{\delta H}{\delta Y} &= \tfrac{-X}{Y^2}, \\ \tfrac{\delta^2 H}{\delta X^2} &= 0, \\ \tfrac{\delta^2 H}{\delta Y^2} &= \tfrac{2X}{Y^3}. \end{aligned} \tag{2.7}$$

The expressions in (2.7) can be substituted into (2.5) giving

$$\begin{aligned} E(Z) \simeq & \quad \tfrac{\mu_X}{\mu_Y} + \tfrac{1}{2}\left[0 + \tfrac{\mu_X}{\mu_Y^3}\sigma_Y^2\right], \\ V(Z) \simeq & \quad \left[\tfrac{1}{\mu_Y}\right]^2 \sigma_X^2 + \left[\tfrac{-\mu_X}{\mu_Y^2}\right]^2 \sigma_Y^2. \end{aligned}$$

An approximate 95% confidence interval for Z, assuming X and Y in (2.4) are independent and $\kappa(t)$ are normally distributed, can now be calculated as

$$E(Z) \pm q \times \sqrt{V(Z)}. \tag{2.8}$$

In (2.8) $q = 1.96$ is the 97.5th quantile of a standard normal distribution, N(0,1), thus giving a 95% confidence interval at several discrete values on the x-axis. Since multiple confidence intervals are calculated it would be better to make an adjustment to the quantile value so that intervals are not so small as to induce spurious significant results. Chaudhuri and Marron (1999) and Hannig and Marron (2006) show how to calculate a suitable quantile value that takes into account multiple comparisons for both the first and second derivatives. Since the second derivative is a good approximation to curvature, the 97.5th quantile from the second derivative calculations could be used rather than from the standard normal distribution to give more conservative estimates of the confidence intervals.

The confidence interval calculations could be further improved by considering the distribution theory described in Chaudhuri and Marron (1999) and Hannig and Marron (2006). Adjustments to the calculation could also be made by including a covariance term in the Taylor expansion of the curvature formula to account for dependence between the first and second derivatives.

2.2.1 Common temporal trends and seasonal patterns

This section will show how the DBF statistic and curvature can be used to compare temporal trends and seasonal patterns estimated for each of the 59 LHA's in England and Wales. The aim is to provide some exploratory data analysis of the large data set provided by the EA and to investigate whether common patterns can be found among LHA's. This is a first step towards finding common patterns within LHA's: LHA's are defined to be spatially independent and this assumption means that the complex spatial correlation based on river network structure found within an LHA does not need to be incorporated in this analysis. Later in this thesis (Chapters 3-5), spatial dependence within an LHA will be considered in detail and methods will be developed to incorporate the spatial structure found within an LHA into common patterns methodology.

Bowman et al. (2010) and Miller et al. (2014) modelled smooth temporal trends and seasonal patterns of OP and TON for each of the 59 LHA's in England and Wales

using additive models as in Section 1.3.1. Functional mean curves were estimated for temporal trend and seasonal pattern with corresponding 95% confidence intervals and LHA's whose smooth functions did not fall within the intervals were classed as behaving differently from average. In this section, the smooth functions will be compared using dynamic factor analysis to investigate common patterns shared among the 59 LHA's for both temporal trend and seasonality.

Temporal trends and seasonal patterns of log(TON) were estimated for each LHA using additive models fitted using the `mgcv` package in `R`. The data from all monitoring sites within an LHA were modelled using (2.9), where m indexes spatial site ($m = 1, ..., m_L$ and m_L is the number of monitoring sites in $\text{LHA}_L, L = 1, ..., 59$), t indexes date in decimal date format ($t \in [1995.000, ..., 2010.997]$), r indexes day of year ($r = 1, ..., 365$), $\varepsilon_{mtr} \overset{iid}{\sim} N(0, \sigma^2)$, $s_i()$ are smooth terms and k_i is the dimension of the basis used to represent the $i = 1, 2, 3$ smooth terms. The spatial term, $s_1()$, is estimated using a thin plate regression spline since this type of spline is optimal for fitting a smooth spatial surface (Wood, 2006). The trend term $s_2()$ is estimated with a cubic regression spline and the seasonal term $s_3()$ is a cyclic cubic regression spline since this type of spline forces the endpoints of the estimated curve to be equal thus ensuring no gap between the end of one year and the beginning of the next. Two methods - GCV and REML - were used to estimate k and both gave similar results. In Equation (2.9) $k_1 = 12, k_2 = 9, k_3 = 8$. These values for k in (2.9) also result in similar degrees of freedom to the models fitted in Bowman et al. (2010). Temporal trends and seasonal patterns were modelled for the time period 1995 to 2010 since few data points were available for 1990-1994.

$$\begin{aligned} \log(\text{TON})_{mtr} &= s_1(\text{Easting}_m, \text{Northing}_m, k_1) \\ &+ s_2(t, k_2) \\ &+ s_3(r, k_3) \\ &+ \varepsilon_{mtr} \end{aligned} \tag{2.9}$$

After extracting spatial information from (2.9), values of log(TON) were predicted for 250 equally spaced time points for $s_2()$ and for 300 equally spaced time points for $s_3()$. Each point had noise added as a random draw from the distribution of residuals from each LHA model and were drawn from a normal distribution with mean of zero and standard deviation of the residuals from the fitted model. These noisy temporal trends

and seasonal patterns were modelled for all 59 LHA's using DFA. Three types of DFA model were fitted: two assuming LHA's are independent and one incorporating spatial dependence between LHA's. The models assuming independence were fitted using a diagonal covariance matrix of the form in (2.10). The diagonal elements of this covariance matrix are either allowed to be different or constrained to be equal and each element of the diagonal reflects the variance of log(TON) in an individual LHA. Spatial dependence is assumed if the model is fitted using a covariance matrix of the form in (2.11). The off-diagonal elements of this matrix represent spatial relationships in the data not captured by the common trends.

$$\begin{bmatrix} \sigma_{1,1} & 0 & \dots & 0 \\ 0 & \sigma_{2,2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{n,n} \end{bmatrix} \tag{2.10}$$

$$\begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \dots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_{2,2} & \dots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \dots & \sigma_{n,n} \end{bmatrix} \tag{2.11}$$

DFA was implemented as follows:

1. Fit a model with 1 common trend and a diagonal covariance matrix and record the AIC value.

2. Fit a model with 1 common trend and a non-diagonal covariance matrix and record the AIC value.

3. Repeat (1) and (2) for models with 2 common trends and record the AIC values.

4. If the AIC values from (3) are greater than in (1) and (2) then stop and conclude that 1 common trend is sufficient to describe the temporal trend. The 'best' model from (1) and (2) is the one with the smallest AIC value i.e. the most appropriate covariance structure (diagonal or non-diagonal) is also determined using AIC.

5. If the AIC values from (3) are less than in (1) and (2) then continue fitting DFA models by increasing the number of common trends to be estimated until AIC values start to increase. The best model is the one with the lowest AIC value.

2.2.1.1 Temporal trend

DFA models were fitted to the 59 noisy temporal trends fitted using (2.9). Up to 5 common trends were estimated and diagonal and equal, diagonal and unequal, and unconstrained covariance structures were used. For the unconstrained covariance models only 1 and 2 common trends were estimated since these models are computationally demanding. Table 2.1 contains the log likelihood and the difference between AIC value for the best fitting model and AIC for all other models fitted. The best model (highlighted in bold) as determined by AIC value is model 1 which has one common trend and a diagonal and equal covariance structure indicating equal variance for all LHA's. The left panel of Figure 2.1 shows the loadings for model 1. LHA's with a loading less than 0.001 are not shown. The plot on the right shows the pattern of the common trend and indicates peaks in log(TON) at 1997 and 2003-2005 but with an overall decrease between 1995 and 2010. The map in Figure 2.2 shows the LHA's coloured by loading value.

Model	Covariance structure	m	logLik	delta.AIC
1	**diagonal and equal**	**1**	**-20,745**	**0**
2	diagonal and equal	2	-20,705	37
3	diagonal and equal	3	-20,672	87
6	diagonal and unequal	1	-20,741	108
7	diagonal and unequal	2	-20,701	146
4	diagonal and equal	4	-20,647	152
8	diagonal and unequal	3	-20,667	197
5	diagonal and equal	5	-20,620	212
9	diagonal and unequal	4	-20,641	261
10	diagonal and unequal	5	-20,615	323
11	unconstrained	1	-19,815	2195
12	unconstrained	2	-19,773	2262

TABLE 2.1: Results from DFA for temporal trend. m: number of common trends estimated, logLik: log likelihood, delta.AIC: difference between AIC values of best model and other model.

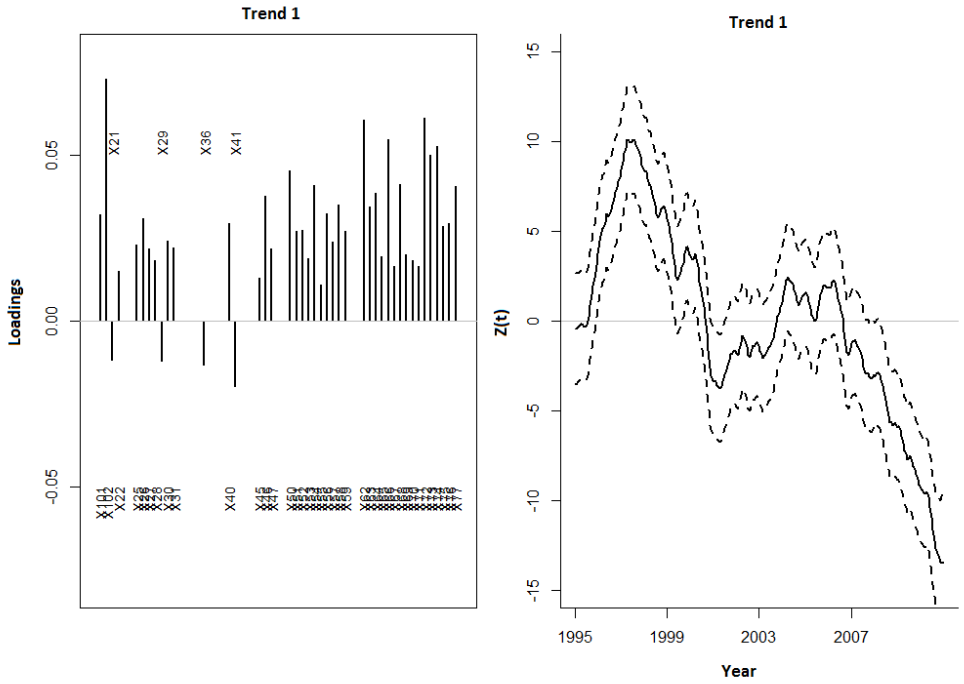


FIGURE 2.1: Loadings for model of temporal trend with 1 common trend and diagonal and equal covariance structure. Text indicates LHA identifier e.g. X102 represents LHA 102 (left) (see Figure 1.1). Pattern of common trend (right, solid line) with ±2 standard errors (dashed line). The y-axis shows the value of the common trend at time t ($\mathbf{Z}_t$ in Equation(1.17)) and is unitless.

2.2.1.2 Seasonal pattern

DFA models were fitted to the 59 noisy seasonal patterns fitted using (2.9). Up to 5 common trends were estimated and diagonal and equal, diagonal and unequal, and unconstrained covariance structures were used. The AIC values for the DFA models can be found in Table 2.2 where it can be seen that the DFA model with 1 common trend and a diagonal and equal covariance structure is the best fitting model. Figure 2.3 shows the loadings (left panel) and the shape (right panel) of the common seasonal pattern estimated for the 59 LHA's. As expected, the common seasonal pattern indicates that log(TON) is at a minimum between August and September and has maximum values at the beginning and end of the year in winter months. A map of the loadings is given in Figure 2.4 and the pattern is similar to that seen for the temporal trend so the spatial pattern is the same across the country for both temporal trend and seasonal pattern.

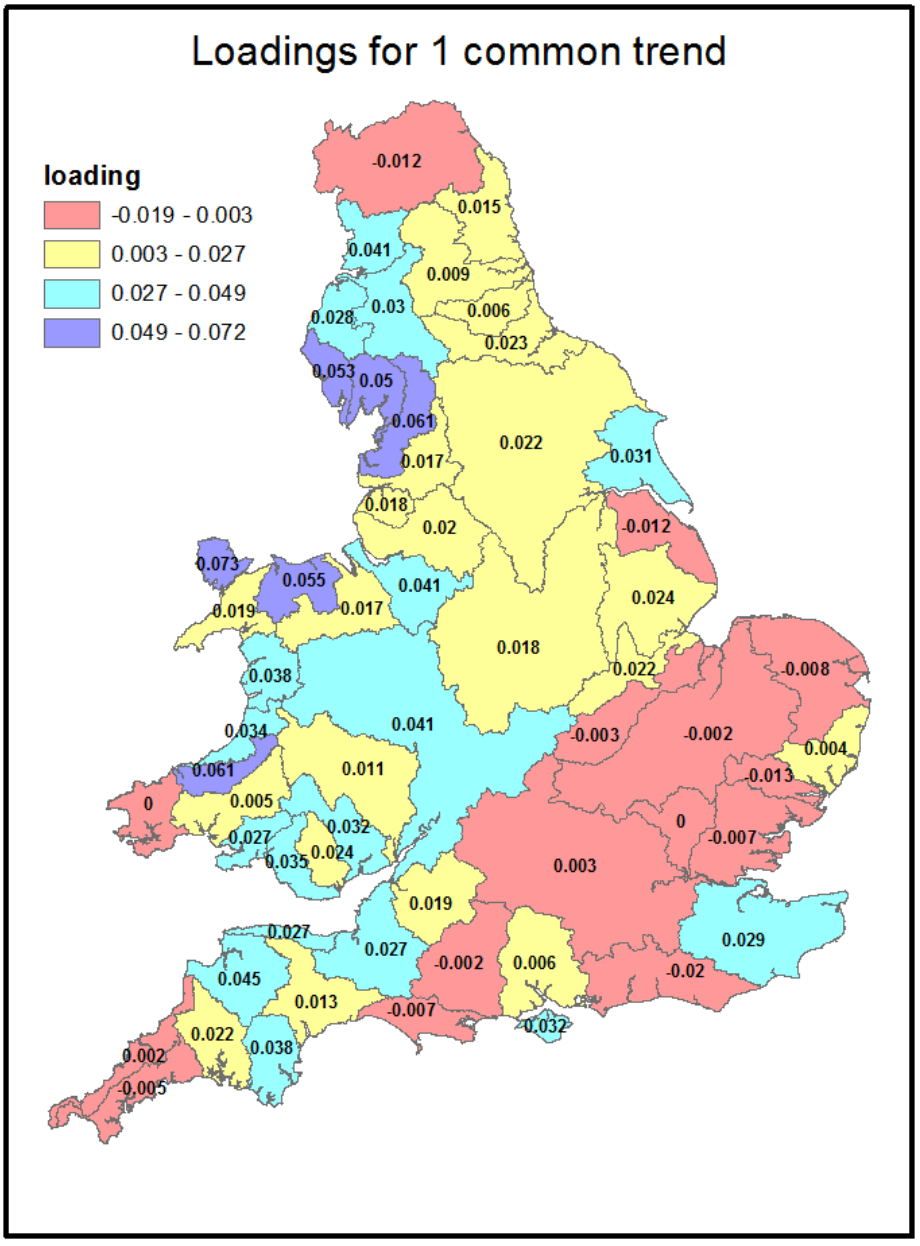


FIGURE 2.2: Map showing loadings for each LHA for model with 1 common trend and diagonal and equal covariance matrix.

model	Covariance structure	m	logLik	delta.AIC
1	**diagonal and equal**	**1**	**-24,295**	**0**
2	diagonal and equal	2	-24, 254	35
6	diagonal and unequal	1	-24, 273	71
3	diagonal and equal	3	-24, 221	84
7	diagonal and unequal	2	-24, 229	102
4	diagonal and equal	4	-24, 192	141
8	diagonal and unequal	3	-24, 195	151
9	diagonal and unequal	4	-24, 168	211
5	diagonal and equal	5	-24, 172	214
10	diagonal and unequal	5	-24, 147	285
11	unconstrained	1	-23, 329	2027
12	unconstrained	2	-23, 281	2076

TABLE 2.2: Results from DFA for seasonal pattern. m: number of common trends estimated, logLik: log likelihood, delta.AIC: difference between AIC values of best model and other model.

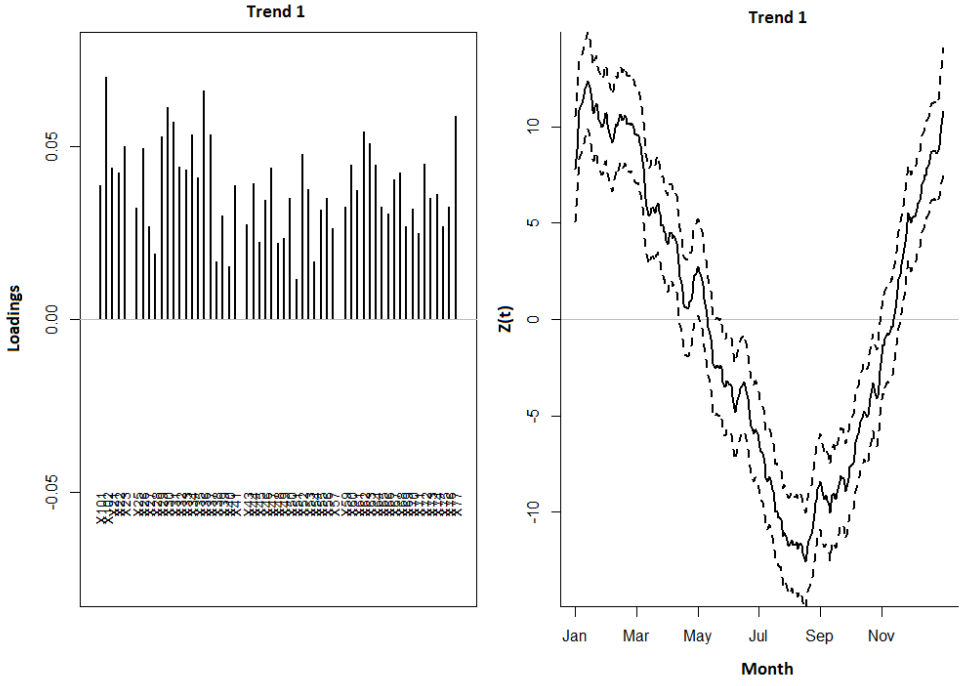


FIGURE 2.3: Loadings for model of seasonal pattern with 1 common trend and diagonal and equal covariance structure. Text indicates LHA identifier e.g. X102 represents LHA 102 (left) (see Figure 1.1). Pattern of common trend (right, solid line) with ±2 standard errors (dashed line). The y-axis shows the value of the common trend at time t ($\mathbf{Z}_t$ in Equation(1.17)) and is unitless.

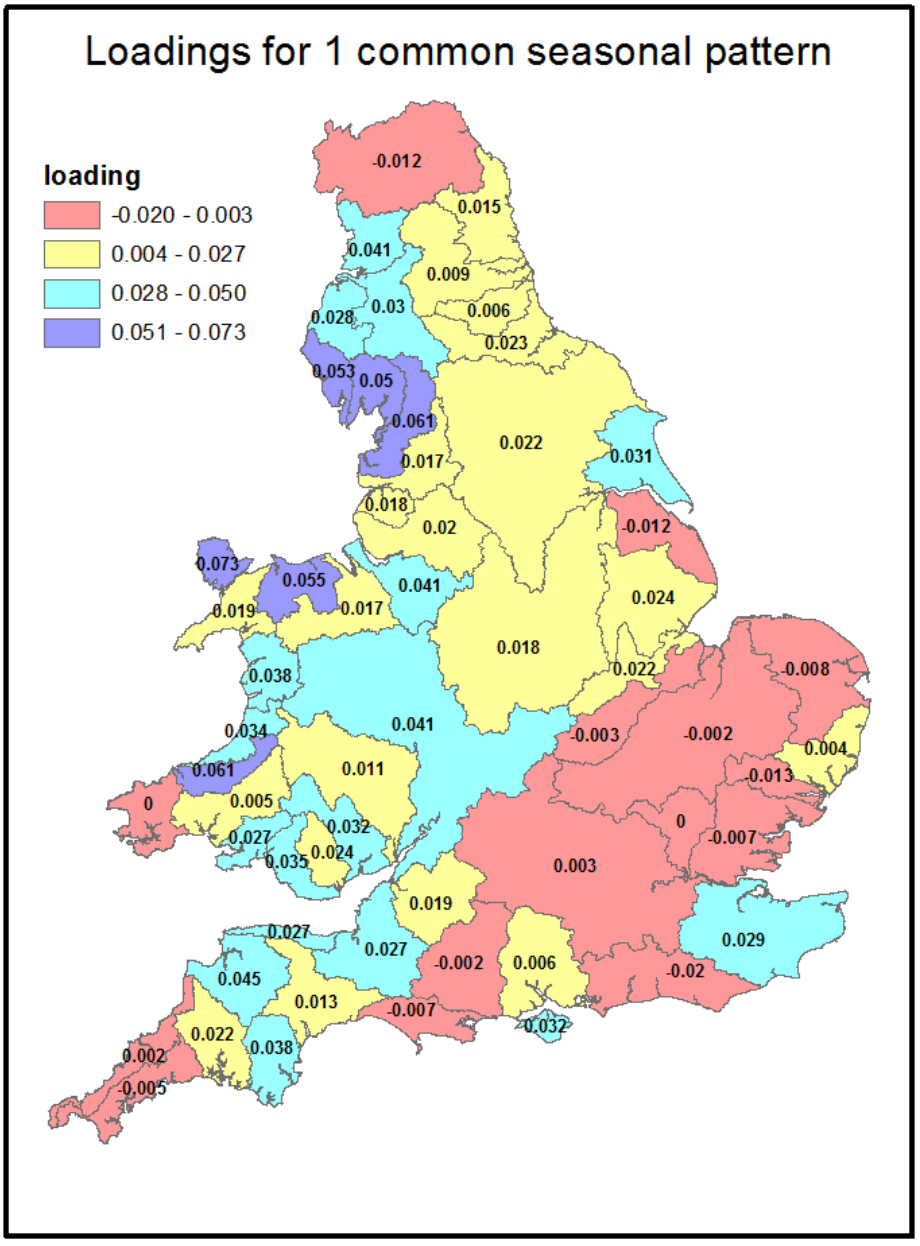


FIGURE 2.4: Map showing loadings for each LHA for model with 1 common trend and diagonal and equal covariance matrix.

2.2.1.3 Summary: common temporal trends and seasonal patterns

A temporal trend and seasonal pattern were modelled for each LHA and these were compared using dynamic factor analysis to provide some initial exploration of the data and to investigate a statistical technique suitable for identifying dominant temporal patterns. It was shown that a model with a single common trend and a diagonal and equal covariance matrix was suitable for temporal trend and seasonal pattern respectively to describe the temporal evolution of nitrates in 59 LHA's. The diagonal covariance matrix suggests independence between the LHA's which is in line with the construction of LHA boundaries. Although DFA does not directly compare curves, this section has been shown DFA to be a useful technique providing evidence of common behaviour among LHA's over time.

2.2.2 Comparing different estimates of temporal trend and seasonal pattern for a single LHA

This section moves from considering temporal trends and seasonal patterns between LHA's to common patterns found between monitoring sites within a single LHA. Section 2.2.1 estimated a common trend for temporal trend and seasonal pattern that was shared among the 59 LHA's in England and Wales. In this section a single LHA will be investigated and comparisons made between the temporal trend and seasonal pattern estimated using some of the methods described in Chapter 1: additive models fitted using the `mgcv` and `INLA` packages in `R`, functional data analysis and dynamic factor analysis (DFA). This is not intended to be a direct comparison of the four modelling procedures which would require knowledge of the true underlying temporal trend and seasonal pattern. The work in this section is a comparison of the shapes of curves estimated from four different models and is included here as an interesting application of the comparison techniques (DBF statistic and curvature) described earlier in this chapter. The aim of this study is to investigate whether the choice of statistical method used by analysts matters and to understand how estimates of temporal trend and seasonal pattern differ between models. A description of how the four modelling procedures under consideration were implemented will be given followed by a novel comparison of the shapes of estimated temporal trends and seasonal patterns. In the final Comments section there is some discussion on the different ways in which spatial structure can be

incorporated into each of the methods compared and general guidelines are proposed to aid selection of an appropriate method to use when estimating the temporal trend and seasonal pattern within an LHA.

In this section, four statistical methods are applied to data from a single LHA and the estimated temporal trends and seasonal patterns are compared using visual distance and curvature, described earlier in this chapter. The data for this investigation come from LHA 61 which contains 28 monitoring sites and was chosen since Brodgar software (`http://www.brodgar.com/`), used to implement DFA, can model approximately 30 time series at most. This LHA also contains several small river networks with few flow connected monitoring sites so spatial correlation based on river network structure is likely to have little influence here.

An additive model as in (2.9) was fitted to all of the data from LHA 61 using the `mgcv` package in `R` and the temporal trend and seasonal pattern extracted from the model. The smooth temporal trend and seasonal pattern were evaluated at 250 and 300 time points respectively. Next, an additive model was fitted using the `INLA` package where time was indexed by month and year rather than day and year.day in (2.9). A cubic interpolation spline was used to smooth the trend curve and interpolate values at the same dates as for the other models. Spatial structure was modelled using a triangulated mesh of the LHA, created with settings chosen for computational efficiency and based on recommendations in Cameletti et al. (2011) and personal communication with one of the authors (Finn Lindgren). For functional data analysis an additive model as in (2.12) was fitted to the data for each of the 28 monitoring sites separately and using the same notation and values for k_2 and k_3 as in (2.9). The temporal trend and seasonal pattern were extracted from each model and evaluated at the same time points as for the additive model fitted to all of the data within LHA 61. An estimate for the LHA temporal trend and seasonal pattern was created by taking the pointwise mean at each time point. Noisy versions of the 28 smooth temporal trends and seasonal patterns were created by adding a random draw from a normal distribution with mean of zero and standard deviation equal to the standard deviation of the residuals from the additive model. Finally, DFA was applied to the noisy temporal trends and seasonal patterns for all combinations of one or two common trends and diagonal and unconstrained covariance structure. The diagonal terms for both covariance structures were allowed to vary and the best fitting

model was chosen using AIC. The curves were smoothed using a cubic smoothing spline with 11 knots for both the trend and seasonal curves.

$$\begin{aligned} \log(\text{TON})_{tr} &= s_2(t, k_2) \\ &+ s_3(r, k_3) \\ &+ \varepsilon_{tr} \end{aligned} \tag{2.12}$$

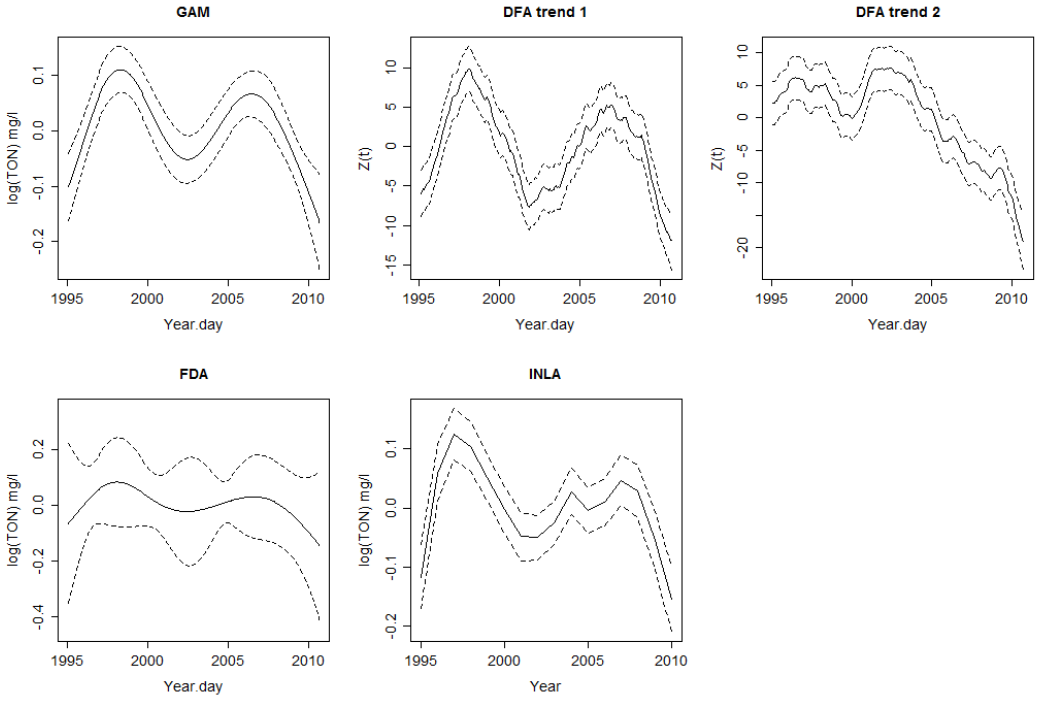


FIGURE 2.5: Temporal trends estimated for LHA61 using GAM, FDA, INLA, and DFA (solid line) with error bands (dashed lines). y-axis for GAM, FDA, and INLA is log(TON) mg/l and for DFA trend 1 and 2 is Z(t) (see Equation (1.17)). x-axis for INLA is Year and for GAM, FDA, and DFA is Year.day.

2.2.2.1 Temporal trend

Figure 2.5 shows plots of the temporal trends estimated from four methods with corresponding 95% confidence intervals and for INLA a 95% credible interval. Note that these plots are not comparable as the y-axis for DFA trend 1 and 2 is Z(t) (Equation (1.17)) while the y-axis for GAM, FDA and INLA is log(TON) mg/l. Also, the x-axis for INLA is Year while the x-axis for GAM, FDA and DFA is Year.day (see Figure 2.6 for normalised curves that are suitable for comparison where all x- and y- axes are on the same scale). The best DFA model has two common trends and a diagonal covariance matrix. The temporal trend estimated from all methods (except DFA trend 2) shows that from

1995 log(TON) increased, followed by a decrease before increasing again around 2002 and then decreasing in recent years. DFA trend 2 looks quite different from the other four curves with a peak in log(TON) between 2000 and 2005 but still suggests a decrease in log(TON) in recent years. The shapes of these curves will now be compared using the DBF statistic and plots of curvature.

Calculation of $\text{DBF}_{\Delta_{d_V}}$ requires normalised data and this is useful here since the trends and seasonal patterns are not all estimated on the same temporal scale. For example, common trends estimated using DFA are not on the log(TON) scale; the temporal trend and seasonal pattern estimated using INLA are based on time points relating to Year and Month respectively whereas those estimated using GAM, DFA and FDA relate to Year.day and Day. Normalising the curves before they are compared means that the shapes of curves estimated on different scales can be compared. The temporal trends estimated from INLA and DFA were smoothed prior to normalisation for shape comparison so as to remove small scale variability thus ensuring comparisons were based on the overall shape of the curves. The normalised curves are shown in Figure 2.6.

The DBF statistic (Minas et al., 2011) is used to compare groups of curves so in order to use this to compare the shapes of the five curves in Figure 2.6 15 replicates are simulated for each curve by adding a random draw from a uniform distribution with parameters equal to the lower and upper values of the 95% confidence bands at each time point and smoothing the points using a cubic smoothing spline followed by normalisation using (2.2). The five trend patterns in Figure 2.6 are tested for similarity using $\text{DBF}_{\Delta_{d_V}}$ and the results show that the five curves are significantly different ($p < 0.001$).

The DBF statistic informs the user if differences exist between curves but not which curves are different and what the nature of any differences are. Curvature can be used to describe and compare shapes of curves and so curvature was calculated for the temporal trends in Figure 2.6 and is plotted in Figure 2.7 with 95% confidence intervals, the calculation of which was described in Section 2.2. High (positive or negative) values of curvature relate to sharp peaks or troughs while low values correspond to flatter ones. Positive curvature corresponds to a U shaped bend and vice versa. From the plot the main differences in shape are the time of the turning points (location of peaks and troughs) and how sharp or shallow the turning points are (height of peaks and troughs). GAM, FDA and DFA trend 1 are most similar in terms of the locations of turning

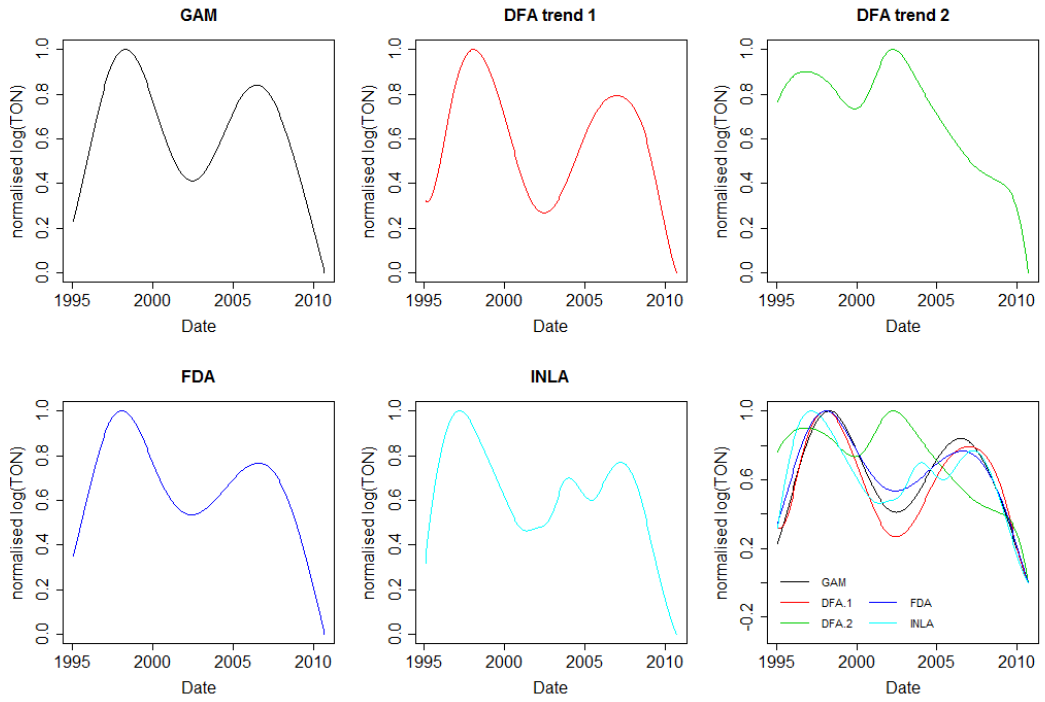


FIGURE 2.6: Normalised smooth temporal trends estimated using GAM, FDA, DFA and INLA.

points. The INLA curve follows a similar pattern (in terms of number of turning points) as GAM, FDA and DFA trend 1 but the locations of the turning points are different. DFA trend 2 follows a similar pattern to the other four curves but the locations of the turning points are much earlier although of similar magnitude to the FDA curve at the second and third turning points. The comparison of DFA trend 1 and DFA trend 2 curvature curves with the GAM, FDA and INLA curvature curves is particularly useful in identifying which of the two DFA trends identifies the main temporal trend. DFA estimates common trends simultaneously so it is not always obvious which of the common trends models the dominant pattern in the data. In the case presented here, DFA trend 1 picks out the main temporal trend as estimated by the other three methods, with DFA trend 2 highlighting a small peak in log(TON) between 2000 and 2005. Interestingly, the temporal trend estimated from INLA looks to be a mixture of the two DFA trends.

2.2.2.2 Seasonal pattern

Figure 2.8 shows the seasonal patterns estimated for LHA 61 using GAM, DFA, FDA and INLA. Note that these plots are not comparable as the y-axis for DFA trend 1 and 2 is Z(t) (Equation (1.17)) while the y-axis for GAM, FDA and INLA is log(TON) mg/l.

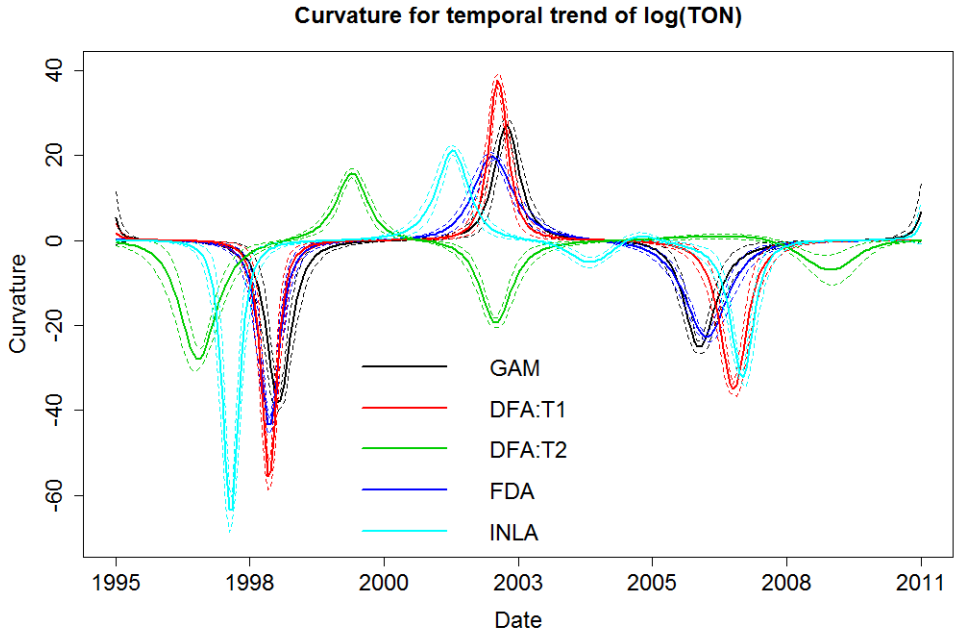


FIGURE 2.7: Curvature of temporal trends estimated using GAM, DFA, FDA and INLA.

Also, the x-axis for INLA is Month while the x-axis for GAM, FDA and DFA is Day of year (see Figure 2.9 for normalised curves that are suitable for comparison where all x- and x- axes are on the same scale). The best DFA model had two common trends and a diagonal covariance matrix. The plots show log(TON) is higher at the beginning and end of the year, with a yearly minimum around day 270. DFA trend 1 estimates minimum log(TON) occurs around day 215 and DFA trend 2 estimates minimum log(TON) occurs around day 270, even though the seasonal pattern for DFA trend 2 looks different from the other four curves.

Figure 2.9 shows the smoothed and normalised seasonal patterns estimated from the four methods. As with the temporal trends, the INLA seasonal pattern has been interpolated from monthly time points to days using interpolating cubic splines but a periodic cubic spline has been used to ensure continuity between the end and beginning of the year. The common trends estimated from DFA have been smoothed using periodic cubic smoothing splines with 11 knots.

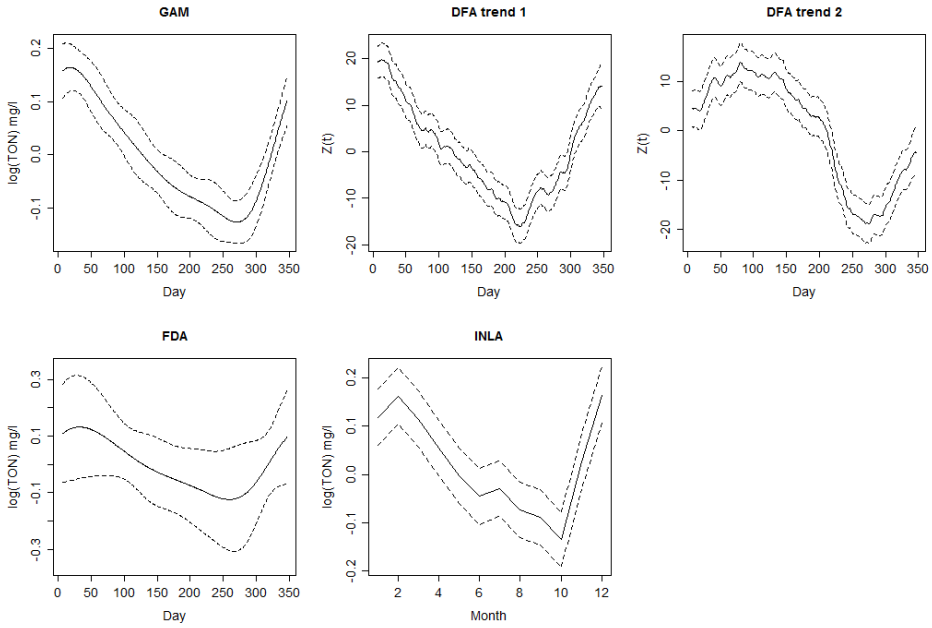


FIGURE 2.8: Seasonal patterns estimated for LHA61 using GAM, FDA, INLA, and DFA (solid line) with error bands (dashed lines). y-axis for GAM, FDA and INLA is log(TON) mg/l and for DFA trend 1 and 2 is Z(t) (see Equation (1.17). x-axis for GAM, FDA, and DFA is Day of year and for INLA is Month).

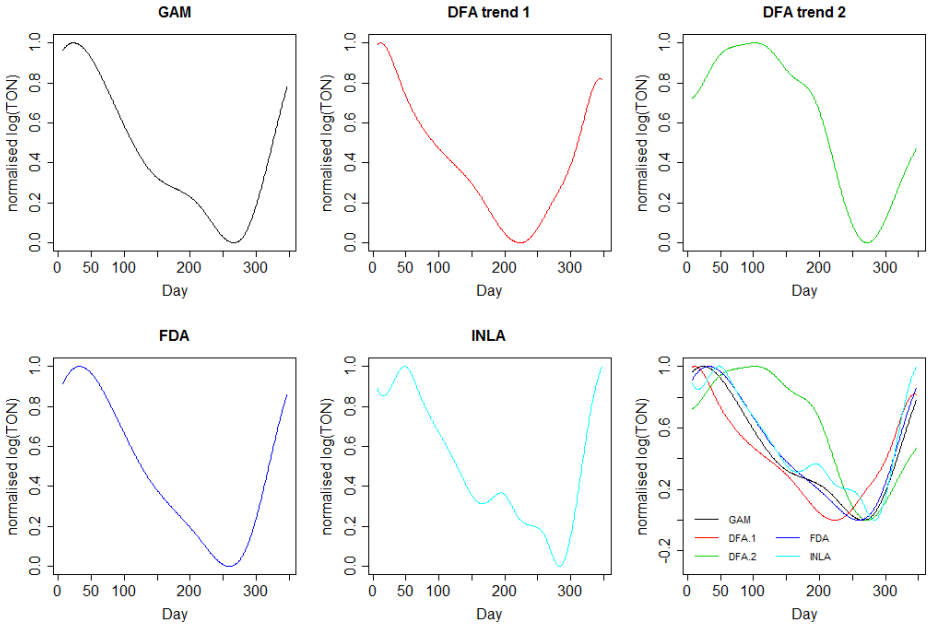


FIGURE 2.9: Normalised smooth seasonal patterns estimated for LHA 61 using GAM, DFA, INLA, and FDA.

The $\text{DBF}_{\Delta_{d_V}}$ test statistic was calculated ($p < 0.001$) and showed that the five curves were significantly different. Curvature was calculated for each of the five seasonal patterns and is plotted with 95% confidence intervals in Figure 2.10. GAM and FDA are very similar with the FDA curve being slightly flatter (lower curvature) than the GAM curve. The INLA seasonal pattern picks up a small dip in log(TON) that occurs about halfway through the year before the minimum log(TON) about three quarters of the way through the year and the GAM seasonal pattern also reflects this but not as prominent as for INLA. DFA trend 2 is most similar to the GAM, FDA and INLA curves in terms of curvature at the time of year where log(TON) is at a minimum but this curve also suggests log(TON) occurs later in the year than the other curves. DFA trend 1 shows minimum log(TON) occuring earlier in the year than the other curves.

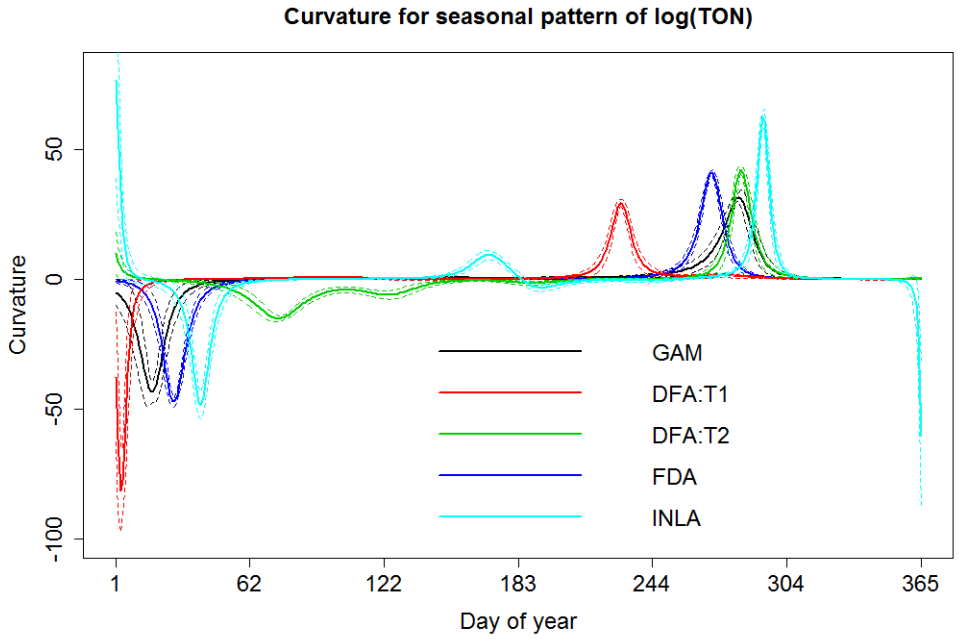


FIGURE 2.10: Curvature of normalised seasonal patterns for LHA 61 estimated using GAM, DFA, FDA and INLA.

2.2.2.3 Summary: comparing different estimates

Temporal trends and seasonal patterns were modelled for a single LHA using four different statistical methods. The estimated temporal trends and seasonal patterns from the four methods were shown to be significantly different using the DBF test statistic and inspection of curvature plots aided identification of the differences between the shapes of

curves. Differences were mostly due to the location of turning points such as the point in the year where minimum nitrate levels occur.

Significant differences between the shapes were identified as sections of the curvature curves with non-overlapping confidence intervals. Some improvements could be made to the calculations used to construct the curvature confidence intervals. Adjustments to the confidence interval calculation could be made for multiple comparisons by using the distribution theory in Chaudhuri and Marron (1999) and Hannig and Marron (2006). Further improvements could be made by taking into account the dependent nature of time series data. See for example Park et al. (2004), Rondonotti et al. (2007) and Park et al. (2009) for details of SiZer for dependent (time series) data. The confidence bands were calculated assuming independence between the numerator and denominator in (2.3) and so estimation could also be improved by including a covariance term in the Taylor expansion of the curvature formula, which would allow for dependence between the numerator and denominator of $\kappa(t)$.

A sensitivity analysis showed the DBF test to be robust to the choice of smoothing parameter. Curvature however, being a measure of 'wiggliness', is sensitive to choices made when constructing smooth estimates of temporal trends and seasonal patterns of log(TON). The choice of smoothing parameter used to estimate these curves was chosen so as to highlight interesting features in the curve and smooth out highly localised variability. This work aims to compare long term trends and seasonal patterns of log(TON) so it is the main features of the curves that are of interest as opposed to variability over a very short time period. It should be noted that changing the level of smoothing would change the curvature plots. The use of confidence intervals developed for curvature in this Chapter however means that significantly non-zero curvature can still be identified for different values of the smoothing parameter.

An explanation for the differences between estimates is that each of the statistical methods incorporate spatial information in different ways. For example, the `mgcv` model captures mean spatial pattern through a bivariate smooth spatial surface and assuming independence between monitoring sites. Although spatial correlation could be incorporated into the model via a mixed model representation there is currently no method for implementing this for spatial correlation due to river network topology and this could form the basis of future research. This chapter aimed to investigate statistical methods

for curve comparison and it was decided to fit an `mgcv` additive model assuming independence between monitoring sites since an appropriate form for the spatial covariance structure is considered in detail in Chapters 3 and 4. The model assuming independence provided an adequate estimate of the temporal trend and seasonal pattern in LHA61 and would be affected by spatial correlation only in that the standard error of the estimates would be narrower than had spatial correlation been accounted for. Since this chapter was interested in the shape of the estimated temporal trend or seasonal pattern and no inference was carried out on these estimates, the model was applied assuming spatial independence. The `inla` additive model incorporates spatial correlation using a neighbourhood matrix and sparse precision matrix rather than a dense Matérn spatial covariance matrix but this does not account for flow direction in applications to river network data. The functional mean was calculated as the pointwise mean of temporal trends and seasonal patterns estimated for each monitoring site in LHA61 individually. This method did not incorporate spatial information but allowed different estimates to be computed for each monitoring site. Finally, the DFA approach incorporates spatial information depending on the specified form of the covariance matrix where off-diagonal elements capture spatial structure not modelled by the common trends.

The four statistical methods compared in Section 2.2.2 estimate temporal trends and seasonal patterns with some similarities and clear differences. Analysts will usually use their preferred method rather than consider different interpretations that might occur from using different statistical methods. From the results in this chapter it seems that GAM and FDA provide estimates that are most similar. The best fitting DFA model for temporal trends and seasonal pattern respectively had two common trends, one of which was similar in terms of curvature to the GAM and FDA estimates with the other highlighting deviations from the mean curve. Interestingly, the INLA curve appears to be a mixture of the two DFA common trends. Going forward, it is recommended that analysts first decide if they wish to estimate a single temporal trend or seasonal pattern to capture nitrate levels across multiple monitoring sites or if they are interested in identifying more than one dominant pattern. Dimension reduction techniques such as DFA are more suitable than the other methods compared in Section 2.2.2 for estimating multiple temporal patterns for a single LHA. For estimating a single mean curve then GAM or INLA models are recommended. The FDA curve is flatter than the other curves and variation is masked by taking the pointwise average of several curves in this example.

The INLA curves capture more variation in shape than the GAM curves but this might be due to the level of smoothing carried out when fitting the GAM model, although the GAM model still captures the main pattern in the data and was computationally simpler to implement than the INLA model. Analysts should therefore choose between GAM and INLA depending on their own computational skills and pay careful attention to the choice of smoothing parameter in the GAM model so that interesting features of the estimated curves are not masked by oversmoothing. DFA is the most suitable method for determining the presence of multiple temporal patterns of interest. DFA however is not suitable for large numbers of time series, especially when an unconstrained covariance matrix is included the model so Chapter 5 of this thesis, while focussing on dimension reduction techniques, considers principal components analysis which is more suitable for large numbers of time series than DFA. In the case of large datasets GAM and INLA models are much more computationally efficient and existing methodology allows models to include spatial correlation. Spatial correlation options at present however do not include the spatial correlation structure specific to river networks, discussed in greater detail in Chapter 3 and Chapter 4.

2.3 Comments

This chapter has presented exploratory analysis of the data provided by the EA and showed that the temporal trend and seasonal pattern for all of the 59 LHA's in England and Wales could be modelled using a DFA model with a single common trend and diagonal and equal covariance matrix. This suggests that LHA's, on average, behave exhibit similar temporal patterns over time in terms of mean pattern and variance.

This Chapter also proposed a novel approach for comparing temporal trends and seasonal patterns of log(TON) by combining methods developed for use in the areas of genetics and computer vision. The DBF test was adapted to include simulated curves rather than replicates. This approach is useful for comparing a small number of curves but it was noted that the curvature plots are sensitive to the choice of smoothing parameter used to construct the curves, although the use of confidence intervals means that significant differences can still be identified. Some suggestions were made to improve the confidence interval calculations.

Finally, a discussion of the ways in which each of the methods compared in Section 2.2.2 incorporates spatial information showed that none of the methods explicitly account for river network structure. It was also noted that different statistical methods provide different estimates of temporal trends and seasonal patterns and this might be because of the way in which spatial information is incorporated by these methods. Some recommendations were also made for future analysts to help them decide which statistical method to use when estimating temporal trends and seasonal patterns. The choice of method depends on the scale of the data (since implementation using `INLA` might have computational advantages for estimating spatial correlation compared to `mgcv`, and DFA, particularly when implemented using Brodgar software can only model up to 30 time series), and whether the analyst was required to estimate an average temporal trend or seasonal pattern, or assess whether more than one dominant pattern was present.

Chapter 3

Spatial modelling in a river network

'Everything is related to everything else, but near things are more related than distant things'

Tobler's First Law of Geography

This chapter aims to introduce statistical methods suitable for modelling spatial correlation for non-Euclidean measures of distance, and demonstrate the application of methods recently developed specifically for data collected on river networks. This is a novel application to data collected on rivers in England since these methods have not been previously applied to UK data. This chapter also aims to investigate spatial correlation at different scales for data collected on river networks and also to show how appropriate spatial covariance models can be used to make predictions at unsampled locations on the network.

Temporal trends and seasonal patterns of water quality data were estimated in Chapter 2 for 59 LHA's, assuming independence in space. This was a reasonable assumption since LHAs are defined to be independent of neighbouring areas. Section 2.2.2 shifted the focus to water quality within a single LHA. Nitrate levels recorded at monitoring sites within an LHA are spatially dependent in two ways: by water flowing from upstream sites to downstream sites and by runoff from land across the LHA. Chapter 1 described spatial covariance functions, used to estimate how data recorded at one monitoring site are related to data recorded at other monitoring sites nearby, based on Euclidean

distance between monitoring sites. Euclidean distance is used to calculate the (binned) semivariance and parameters of the spatial covariance function are estimated. For data recorded on river networks however, Euclidean distance might not be appropriate when describing spatial relationships in the data. Instead, distance measured along the river network, called stream distances, might be more appropriate. It is tempting to simply substitute Euclidean distance with stream distance in the spatial covariance function but ver Hoef et al. (2006) explains why this is not appropriate and introduces the Tail-up model, a special type of spatial covariance function suitable for data collected on river networks. This chapter begins with a discussion of the development of statistical models suitable for river network data and presents an investigation into the most appropriate spatial covariance function to use for data of this type. Spatial covariance functions suitable for non-Euclidean distances (in this case distance along the river network) are introduced, after which several spatial covariance models are fitted to river network data at a variety of spatial scales. Finally, the best fitting spatial covariance model is used to make predictions at unobserved locations on a river network and compared to predictions based on non-spatial and Euclidean distance based models.

3.1 Spatial models for river networks

This section will describe the development of attempts in the literature to account for spatial correlation for non-Euclidean distance measures, followed by a detailed description of recently developed methods that can be used to model spatial correlation for river network data, accounting for flow direction and relative influence of upstream monitoring sites on downstream locations.

As described in Chapter 1, statistical models exist and have been widely described and implemented in the literature to model data collected from monitoring sites placed across a spatial region. The spatial covariance models are based on Euclidean distance between pairs of monitoring sites but Euclidean distance might not always be the best distance metric to use. For example, Curriero (2007) suggests geodetic distances, water distance measures and travel times as possible non-Euclidean distance metrics and shows that naively replacing Euclidean distance in standard spatial covariance functions with a non-Euclidean measure of distance can result in a variance-covariance matrix that is not positive definite - a requirement for a valid variance-covariance matrix. It can be shown

that the Exponential covariance function is the only standard covariance function that yields a positive definite variance-covariance matrix when Euclidean distance is replaced with non-Euclidean distance such as 'city block' (Curriero, 2007) or stream distance (ver Hoef et al., 2006). Curriero (2007) discusses the concept of 'isometric embedding in a Euclidean space' and explains that stream distances between monitoring sites are equivalent to Euclidean distances calculated from a set of locations $\mathbf{s}^* \in \Re^1$, located along the stretched out stream that is now in one dimension, assuming the stream does not branch off.

Little et al. (1997) compare kriging predictions from semivariograms based on Euclidean distance between pairs of sites and 'in water' distance between pairs of sites by simply plugging stream distances into the covariance models being considered and note that there was little difference in predictions from these different distance metrics. They also note that Euclidean distances and stream distances in their data set were very similar and so it is reasonable to assume there will be little difference in results based on Euclidean and 'in water' distances. Rathbun (1998) also discusses the use of water distance compared to Euclidean distance but for observations at monitoring sites in an estuary so once again there is no need to account for the branching structure found in river networks. Cressie and Majure (1997) use stream distance with the exponential covariance model in a spatio-temporal model to investigate livestock waste in streams. The prediction sites however are located on a single stretch of river (with one exception) and the neither the branching structure seen in river networks nor direction of flow are accounted for.

Gardner et al. (2003) model stream temperatures in a branched river network and compare the prediction performance of three distance metrics: Euclidean distance, stream distance and a weighted stream distance based on stream order that accounts for the fact that small tributaries will have little influence on the stream temperature in the main stem of the river. It is noted that the nugget value for the weighted stream distance semivariogram is 0 suggesting that small scale variation has been accounted for by the weights thus allowing large changes in stream temperature over short distances (likely to be a result of moving between main stem and tributary). Gardner et al. (2003) use the spherical covariance function but do not check if this is in fact a valid covariance model when using stream distances, as recommended by Curriero (2007) and Rathbun (1998). Gardner et al. (2003) also discuss the fact that the contribution from a tributary affects

stream temperature on the main stem in both downstream and upstream directions i.e. flow direction has not been accounted for.

ver Hoef et al. (2006) and Cressie et al. (2006) introduce linear models based on a moving average approach that incorporate flow direction, stream distance and suitable weights at confluences, thus improving on earlier attempts to model spatial correlation in stream networks (Gottschalk (1993), Sauquet et al. (2000)). ver Hoef et al. (2006) show that weighted models perform better (in terms of AIC and RMSPE) than unweighted models based purely on stream distance and suggest that reach contributing area or stream order are suitable weights. Peterson et al. (2007) explains in detail how to produce the data required for the Tail-up covariance models (distance matrices based on stream distance and connectedness information). Generalized linear models that incorporate stream distance based covariance models are introduced in Peterson and ver Hoef (2010).

Garreta et al. (2010) develop valid 'tail down models' based on stream distance for flow unconnected points. The tail-down model is proposed as a suitable model for variables of interest that can move upstream such as fish populations. The tail-down model shares properties of the tail up models (stationarity along stream and conditional independence of branches) developed in (Monestiez et al., 2005). Garreta et al. (2010) and ver Hoef and Peterson (2010) discuss the merits of a 'hybrid' structure combining Tail-up and Tail-down covariance models and Peterson and ver Hoef (2010) discuss combining the Tail-up and Euclidean covariance models. Money et al. (2009) introduce 'composite Euclidean-river distance', a weighted sum of stream distance and Euclidean distance between two monitoring sites and use this in an isotropic exponential-power covariance model as part of a spatio-temporal model for dissolved oxygen (DO), but do not use a flow-connected covariance model since few of their monitoring sites are flow connected on any given date. Cressie et al. (2006) also discusses a mixture of Euclidean distance and stream distance covariance functions controlled by smoothing parameter $\lambda \in [0, 1]$.

Peterson and ver Hoef (2010) discuss the merits of using a 'variance components' approach to model spatial autocorrelation in stream networks, accounting for spatial autocorrelation at different spatial scales i.e. from the individual stream segment to the full network scale and is equivalent to adding a random effect term to the model. It is suggested that unobserved or unmeasured explanatory variables might be accounted for

by combining Euclidean distance based covariance models with stream distance models. Peterson and ver Hoef (2010) (Appendix B) recommend that the spatial covariance models are fitted using REML since ML can produce biased covariance estimates.

The work discussed in this section describe efforts to account for stream distance and flow connectedness while calculating spatial covariance but this is not the only way to account for river network structure within statistical models. Whitehead et al. (2011) develop a process based model for phosphorous that can include any number of tributaries of any stream order while Webb and Padgham (2013) represent river networks as directed graphs with nodes being habitat patches and use these graphs to build a theoretical model to investigate the effect of network structure on population abundance and persistence. Skøien et al. (2006) develop a procedure they call 'top-kriging' for interpolation on river networks that accounts for the size of area in which a stream segment exists as well as the nested nature of those areas. Top-kriging is used to model runoff - a combination of continuous spatial processes (rainfall, soil characteristics) and stream network processes (such as flow, stream temperature). Top-kriging is compared to the methods of ver Hoef et al. (2006) and kriging based on Euclidean distance by Laaha et al. (2012) who conclude that neither purely Euclidean distance or purely stream distance based kriging methods are satisfactory and that top-kriging gives a good balance between the two. Laaha et al. (2012) however do not directly compare the work of ver Hoef et al. (2006) to top-kriging using the same dataset.

An alternative approach was recently taken by O'Donnell et al. (2014) who adapt kernel methods for stream distance and penalized splines for a finite discrete approximation to river networks and use these to estimate spatial trend rather than the covariance function. The idea is to borrow strength from nearby sites while respecting the network structure. Rushworth et al. (2015) compares the penalized spline approach to the work in ver Hoef and Peterson (2010) and find that the penalized spline approach is preferable when data are sparse in space and also in terms of computational efficiency since the model described in ver Hoef et al. (2006) relies on a dense distance matrix which causes computational time to increase with increasing numbers of monitoring sites. The work on penalized splines in O'Donnell et al. (2014) would be computationally demanding if space-time interactions were to be modelled, since tensor products of smooths require the estimation of many basis coefficients, but Rushworth et al. (2015) show that computations can be carried out efficiently if the connectedness information in the river

network is represented as a binary neighbourhood matrix for discretised stream segments rather than river distance between monitoring sites. The spatial covariance approach of ver Hoef and Peterson (2010) however has the advantage when estimating fine scale dependence and can also be used for binary or count data. The methods in Rushworth et al. (2015) allow for the estimation of smooth non-linear fixed effects as described in Chapter 1 but require data on flow for every stream segment which might not be available, although this can be modelled and estimates used in place of recorded flow information.

In this chapter the spatial covariance approach of ver Hoef et al. (2006) and ver Hoef and Peterson (2010) is used to investigate which form of spatial covariance function is most suitable for a densely monitored complex monitoring network, following the recommendations in Rushworth et al. (2015). Spatial dependence is modelled for the Trent LHA. A description of the spatial covariance models based on stream distance and flow connectedness is given in Section 3.1.1, followed by a study to investigate the most suitable spatial covariance function under three scenarios (1) a single network, (2) several networks within a single LHA and (3) multiple networks across multiple LHA's. Finally, the best covariance function will be used to predict nitrate levels at unsampled locations to investigate how the predictions and corresponding standard errors differ between different spatial covariance functions.

3.1.1 Valid covariance models incorporating stream distance

Following the notation in ver Hoef and Peterson (2010), a large class of autocovariances can be developed by creating random variables as the integration of a moving average function over a white noise process $W(x)$:

$$Z(s|\boldsymbol{\theta}) = \int_{-\infty}^{\infty} g(x - s|\boldsymbol{\theta}) dW(x), \tag{3.1}$$

where x and s are locations on the real line and $g(x|\boldsymbol{\theta})$ is the moving average function defined on $\mathcal{R}^1$. Cressie and Pavlicová (2002) present a method for expressing any valid autocovariance model as a moving average function. Using this moving average construction, a valid autocovariance between $Z(s)$ and $Z(s+h)$ can be expressed as

$$C(h|\boldsymbol{\theta}) = \int_{-\infty}^{\infty} g(x|\boldsymbol{\theta})g(x-h|\boldsymbol{\theta})dx. \tag{3.2}$$

ver Hoef et al. (2006) and Cressie et al. (2006) use this moving average construction to develop 'Tail-up' models where the moving average functions are positive only upstream from a monitoring site. When the function reaches a confluence it continues upstream but splits according to spatial weights assigned to each branch. The weights can be proportional to flow volume or other proxy variables such as stream order or stream segment catchment area (Peterson, 2011), where catchment area is the area of land that drains directly to a particular stream segment. The integral in (3.1) is calculated piecewise by summing up all segments containing the moving average function $g(x|\boldsymbol{\theta})$. Only the segments that are flow-connected and in U_i, the set of segments upstream from monitoring site x_i including the i_{th} segment, need integrated. The covariance between two flow-connected monitoring sites r_i and s_j, where r_i is downstream of s_j can be calculated as

$$C(r_i, s_j|\boldsymbol{\theta}) = \begin{cases} \pi_{ij}C_t(h|\boldsymbol{\theta}) & \text{if } r_i \text{ and } s_j \text{ are flow-connected} \\ 0 & \text{if } r_i \text{ and } s_j \text{ are flow-unconnected} \end{cases} \tag{3.3}$$

where π_{ij} are weights and h is stream distance. $C_t(h|\boldsymbol{\theta})$ can take many forms, with the exponential, spherical (ver Hoef and Peterson, 2010) and Epanechnikov (Garreta et al., 2010) being explored in this chapter. Their form is given in Table 3.1. The factor 3 in the exponential model causes the autocorrelation to be approximately 0.05 when h equals the range parameter α_u. This helps compare range parameters between models which approach zero asymptotically (exponential) with those that reach zero at α_u (ver Hoef et al., 2014).

The moving average function splits when it encounters branching as it moves upstream. Weights are assigned to each stream segment and are based on additive functions that ensure stationarity of the variance (ver Hoef and Peterson, 2010). The additive function is constant within a stream segment and is the sum of the value of two segments when they join at a confluence point. Cressie et al. (2006) defines a spatial process on a stream network as

Model	
Exponential	$C_t(h\|\boldsymbol{\theta}) = \sigma_u^2 \exp(-3h/\alpha_u)$
Spherical	$C_t(h\|\boldsymbol{\theta}) = \theta_v \left(1 - \frac{3}{2}\frac{h}{\alpha_u} + \frac{1}{2}\frac{h^3}{\alpha_u^3}\right) I\left(\frac{h}{\alpha_u} \leq 1\right)$
Epanechnikov	$C_t(h\|\boldsymbol{\theta}) = \frac{\sigma_u^2 (h-\alpha_u)^2 f_{eu}(h;\alpha_u)}{16\alpha_u^5} I\left(\frac{h}{\alpha_u} \leq 1\right)$

TABLE 3.1: Tail-up models used in this chapter. $f_{eu}(h;\alpha_u) = 16\alpha_u^2 + 17\alpha_u^2 h - s\alpha_u h^2 - h^3$, $I(\cdot)$ is the indicator function (equal to 1 if true), σ_u^2 is the overall variance parameter also know as the partial sill, $\alpha_u > 0$ is the range parameter (in stream distance), and $\boldsymbol{\theta}_u = (\sigma_u^2, \alpha_u)^\top$.

$$Z(s_i|\boldsymbol{\theta}) = \int_{\vee s_i} g(x - s_i|\boldsymbol{\theta}) \sqrt{\frac{\Omega(x)}{\Omega(s_i)}} dW(x),$$

where $\vee_{s_i}$ is the domain upstream of the point s_i. In this case, π_{ij} in (3.3) are $\sqrt{\frac{\Omega(x)}{\Omega(s_i)}}$. Alternatively, ver Hoef et al. (2006) define a spatial process on a stream network as

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

$B_{ij} = D_j \backslash D_i$ is the set of segments between the i_{th} and j_{th} (including the j_{th} but excluding the i_{th}). Also, for each fork upstream of the i_{th} segment, where the branches are denoted j and k, $0 \leq \omega_j, \omega_k \leq 1$ and $\omega_j + \omega_k = 1$. This ensures stationarity of variances and π_{ij} in (3.3) will be $\prod_{k \in B_{i,j}} \sqrt{\omega_k}$. ver Hoef and Peterson (2010) show that this weighting scheme is equivalent to that in Cressie et al. (2006).

ver Hoef and Peterson (2010) introduce the Tail-down model which models autocorrelation between monitoring sites that are flow-unconnected (as in Figure 3.1). The moving average function is defined so that it is nonzero only downstream from a monitoring site. The form of the covariance function for the Tail-down model depends on whether sites are flow-connected (c) or flow-unconnected (n). If sites are flow-connected, then for s_2 upstream of r_1 and $h = s_2 - r_1 > 0$,

$$C_c(h|\boldsymbol{\theta}) = \int_{-\infty}^{-h} g(-x|\boldsymbol{\theta}) g(-x - h|\boldsymbol{\theta}) dx. \tag{3.4}$$

If r_1 and s_2 are flow-unconnected and the distance from r_1 to the nearest confluence with s_2 downstream of which they share flow is a and the distance from s_2 to the nearest confluence with r_1 is b, then for $b \geq a$

$$C_n(a, b|\boldsymbol{\theta}) = \int_{-\infty}^{-b} g(-x|\boldsymbol{\theta})g(-x-(b-a)|\boldsymbol{\theta})dx. \tag{3.5}$$

As with the Tail-up models, (3.4) and (3.5) can be reparameterized as follows for the spherical Tail-down model:

$$C_d(a, b, h|\boldsymbol{\theta}) = \begin{cases} \sigma_d^2 \left(1 - \frac{3}{2}\frac{h}{\alpha_d} + \frac{1}{2}\frac{h^3}{\alpha_d^3}\right) I\left(\frac{h}{\alpha_d} \leq 1\right) & \text{if flow-connected,} \\ \sigma_d^2 \left(1 - \frac{3}{2}\frac{a}{\alpha_d} + \frac{1}{2}\frac{b}{\alpha_d}\right)\left(1 - \frac{b}{\alpha_d}\right)^2 I\left(\frac{b}{\alpha_d} \leq 1\right) & \text{if flow-unconnected.} \end{cases} \tag{3.6}$$

The Tail-down model has not been applied in this chapter and details of other Tail-down covariance models can be found in ver Hoef and Peterson (2010).

Random effect models can also be fitted (ver Hoef et al., 2014) for factor variables where it is assumed that $\gamma_k(x)$ is the factor level at monitoring site x. Each level of γ_k is a random quantity with zero mean and variance σ_k^2. Sites with the same level of k are correlated and

$$C_k(r_i, s_j) = \begin{cases} \sigma_i^2 & \text{if } \gamma_k(r_i) = \gamma_k(s_j), \\ 0 & \text{if } \gamma_k(r_i) \neq \gamma_k(s_j). \end{cases} \tag{3.7}$$

The most general linear model that can be fitted using the `SSN` package in `R` is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}_u + \mathbf{z}_d + \mathbf{z}_e + \mathbf{W}_1\boldsymbol{\gamma}_1 + ... + \mathbf{W}_p\boldsymbol{\gamma}_p + \boldsymbol{\varepsilon}$$

where $\mathbf{X}$ is the design matrix of fixed effects with corresponding parameters $\boldsymbol{\beta}$. $\mathbf{z}_u$, $\mathbf{z}_d$ and $\mathbf{z}_e$ are vectors containing spatially autocorrelated random variables with Tail-up, Tail-down and Euclidean autocovariance, respectively, with $\text{var}(\mathbf{z}_u) = \sigma_u^2\mathbf{R}(\alpha_u)$, $\text{var}(\mathbf{z}_d) = \sigma_d^2\mathbf{R}(\alpha_d)$ and $\text{var}(\mathbf{z}_e) = \sigma_e^2\mathbf{R}(\alpha_e)$. $\mathbf{R}(\alpha_u)$, $\mathbf{R}(\alpha_d)$ and $\mathbf{R}(\alpha_e)$ are correlation matrices depending on range parameters α_u, α_d and α_e. $\mathbf{W}_k$ is a design matrix of random variables effects γ_k and $\text{var}(\gamma_k) = \sigma_k^2\mathbf{I}$; $k = 1, .., p$. $\boldsymbol{\varepsilon} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ Spatial prediction using

this model is 'universal kriging' and 'ordinary kriging' is the special case where there are no covariates and the design matrix $\mathbf{X}$ is a single column of 1's. The most general form of the covariance matrix is

$$\text{cov}(\mathbf{Y}) = \mathbf{\Sigma} = \sigma_u^2\mathbf{R}(\alpha_u) + \sigma_d^2\mathbf{R}(\alpha_d) + \sigma_e^2\mathbf{R}(\alpha_e) + \sigma_1^2\mathbf{W}_1\mathbf{W}_1^\top + ... + \sigma_p^2\mathbf{W}_p\mathbf{W}_p^\top + \sigma_0^2\mathbf{I}$$

where u, d and e refer to the Tail-up, Tail-down and Euclidean covariance models respectively, and σ_0^2 is the nugget effect representing independent error. ver Hoef et al. (2014) gives details of the spatial generalized linear mixed model where data follow either a binomial or poisson distribution.

3.2 Investigating spatial relationships

The data provided by the EA have a hierarchical structure with spatial correlation at different levels:

- Monitoring sites located within a single river network
- Multiple networks within a single LHA
- Spatially connected LHA's

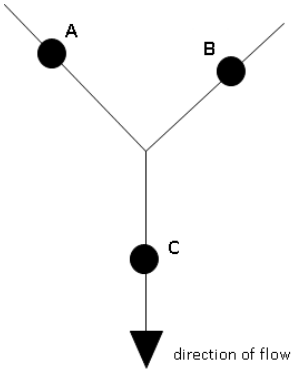
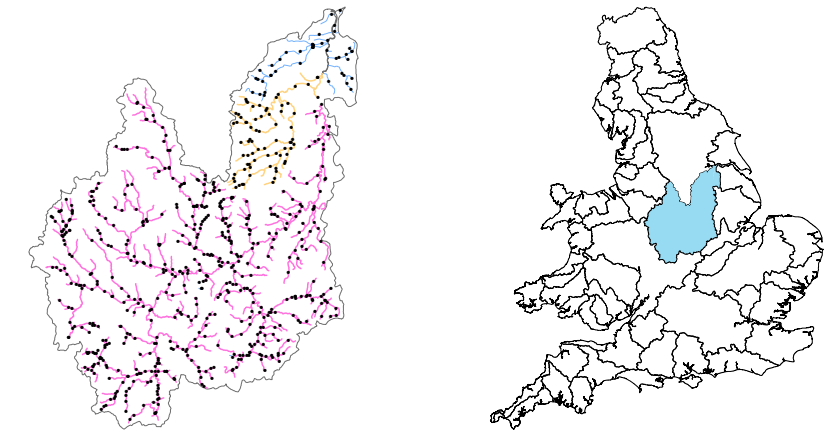


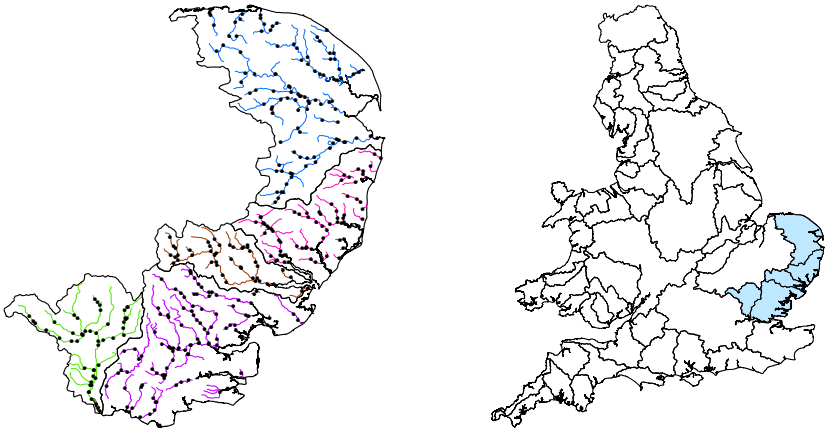
FIGURE 3.1: Illustration of the definition of 'flow connected' and 'flow unconnected'. The diagram contains monitoring sites (circles), stream segments (lines) and direction of flow (arrow). (A and C) and (B and C) are 'flow connected' whereas (A and B) are 'flow-unconnected'.

A 'network' is defined here as a set of flow connected (Figure 3.1) rivers with multiple sources that drain into a single outlet. LHA 28 (Trent) is used here to investigate the nature of spatial correlation in a single network and between multiple networks within a single LHA. LHA 28 is a nitrate vulnerable zone and has many flow-connected and flow-unconnected monitoring sites so was considered suitable for the investigations being carried out. LHA 28 consists of 16 separate networks and has a dominant network with 564 monitoring sites. 15 smaller networks contain between 1 and 73 monitoring sites, as shown in Figure 3.2(a). The position of LHA 28 relative to the rest of England and Wales is shown in Figure 3.2(b). LHA's 34 (Bure, Waveney), 35 (Gipping), 36 (Stour, East Anglia), 37 (Blackwater, Chelmer) and 38 (Lee) are used to investigate spatial correlation between multiple LHA's (Figures 3.2(c) and 3.2(d)).



(a) LHA 28. Dominant and second largest network indicated by pink and orange lines respectively.

(b) Location of LHA 28.



(c) LHA's 34-38.

(d) Location of LHA's 34-38.

FIGURE 3.2: LHA's used for investigation of spatial correlation in river networks at different spatial scales. Monitoring sites are black dots (a) and (c).

3.2.1 Spatial relationships in a single network

The dominant network of LHA 28 (Figure 3.2(a)) is used in this section to investigate spatial correlation in a single network. Figure 3.3 show log(TON) averaged by season for 1990-2010, 1990-2000 and 2003-2010, referred to as 'all', 'early' and 'late' respectively. The date ranges were chosen so a value existed for every monitoring site to make the plots directly comparable. There are small differences between the plots in the middle and right columns suggesting there has been a small decrease over time which is most noticeable for the winter months. The analyses in this section will be applied to the seasonal data averaged over all/early/late years to investigate (1) an appropriate spatial covariance structure to model these data and (2) if the parameters of the spatial covariance function change over time.

Four types of model were fitted to the data from the dominant network in LHA 28: non-spatial (3.8), covariance structure based on stream distance (Tail-up model) (3.9), covariance structure based on Euclidean distance (3.10) and hybrid models combining both Euclidean distance and stream distance covariance structures (3.11). For each type of covariance structure, log(TON) was modelled for $m = 1, ..., 566$ monitoring sites and coefficients β_1 and β_2 were estimated using REML. Other notation is as defined in Section 3.1. In each of these four models it is assumed that $\varepsilon_m \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

$$\log(\text{TON})_m = \beta_1 \text{Easting}_m + \beta_2 \text{Northing}_m + \varepsilon_m \tag{3.8}$$

$$\begin{aligned} \log(\text{TON})_m &= \beta_1 \text{Easting}_m + \beta_2 \text{Northing}_m + \text{z}_{u_m} + \varepsilon_m \\ \text{var}(\mathbf{z}_u) &= \sigma_u^2 \mathbf{R}(\alpha_u) \end{aligned} \tag{3.9}$$

$$\begin{aligned} \log(\text{TON})_m &= \beta_1 \text{Easting}_m + \beta_2 \text{Northing}_m + \text{z}_{e_m} + \varepsilon_m \\ \text{var}(\mathbf{z}_e) &= \sigma_e^2 \mathbf{R}(\alpha_e) \end{aligned} \tag{3.10}$$

$$\begin{aligned} \log(\text{TON})_m &= \beta_1 \text{Easting}_m + \beta_1 \text{Northing}_m + \text{z}_{u_m} + \text{z}_{e_m} + \varepsilon_m \\ \text{var}(\mathbf{z}_u) &= \sigma_u^2 \mathbf{R}(\alpha_u) \\ \text{var}(\mathbf{z}_e) &= \sigma_e^2 \mathbf{R}(\alpha_e) \end{aligned} \tag{3.11}$$

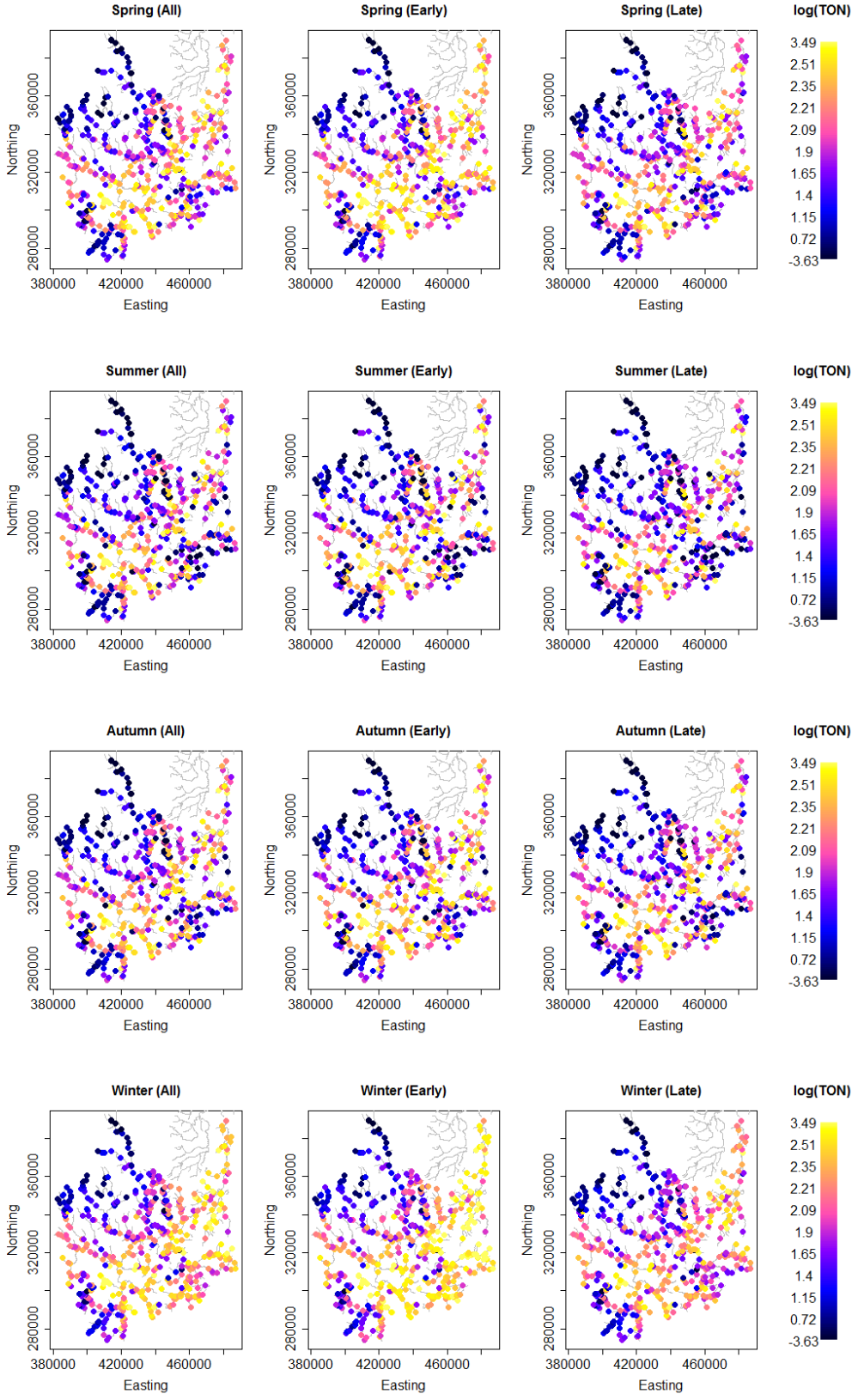


FIGURE 3.3: Log(TON) for Spring, Summer, Autumn and Winter averaged over 1990-2010 (left), 1990-2000 (middle) and 2000-2003 (right).

Table 3.2 lists all models fitted to the dominant network in LHA 28. The models were chosen from among those currently implemented in the `SSN` package in `R`. Ideally, the fit should not depend strongly on the choice of particular covariance model but rather on choosing the appropriate covariance structure. All models were fitted using REML (Peterson and ver Hoef, 2010) and include a nugget effect as a measure of independent error. Model performance was compared using Root Mean Square Prediction Error (RMSPE) (3.12), where n is the sample size and $\hat{z}_i$ is the prediction of the i_{th} datum from the model fitted to all of the data points except i, $\mathbf{z}_{-i}$.

$$\text{RMSPE} = \sqrt{\sum_{i=1}^{n} \left(\hat{z}_i - z_i\right)^2 / n} \tag{3.12}$$

Model number	Model type	Covariance model
1	non-spatial	NULL
2	stream distance	Exponential.tailup
3	stream distance	Spherical.tailup
4	stream distance	Epanechnikov.tailup
5	Euclidean distance	Exponential.Euclidean
6	Euclidean distance	Gaussian.Euclidean
7	Euclidean distance	Spherical.Euclidean
8	hybrid	Exponential.Euclidean + Exponential.tailup
9	hybrid	Exponential.Euclidean + Spherical.tailup
10	hybrid	Exponential.Euclidean + Epanechnikov.tailup
11	hybrid	Gaussian.Euclidean + Exponential.tailup
12	hybrid	Gaussian.Euclidean + Spherical.tailup
13	hybrid	Gaussian.Euclidean + Epanechnikov.tailup
14	hybrid	Spherical.Euclidean + Exponential.tailup
15	hybrid	Spherical.Euclidean + Spherical.tailup
16	hybrid	Spherical.Euclidean + Epanechnikov.tailup

TABLE 3.2: Models fitted to the dominant network in LHA 28. All models also include a nugget effect representing independent error.

Model selection here focuses on choosing a suitable spatial correlation structure rather than covariate selection since no covariates other than Easting and Northing were available. Easting and Northing are included as linear terms but recent developments in O'Donnell et al. (2014) mean that Easting and Northing could be modelled as non-linear terms or a smooth spatial surface, although their method incorporates flow connectedness information into the deterministic part of the model rather than the covariance

structure. Peterson and ver Hoef (2010) recommend fitting models with the same correlation structure but different covariate combinations using ML and using AIC as the criterion for best model. Garreta et al. (2010) state that in the geostatistical context where data are dependent, it is not appropriate to use BIC or corrected versions of AIC which use sample size for weighting as a model selection criteria. Once a suitable subset of covariates has been selected, the model should be refitted for a variety of covariance structures using REML (ver Hoef et al., 2014). The best model is selected with RMSPE as the criterion. Since the model will be used for kriging at unobserved locations, RMSPE is the criterion used for model selection, where lower values of RMSPE represent better models.

Figure 3.4 shows RMSPE plotted for each model, where data are averaged by season for all/early/late years. In almost all cases, the hybrid model with Epanechnikov Tail-up + Gaussian Euclidean covariance structure has the lowest RMSPE, or is within the best 3 of all models fitted, for each subset of the data. The hybrid models with Gaussian Euclidean covariance model tend to perform better than other hybrid models. It should be noted however that among the hybrid models, there is little difference between RMSPE and this is encouraging as it confirms that the particular choice of covariance model is not as important or influential as the choice of covariance structure i.e. whether to have a non-spatial, stream distance, Euclidean distance or hybrid model. It is clear from Figure 3.4 that the hybrid models perform better than the non-spatial, stream distance or Euclidean distance models.

The covariance parameters estimated for the hybrid models with Gaussian Euclidean structure were examined to choose the best hybrid model. Table 3.3 contains the estimated range parameters. In the dominant network in LHA 28 the greatest flow-connected stream distance is approximately 216km and the greatest Euclidean distance is approximately 140km. The Exponential Tail-up component of the hybrid model has an estimated effective range parameter greater than 100km and the estimated range parameter of the Spherical Tail-up component is often greater than 100km. The range for the Epanechnikov Tail-up component is 50-60km on average. The estimated effective range parameter for the Gaussian Euclidean component of the three models does not appear to vary much with the choice of Tail-up model. Table 3.4 shows the partial sill for each of the three hybrid models with a Gaussian Euclidean distance component.

The partial sill does not vary much between Tail-up models, although the partial sill is greater for summer and autumn than in spring and winter.

Data	Epa TU	Gau Euc	Sph TU	Gau Euc	Exp TU	Gau Euc
Spr.A	49.0	55.5	58.2	55.0	127.6	50.5
Spr.E	48.8	45.4	58.4	45.6	126.5	42.9
Spr.L	46.3	58.3	52.8	58.0	106.9	54.9
Sum.A	59.6	50.6	121.3	46.4	195.2	46.5
Sum.E	57.3	48.3	77.6	46.3	169.7	44.2
Sum.L	72.3	50.1	132.0	48.2	260.6	47.7
Aut.A	58.7	47.4	123.1	43.6	198.9	43.5
Aut.E	58.2	43.5	118.0	40.6	176.9	40.6
Aut.L	78.2	45.0	134.2	43.9	259.7	43.4
Win.A	54.8	50.9	119.7	48.1	180.5	47.9
Win.E	54.8	48.6	120.6	46.4	168.9	46.3
Win.L	52.7	53.1	80.3	51.7	165.9	50.0

TABLE 3.3: Range parameter for hybrid models with Gaussian Euclidean component. Values are stream distance given in km. Season.A = 1990-2010, season.E = 1990-2000 and season.L = 2003-2010. Epa = Epanechnikov, Sph = Spherical, Exp = Exponential, Gau=Gaussian, TU = Tail-up component, Euc = Euclidean component.

Data	Epanechnikov	Spherical	Exponential
Spring.A	0.15	0.14	0.15
Spring.E	0.14	0.14	0.14
Spring.L	0.15	0.15	0.15
Summer.A	0.28	0.28	0.28
Summer.E	0.31	0.31	0.31
Summer.L	0.25	0.25	0.25
Autumn.A	0.21	0.21	0.21
Autumn.E	0.22	0.22	0.22
Autumn.L	0.20	0.20	0.20
Winter.A	0.09	0.09	0.09
Winter.E	0.09	0.09	0.09
Winter.L	0.09	0.10	0.10

TABLE 3.4: Partial sill parameter for Tail-up component of hybrid models with Gaussian Euclidean component. Season.A=1990-2010, season.E=1990-2000 and season.L=2003-2010.

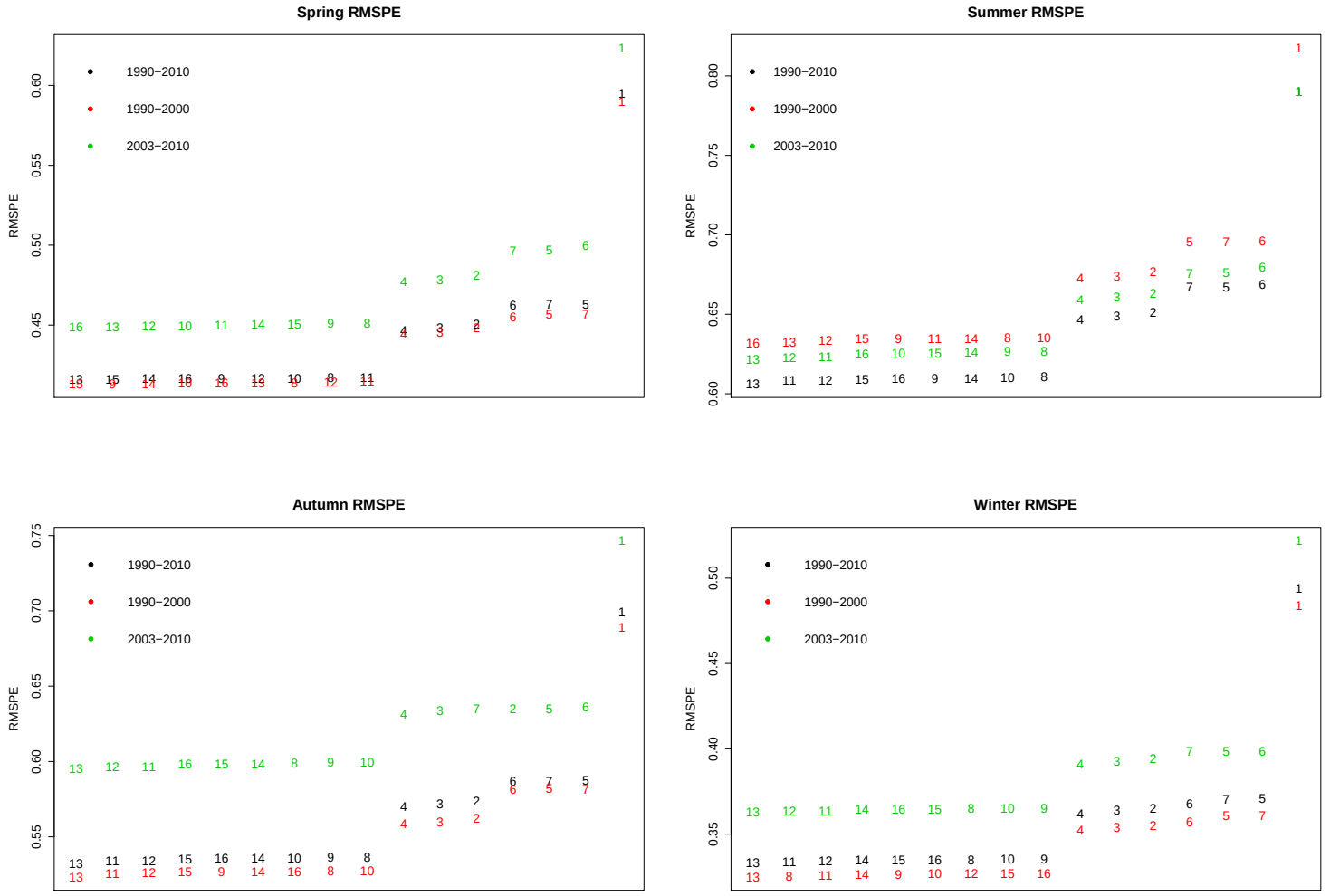


FIGURE 3.4: RMSPE for log(TON) averaged over the periods 1990-2010 (black), 1990-2000 (red) and 2003-2010 (green), by season for the dominant network in the Trent. The labels 1-16 correspond to the model numbers in Table 3.2.

As discussed in Chapter 1, the semivariogram can be used to estimate the parameters in the covariance model ((effective) range, sill, nugget) but this is based on Euclidean distance between pairs of monitoring sites. ver Hoef et al. (2006) introduce the Torgegram - an adaptation of the variogram and is based on stream distance. The semivariance is displayed separately for flow-connected and flow-unconnected sites. ver Hoef et al. (2006) state that the Torgegram should not be used to estimate covariance model parameters as the points are not weighted correctly for the flow connected sites. It is still a useful tool however to gain some insight into reasonable values for covariance parameters in the Tail-up model. Figure 3.5 contains Torgegrams calculated for spring, summer, autumn and winter where log(TON) is averaged over 1990-2010. From looking at the Torgegrams it seems that the Exponential Tail-up model puts the effective range around the centre of the last bin whereas the Epanechnikov and Spherical Tail-up models estimate the range to be around one third or one half of the binned distances, respectively. Either of these two Tail-up models would be appropriate but since the model with the Epanechnikov Tail-up component tended to give a smaller RMSPE than the Spherical Tail-up component, the hybrid model with the Epanechnikov Tail-up + Gaussian Euclidean models was selected as the best model.

Table 3.5 shows the estimated covariance parameters for the final model. The partial sill tends to be lower in the spring and winter compared to summer and autumn for the Tail-up component and the range is greater in the summer and autumn months for the Tail-up component. The range and partial sill are fairly constant across all seasons for the Euclidean component. The nugget is lowest in the winter and highest in the summer for all time averaged subsets of the data. It is interesting to note that the Euclidean distance component dominates the variance (Table 3.6) in Spring and Winter, accounting for approximately 50% of the variability in the model. This makes sense when thinking about the seasonal pattern of log(TON). Values are higher in Spring and Winter and Lord and Antony (2000) suggest that this is a characteristic related to autumn and winter rainfall resulting in the mobilisation of mineralised nitrogen from land.

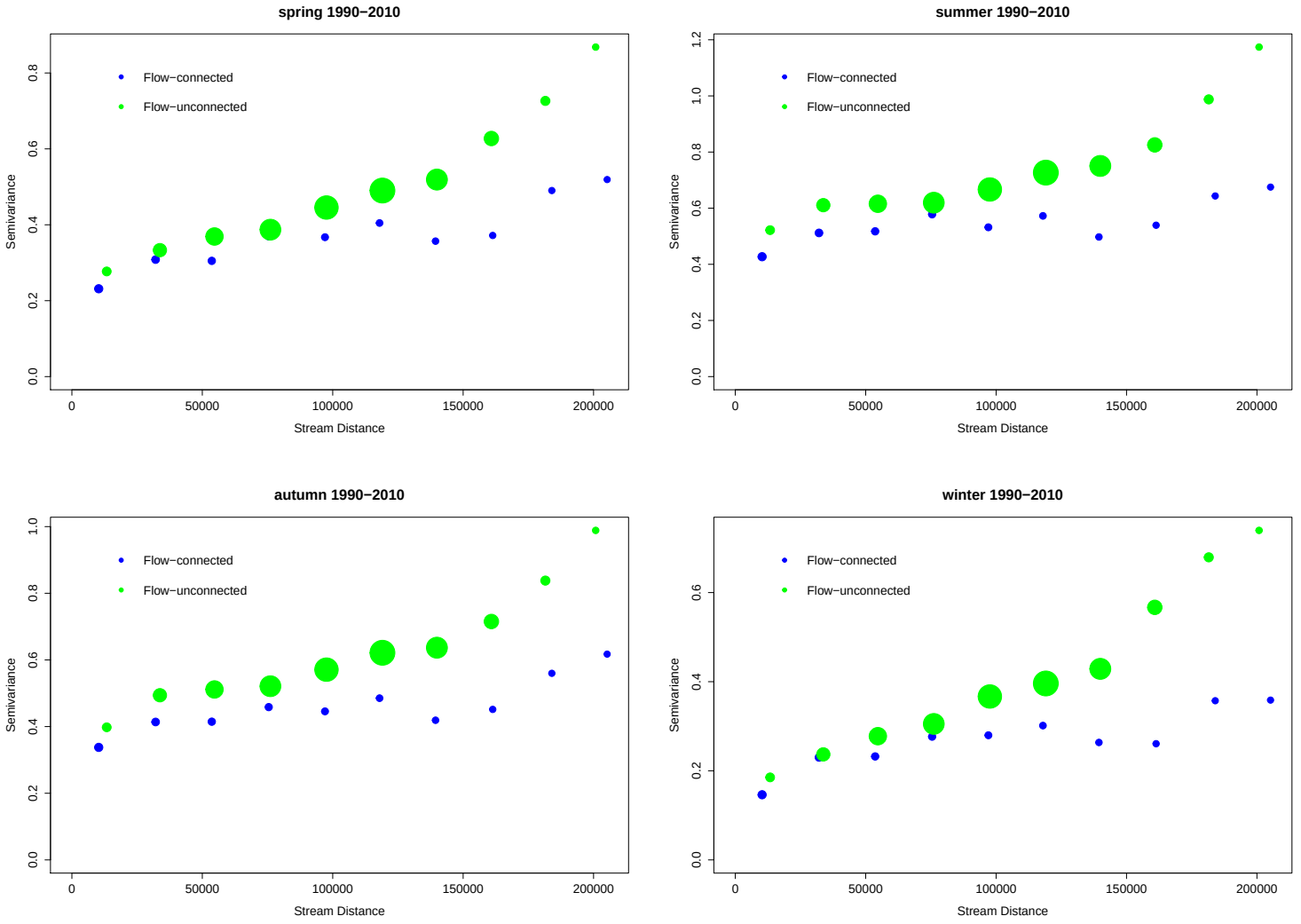


FIGURE 3.5: Torgegrams calculated for spring (a), summer (b), autumn (c) and winter (d) where log(TON) is averaged over 1990-2010. The semivariance is calculated for flow-connected (blue) and flow-unconnected (green) sites separately. Point sizes indicate relative number of pairs of monitoring sites in each bin.

Data	Epanech.TU partial sill	Epanech.TU range	Gaussian.Euc partial sill	Gaussian.Euc range	Nugget
Spring.A	0.15	49.0	0.29	55.5	0.12
Spring.E	0.14	48.8	0.22	45.4	0.12
Spring.L	0.15	46.3	0.33	58.3	0.14
Summer.A	0.28	60.0	0.39	50.6	0.25
Summer.E	0.31	57.3	0.35	48.3	0.27
Summer.L	0.25	72.3	0.39	50.1	0.27
Autumn.A	0.21	58.7	0.28	47.4	0.19
Autumn.E	0.22	58.2	0.23	43.5	0.18
Autumn.L	0.20	78.2	0.28	45.0	0.26
Winter.A	0.09	54.8	0.21	50.9	0.08
Winter.E	0.09	54.8	0.17	48.6	0.07
Winter.L	0.09	52.7	0.25	53.1	0.09

TABLE 3.5: Covariance model parameters for final model for dominant network in LHA 28. (effective) range is distance in km.

Data	Covariates (R-sq)	Epanech.TU	Gaussian.Euc	Nugget
Spring.A	0.01	0.26	0.52	0.21
Spring.E	0.02	0.29	0.45	0.24
Spring.L	0.01	0.24	0.53	0.22
Summer.A	0.01	0.30	0.42	0.27
Summer.E	0.01	0.33	0.38	0.28
Summer.L	0.01	0.27	0.42	0.29
Autumn.A	0.01	0.31	0.40	0.28
Autumn.E	0.02	0.34	0.36	0.28
Autumn.L	0.01	0.26	0.38	0.34
Winter.A	0.02	0.24	0.55	0.20
Winter.E	0.03	0.26	0.51	0.21
Winter.L	0.02	0.21	0.57	0.20

TABLE 3.6: Variance components of final model for dominant network in LHA 28.

The models described in Table 3.2 were also fitted to the data from the second largest network in LHA 28. This network contains 73 monitoring sites, is spread over a smaller geographical area than the dominant network and is shown by the orange line in Figure 3.2(a). These models were fitted to this network to investigate consistency of models for different networks within the same LHA. It seems from Figure 3.6 that this smaller network favours hybrid models with an Exponential Euclidean structure, regardless of the Tail-up structure but as with the dominant network, the hybrid model with Epanechnikov Tail-up component has lowest RMSPE in most cases. Tables 3.7 and 3.8 show the

range and partial sill parameters for the Tail-up component of the hybrid models with Exponential Euclidean component. There is little difference between the three models for the partial sill but the Exponential Tail-up model estimates the range parameter to be far greater than the Spherical or Epanechnikov models. The partial sill is higher in the summer and autumn months as was seen in the dominant network. The Euclidean component explains almost all of the variance for the smaller network (Table 3.9). This might be because the network is fairly homogeneous in terms of observed values, or because the network is spread over a smaller geographic area compared to the dominant network, or perhaps a mixture of both.

	Exponential	Spherical	Epanechnikov
SprA	222.4	61.0	49.2
SprE	112.6	55.0	50.5
SprL	178.2	46.5	53.9
SumA	96.0	55.8	51.8
SumE	80.2	48.7	42.8
SumL	329.5	88.7	53.4
AutA	89.1	57.0	53.3
AutE	112.2	57.8	55.5
AutL	105.5	53.9	47.9
WinA	117.7	55.3	56.3
WinE	110.0	55.8	52.7
WinL	115.0	140.6	39.6

TABLE 3.7: Range parameter (km) for Tail-up component of hybrid models with Exponential Euclidean component, fitted to the second largest network in the Trent catchment area.

	Exponential	Spherical	Epanechnikov
SprA	0.03	0.04	0.03
SprE	0.04	0.04	0.04
SprL	0.01	0.00	0.02
SumA	0.12	0.12	0.13
SumE	0.22	0.22	0.22
SumL	0.01	0.03	0.01
AutA	0.11	0.12	0.12
AutE	0.14	0.14	0.14
AutL	0.08	0.08	0.08
WinA	0.02	0.02	0.02
WinE	0.02	0.02	0.03
WinL	0.00	0.00	0.00

TABLE 3.8: Partial sill for Tail-up component of hybrid models with Exponential Euclidean component, fitted to netID6.

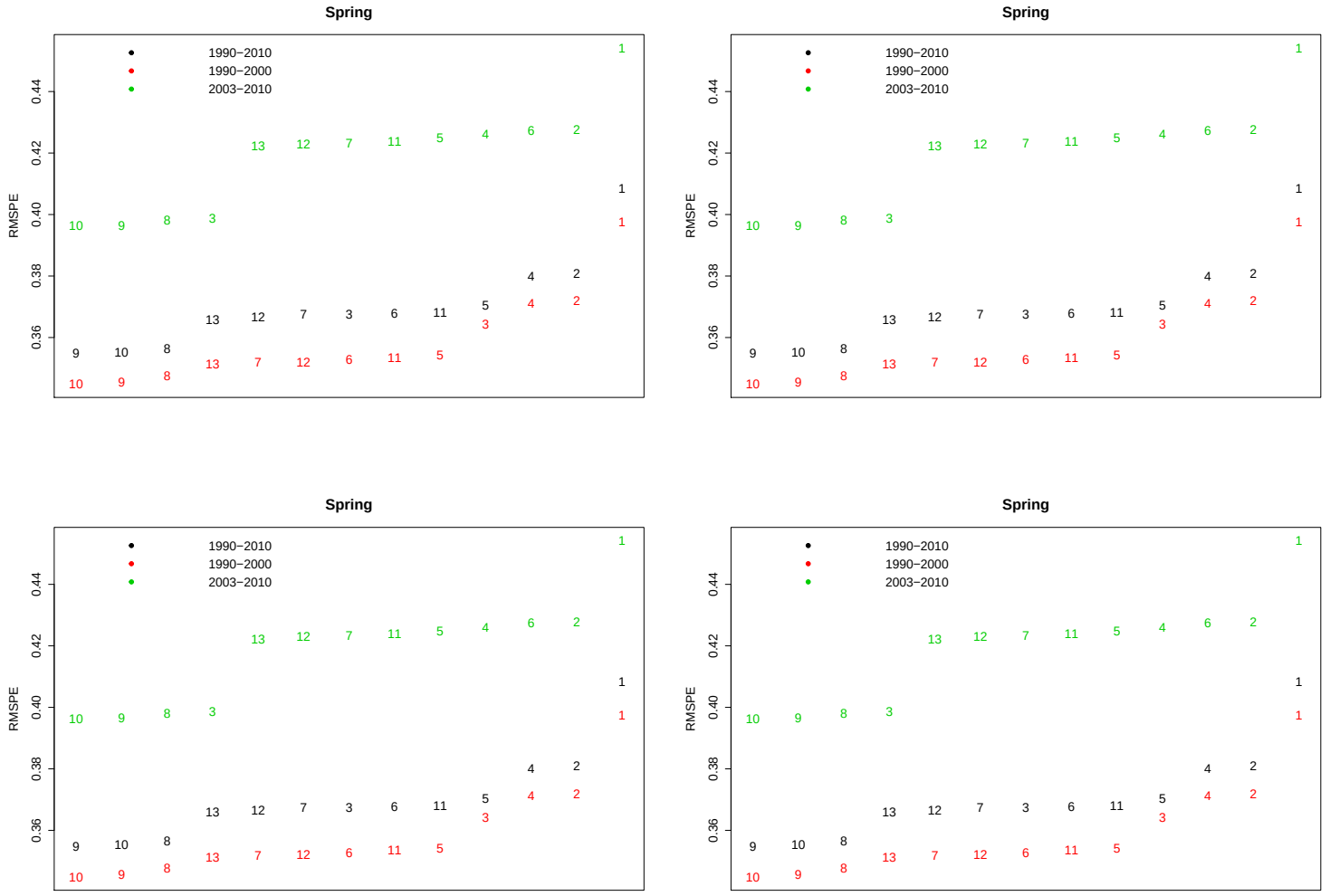


FIGURE 3.6: RMSPE for non-spatial, Euclidean distance and hybrid covariance structures for the second largest network in LHA 28. The plotted numbers correspond to the model numbers in Table 3.2

Data	Covariates (R-sq)	Epanech.tailup	Exponential.Euclid	Nugget
SprA	0.02	0.09	0.65	0.24
SprE	0.03	0.13	0.59	0.25
SprL	0.02	0.03	0.76	0.19
SumA	0.02	0.10	0.67	0.21
SumE	0.05	0.35	0.31	0.28
SumL	0.02	0.01	0.70	0.26
AutA	0.01	0.09	0.76	0.14
AutE	0.01	0.08	0.81	0.09
AutL	0.01	0.08	0.66	0.25
WinA	0.01	0.08	0.70	0.21
WinE	0.01	0.09	0.74	0.16
WinL	0.00	0.00	0.82	0.18

TABLE 3.9: Variance components for the Epanechnikov Tail-up + Exponential Euclidean covariance structure, fitted to the second largest network in LHA 28.

This section has shown that a hybrid structure covariance model with an Epanechnikov Tail-up component is suitable for the dominant and second largest networks in the Trent area although it could be argued that there is some evidence that the particular choice of Euclidean component differs between the two networks. It has been discussed that the choice of covariance structure is more important than the particular choice of Tail-up or Euclidean covariance function.

3.2.2 Spatial relationships among several networks in a single LHA

As well as modelling spatial correlation within a single river network, it is also of interest to investigate spatial correlation among multiple networks in a single LHA. Models were fitted to the data for all monitoring sites in LHA 28 and log(TON) was averaged by season for 1990-2010, 1990-2000 and 2003-2010. Three types of model were fitted as shown in Table 3.10. The first model has Easting and Northing as fixed effects and a hybrid covariance structure with an Epanechnikov Tail-up component and a Gaussian Euclidean component - this is the model that was chosen in the previous section as best representing spatial correlation in the networks being investigated and is used here as a baseline for comparison. The second model has network ID (netID) as a fixed effect, allowing a different mean to be estimated for each network within LHA 28. The third model includes netID as a random effect in the covariance structure which allows a different variance to be associated with each network.

Model	Fixed effects	Covariance		
		Tail-up	Euclidean	Random
Baseline	Easting + Northing	Epanechnikov	Gaussian	
Fixed	Easting + Northing + netID	Epanechnikov	Gaussian	
Random	Easting + Northing	Epanechnikov	Gaussian	netID

TABLE 3.10: Models fitted to investigate spatial correlation among multiple networks in a single LHA.

It was not possible to fit a fixed effects model to all of the data in LHA28 with netID as a fixed effect since this resulted in a singular design matrix that could not be inverted, a necessary step in the model estimation procedure. This was due to spatial sparsity in that eight of the networks in LHA28 contain only one monitoring site. Instead the models in Table 3.10 were fitted to two networks in LHA 28: the dominant network (netID7) and the smaller network (netID6), modelled separately in the previous section.

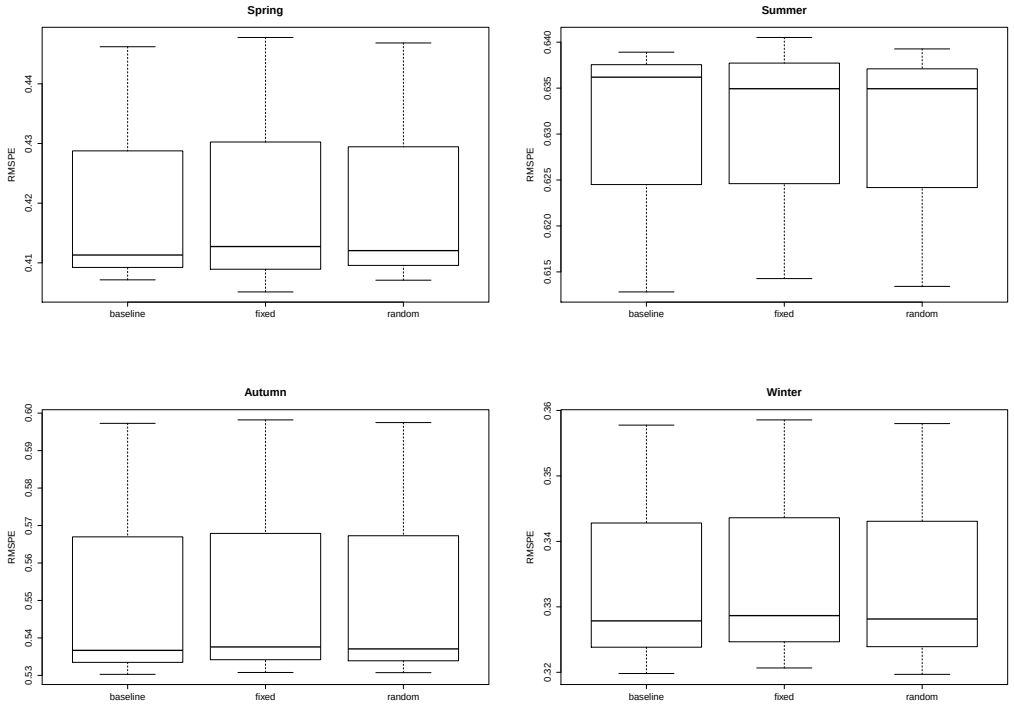


FIGURE 3.7: Boxplots of RMSPE for models fitted to the dominant and smaller network in LHA 28. Each box represents RMSPE for log(TON) averaged over 1990-2010, 1990-2000, 2003-2010.

By considering RMSPE (Figure 3.7) it appears that it is not necessary to account for different networks in the same LHA in the model since RMSPE seems fairly constant

across all models. Adding additional complexity to the fixed effects or covariance structure has not improved the predictive power of the model. This might be because LHA's are defined as a group of networks drained into by the same land. The previous section showed that the covariance structure is dominated by the Euclidean distance component which is likely affected by land use and is investigated further in Chapter 4. Modelling netID as a fixed effect means that an allowance is made for different mean values in different networks but it might be that the difference in mean log(TON) in different networks is too small to be significant in LHA 28 since this LHA is dominated by a large network. Modelling netID as a random effect allows for the estimation of a different variance on each network and again, differences might be masked by the dominance of the main network in LHA 28.

This section has shown that a hybrid covariance structure is the best way to model data collected on rivers in the Trent catchment area. The Tail-up covariance function contributed much less to the variance components for the second largest network in the Trent catchment area compared to the dominant network but an explanation for this is that the second largest network is much smaller than the dominant network and there is less evidence of a branched structure compared to the dominant network. It would be interesting in future to assess the consistency of optimal choice of covariance structure and functions for two networks of similar size and branching structure within the same catchment area to investigate whether it is necessary to account for different networks either in the deterministic or error structure of the model.

They key finding here is that although the particular choice of Tail-up covariance function and Euclidean covariance function differed between networks in the same catchment area, the choice of covariance structure (in this case the hybrid structure) was consistent.

3.2.3 Spatial relationships between multiple LHA's

When modelling log(TON) for the whole of England and Wales it might be necessary to account for different levels of variability of log(TON) for each LHA. LHA's 34, 35, 36, 37 and 38 are modelled here to investigate this third level of spatial correlation in river networks. These five LHA's have 367 monitoring sites in total and are shown in Figures 3.2(c) and 3.2(d). A random effect term was used to represent LHA number

in the spatial stream network model since by definition LHA's are independent and the random effects term assumes that sites with different levels of factor γ_k are independent (ver Hoef et al., 2014). Sites with the same level of factor γ_k are assumed to be correlated. RMSPE for models fitted with and without a random effect for LHA number are given in Table 3.11.

Data	Covariance Structure	RMSPE
Spr.A	Epanech.TU + Gaussian.Euc + Nugget	0.358
Spr.E	Epanech.TU + Gaussian.Euc + Nugget	0.401
Spr.L	Epanech.TU + Gaussian.Euc + Nugget	0.352
Sum.A	Epanech.TU + Gaussian.Euc + Nugget	0.690
Sum.E	Epanech.TU + Gaussian.Euc + Nugget	0.748
Sum.L	Epanech.TU + Gaussian.Euc + Nugget	0.667
Aut.A	Epanech.TU + Gaussian.Euc + Nugget	0.492
Aut.E	Epanech.TU + Gaussian.Euc + Nugget	0.511
Aut.L	Epanech.TU + Gaussian.Euc + Nugget	0.493
Win.A	Epanech.TU + Gaussian.Euc + Nugget	0.236
Win.E	Epanech.TU + Gaussian.Euc + Nugget	0.238
Win.L	Epanech.TU + Gaussian.Euc + Nugget	0.239
Spr.A	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.360
Spr.E	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.402
Spr.L	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.352
Sum.A	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.690
Sum.E	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.748
Sum.L	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.667
Aut.A	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.492
Aut.E	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.511
Aut.L	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.522
Win.A	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.236
Win.E	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.238
Win.L	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.240

TABLE 3.11: RMSPE for two spatial covariance structures fitted to seasonally averaged data. Spr=Sring, Sum=Summer, Aut=Autumn, Win=Winter. Season.A=seasonal average 1990-2010, Season.E=seasonal average 1990-2000, Season.L=seasonal average 2003-2010. TU=Tail-up, Euc=Euclidean. Models were fitted to LHA's 34, 35, 36, 37 and 38.

It does not seem to make any difference including the LHA as a random effect in terms of RMSPE since RMSPE is the same to 3 decimal places for almost all of the datasets (Table 3.11). This might be because there is no strong contrast of within LHA variability to between LHA variability. Lord and Antony (2000) show that land use and nitrate levels in these five LHA's are very similar so it is reasonable to conclude that the variability in log(TON) is similar for these areas. This analysis was repeated a second time to include the data for LHA 28. LHA 28 is classed by the Environment Agency as a

nitrate vulnerable zone while the other five LHA's are not. LHA's 34, 35, 36, 37 and 38 have similar land use but LHA 28 has a mixture of arable and non-arable land use (Lord and Antony (2000) and Chapter 4). Table 3.12 shows RMSPE for Spring, Summer, Autumn, and Winter log(TON) averaged over 1990-2010, 1990-2000 and 2003-2010 where models were fitted to the data from six LHA's. RMSPE was similar between models with and without the random effects term accounting for different levels of variability for each LHA. This is likely to be for similar reasons as discussed for LHA's 34-38 where variability of log(TON) was shown to be similar for all LHA's considered.

Data	Covariance Structure	RMSPE
Spr.A	Epanech.TU + Gaussian.Euc + Nugget	0.399
Spr.E	Epanech.TU + Gaussian.Euc + Nugget	0.411
Spr.L	Epanech.TU + Gaussian.Euc + Nugget	0.420
Sum.A	Epanech.TU + Gaussian.Euc + Nugget	0.651
Sum.E	Epanech.TU + Gaussian.Euc + Nugget	0.687
Sum.L	Epanech.TU + Gaussian.Euc + Nugget	0.666
Aut.A	Epanech.TU + Gaussian.Euc + Nugget	0.527
Aut.E	Epanech.TU + Gaussian.Euc + Nugget	0.534
Aut.L	Epanech.TU + Gaussian.Euc + Nugget	0.571
Win.A	Epanech.TU + Gaussian.Euc + Nugget	0.297
Win.E	Epanech.TU + Gaussian.Euc + Nugget	0.294
Win.L	Epanech.TU + Gaussian.Euc + Nugget	0.320
Spr.A	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.399
Spr.E	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.411
Spr.L	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.420
Sum.A	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.651
Sum.E	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.687
Sum.L	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.665
Aut.A	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.526
Aut.E	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.534
Aut.L	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.573
Win.A	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.297
Win.E	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.294
Win.L	Epanech.TU + Gaussian.Euc + NUMBER + Nugget	0.320

TABLE 3.12: RMSPE for two spatial covariance structures fitted to seasonally averaged data. Spr=Sring, Sum=Summer, Aut=Autumn, Win=Winter. Season.A=seasonal average 1990-2010, Season.E=seasonal average 1990-2000, Season.L=seasonal average 2003-2010. TU=Tail-up, Euc=Euclidean. Models were fitted to LHA's 28, 34, 35, 36, 37 and 38.

Table 3.13 shows the variance components for the models in Table 3.12. It seems that once stream distance and Euclidean distance have been accounted for, there is almost no

between LHA variability to be accounted for. The variability within LHA's dominates the covariance structure.

Data	Covariates (Rsq)	Tail-up	Euclidean	Number	Nugget
Spr.A	0.0010	0.262	0.538		0.200
Spr.E	0.0010	0.253	0.538		0.209
Spr.L	0.0020	0.269	0.507		0.223
Sum.A	0.0010	0.318	0.378		0.303
Sum.E	0.0020	0.314	0.368		0.317
Sum.L	0.0010	0.311	0.323		0.364
Win.A	0.0030	0.213	0.628		0.156
Win.E	0.0020	0.210	0.641		0.147
Win.L	0.0040	0.202	0.606		0.187
Spr.A	0.0010	0.262	0.537	0.00010	0.199
Spr.E	0.0005	0.253	0.538	0.00002	0.209
Spr.L	0.0020	0.269	0.506	0.00030	0.223
Sum.A	0.0010	0.318	0.378	0.00004	0.303
Sum.E	0.0020	0.313	0.367	0.00100	0.316
Sum.L	0.0010	0.315	0.331	0.00003	0.352
Win.A	0.0030	0.213	0.628	0.00002	0.156
Win.E	0.0020	0.210	0.641	0.00030	0.147
Win.L	0.0040	0.202	0.606	0.00030	0.187

TABLE 3.13: Variance components for the models listed in Table 3.12

They key message here is that for the LHA's considered in this section, it was not necessary to account for different levels of variance in each of the LHA's. Variance in this selection of catchment areas in England between 1990 and 2010 is dominated by variance due to Euclidean distance between monitoring sites. The rest of the variance is explained by variance due to separation along the river network, and residual variance (variance when distance between sites is zero).

3.3 Predicting at unsampled locations

One of the aims of geostatistics is to make predictions at unobserved locations. The statistical models in (3.8), (3.9), (3.10) and (3.11) can be used to predict log(TON) at unsampled locations and produce a map of log(TON) for the whole network. To provide an example of this, estimates of winL were calculated for approximately 11,000 unobserved locations using kriging, described in Chapter 1 and can be found in Figure 3.8

along with corresponding standard errors in Figure 3.9. winL is of particular interest since nitrates are higher in winter months compared to other months so this is when log(TON) is most likely to breach the limits set by the Water Framework Directive and Nitrates Directive. The Euclidean component is based on a Gaussian covariance function and the Tail-up component is the Epanechnikov function, both used earlier in this chapter. The maps show that there are many differences in the spatial pattern of winL estimated from a non-spatial model or Euclidean distance covariance model compared to predicted values based on a Tail-up covariance model. The Tail-up model allows for abrupt changes in value of winL at confluences which is not seen in the non-spatial or Euclidean models. The hybrid model in the bottom right map in Figure 3.8 shows predicted values from the hybrid model combining both stream and Euclidean distance based covariance functions but the differences between this map and the predictions based on the Tail-up model are small. More noticeable differences between the Tail-up and hybrid model can be found in the standard errors of the predictions as shown in Figure 3.9 where it can be clearly seen that the hybrid model produces predicted values with lower standard errors than the other covariance structures considered in this chapter.

The key message here is that predictions at unsampled locations on river networks with the least uncertainty can be estimated using a hybrid covariance function combining Euclidean and stream distances. The predicted values show abrupt changes as confluences which is to be expected and predictions with this feature can be found when based on models using either stream distance only or stream distance combined with Euclidean distance. The uncertainty of the predictions is however lower for predictions based on a combination of stream and Euclidean distances since relationships based on Euclidean distance provide useful information for stream segments with few or no monitoring sites. This property is very useful when the variable being predicted is related to both the structure of the river network and the land in which the river network is embedded, as is the case for log(TON).

3.4 Comments

The models considered in this chapter were fitted using the `SSN` package in `R`. One disadvantage of this approach is that extensive data processing is required before models

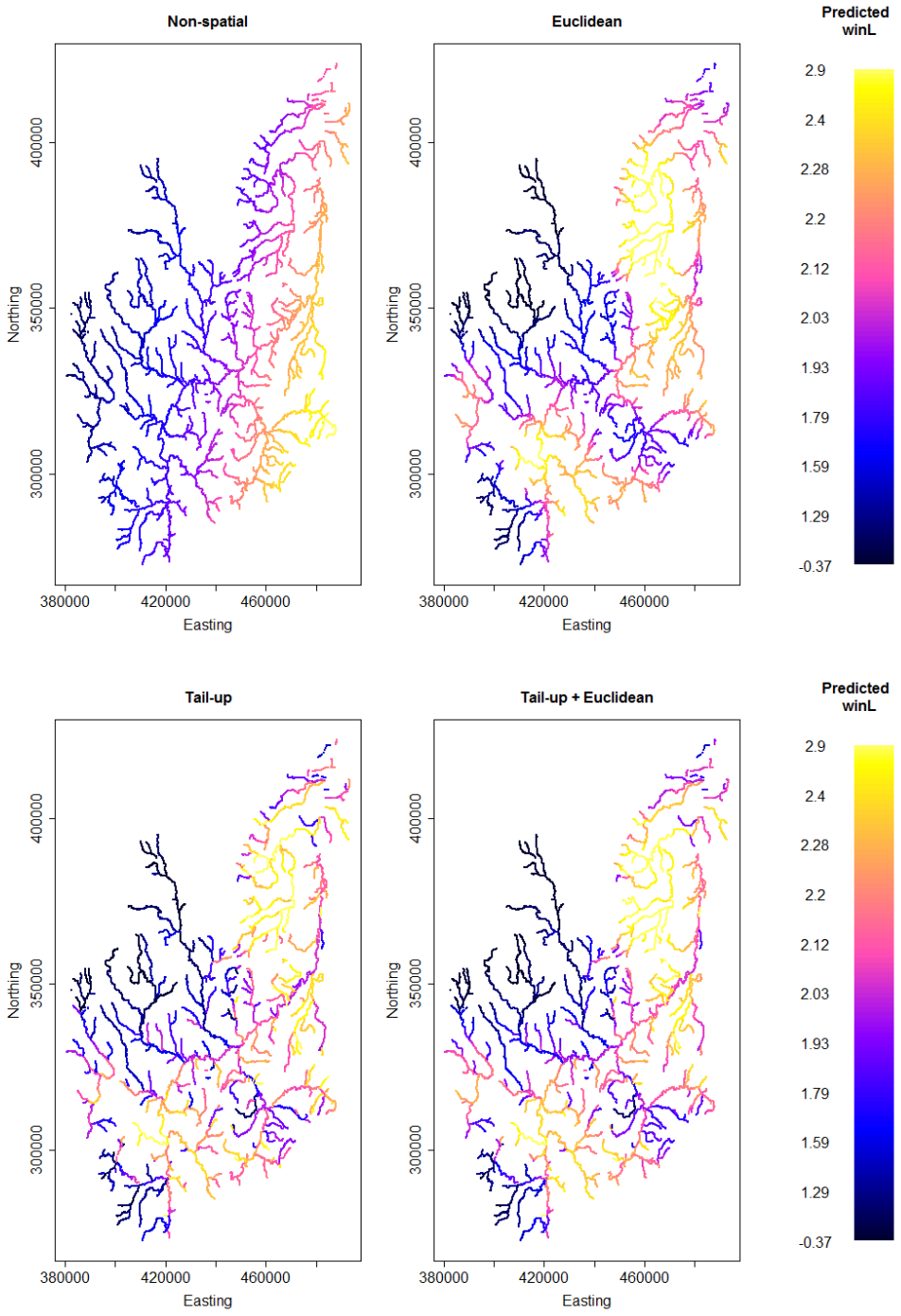


FIGURE 3.8: Predicted values of winL for non-spatial, Euclidean, Tail-up and hybrid covariance structures.

can be fitted and requires access to a commercial software license although Peterson and Ver Hoef (2014) explain that all of the necessary information is provided to enable the user to construct a `.ssn` object without pre-processing the data using the STARS toolkit (Peterson and Ver Hoef, 2014) in ArcGIS. A further drawback is that at present the SSN models can only accommodate linear relationships of covariates with the response

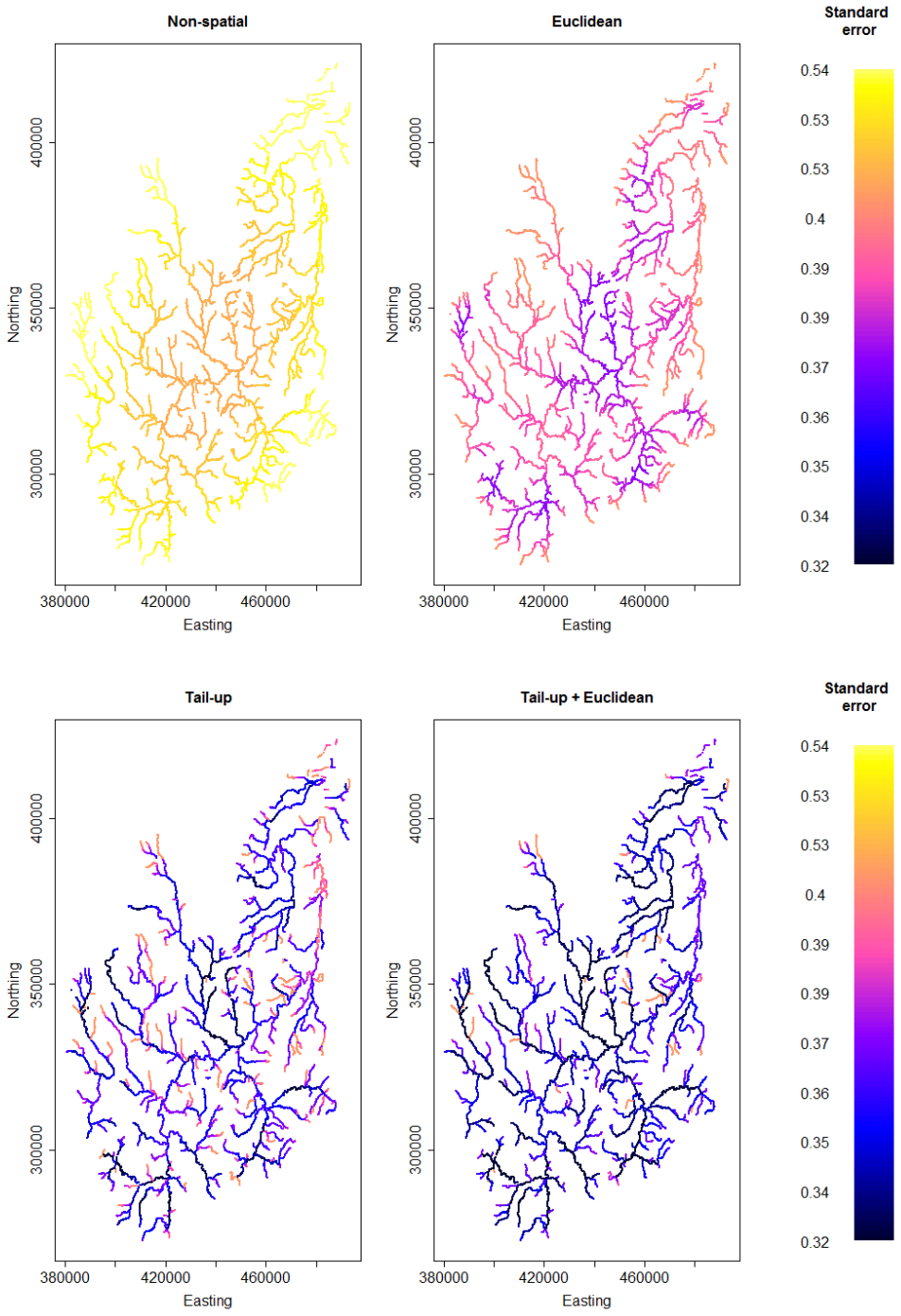


FIGURE 3.9: Standard errors for predicted values of winL for non-spatial, Euclidean, Tail-up and hybrid covariance structures.

variable and models must be fitted to a single time point or time average (such as seasonal average over several years). Recent developments in O'Donnell et al. (2014) and Rushworth et al. (2015) mean that in the future spatiotemporal models with flexible relationships between covariates and the response variable might also be considered, although their work uses the river network structure in the deterministic part of the

model rather than the error structure as in ver Hoef et al. (2006). A drawback to this approach is that computational time increases as the number of stream segments increases. Using simulated data, Rushworth et al. (2015) show that the methods in ver Hoef et al. (2006) have greater computational time than the penalized spline method of O'Donnell et al. (2014) for all sizes of river network considered, there is evidence that the methods in ver Hoef et al. (2006) would have lower computation time for river networks with large numbers of stream segments. The penalized spline approach also requires that the river network is represented as a collection of stream units which might not be easily adapted to river networks not governed by the Scottish Environment Protection Agency.

This chapter has investigated the most appropriate covariance structure to use when modelling data recorded on river networks. Models were fitted for a variety of time averaged subsets of the data which allowed some discussion on the consistency of parameters over time. A variety of covariance structures were considered to account for spatial correlation at different scales: (1) a single network, (2) multiple networks within an LHA and (3) multiple networks across multiple LHA's. It was concluded that a hybrid covariance structure was the best in terms of lowest RMSPE and inspection of the variance components showed that the covariance structure is dominated by the Euclidean distance component. By considering the RMSPE values, it was shown that modelling spatial correlation in stream networks using a standard Euclidean distance model is an improvement on the model assuming independent errors. The use of the Tail-up and hybrid covariance models reduces RMSPE further, justifying the use of a more complex model. It was shown that there is little difference between RMSPE for each of the hybrid models fitted meaning that the ability of the model to predict at unobserved locations is not greatly affected by the particular choice of covariance function once the covariance structure has been chosen.

It was also shown that for the LHA's considered in this chapter it was not necessary to account for different networks or LHA's using either a fixed or random effects structure and it was proposed that one reason for this was due to similar land use among the LHA's modelled. The effect of land use and other covariates on the covariance structure are discussed in further detail in Chapter 4. Finally, kriging was performed for the Trent area (LHA 28) and some differences between predictions based on a variety of covariance structures was discussed. Maps of kriging predictions were used to demonstrate

the improvement in prediction error as a result of fitting a model with a complicated covariance structure based on river network topology. By viewing maps produced at several time points in chronological order, it is possible to understand how winL changes across the whole network over time rather than just estimating an average trend for the whole LHA as in Chapter 2. This detailed view of the long term trend of log(TON) enables the user to identify particular areas of the river network requiring attention. Maps of predicted values can be produced for any time point or month/season/year for which there are sufficient data in space to estimate the model. Kriging depends on the covariance function parameter estimates and the effect of sample size on the parameter estimates is also considered in greater detail in Chapter 4.

The best fitting spatial statistical model with lowest RMSPE considered in this chapter has a complex hybrid covariance structure and does not include any covariate information. This could be used to reduce the cost of maintaining a monitoring network in two ways: first, it might be possible to reduce the number of monitoring sites located on a densely monitored river network if the spatial models with hybrid covariance structure can be shown to perform equally well with fewer monitoring sites. Secondly, it might be possible for regulatory agencies to reduce costs by reducing the need for covariate information which increases the costs of monitoring. A novel comparison between predictions made form covariance based models and covariate based models is presented in Chapter 4 to investigate whether this might be possible.

The key result from the investigations described in this chapter is that the most important thing to consider when choosing a model to be used for spatial prediction at unsampled locations on a river network is the choice of covariance structure i.e. Euclidean distance, stream distance, or hybrid. Choosing an appropriate covariance structure results in lower uncertainty for predicted values, particularly on stream segments with no monitoring sites. The particular Euclidean or stream distance covariance function (e.g. Gaussian, exponential, Epanechnikov) used is less crucial.

Chapter 4

Further investigation of the spatial covariance function

This chapter aims to further investigate the hybrid spatial covariance function with Epanechnikov Tail-up and Gaussian Euclidean components, identified as most suitable for the Trent catchment area in Chapter 3, and to provide suggestions to reduce the monitoring budget. Specifically, Section 4.1 aims to investigate how reducing the number of monitoring sites in the Trent affects the covariance function parameter estimates and predictions at unobserved locations made using the hybrid spatial covariance function. Following this, Section 4.2 considers the trade-off between modelling the mean structure using covariate information and modelling the covariance structure using a suitable function. The aim of this section is to investigate whether it is necessary to model both covariates and covariance or whether the cost of gathering data can be reduced by focusing on one of these.

The application of dynamic factor analysis in Chapter 2 suggested common temporal patterns of log(TON) are present and so it might be possible to reduce the size of the monitoring networks in England and Wales since there are monitoring sites recording the same information. A simulation study is designed and implemented in this chapter to explore how well a reduced monitoring network would perform in terms of predictions made using a spatial statistical model based on a subset of the data, and corresponding uncertainty of those predictions. Chapter 3 aimed to find the most appropriate spatial

covariance structure for modelling data recorded on river networks and found that predictions with the lowest uncertainty were made using a statistical model with a hybrid covariance structure, a combination of spatial covariance functions based on Euclidean and stream distance. This chapter also presents a novel study designed to investigate how predictions and their associated uncertainty made using a spatial statistical model with a complex hybrid covariance structure compare to predictions made using a statistical model built using covariate information. Recording, storing and processing covariate data adds to the cost of maintaining a monitoring network and producing the data required for the hybrid covariance structure is time intensive so there is interest in assessing whether the complex spatial covariance structure and (possibly expensive) covariate information are interchangeable. The two studies presented in this chapter aim to assess how cost savings could be made either by reducing the size of the monitoring network or substituting covariate information with a complex spatial covariance structure.

4.1 Sampling on a river network

This section will investigate the effect of reducing the number of sites in the monitoring network on the spatial covariance function parameter estimates and predictions from models based on a reduced monitoring network. First, a discussion of the literature is provided to explore existing approaches to reducing the size of monitoring networks, with particular emphasis on water quality monitoring networks. Next, a simulation study is implemented to investigate reducing the size of the monitoring network in the Trent catchment area. Finally, the results are discussed and general points to consider when reducing the size of a monitoring network are suggested. In particular, recommendations for reducing the monitoring network in the Trent catchment area are provided.

Government agencies frequently come under pressure to reduce costs and one way to do this is to reduce the size of monitoring networks. Chapter 2 showed that common patterns exist among large hydrological areas suggesting that duplicate information is being recorded by monitoring sites and strengthening the argument to reduce the number of monitoring sites. Ferreyra et al. (2002) state that "The density reduction of an existing spatial network is...relevant in many regions of the world where funding for environmental monitoring is decreasing". Diggle and Ribeiro (2007) refer to this reduction of an existing monitoring network as 'retrospective design'. If monitoring networks are

to be reduced then it is important to understand what effect this will have on inferences from the reduced data. Fuentes et al. (2007) state that "The proposed reduced network should maintain sufficient spatial information to ensure reasonable statistical inference" in the context of air pollution monitoring networks.

Ferreyra et al. (2002) discuss reducing the size of a soil water monitoring network and aim to find the optimal subset of monitoring points that will best describe the spatial pattern as well as accounting for variability over time. Wu et al. (2010) investigate reducing the size of an ozone monitoring network in France using a simulated annealing algorithm to select the optimal subnetwork of a given size and note that in the optimal subnetworks there are clusters of points in areas where observed ozone concentrations are spatially heterogeneous. Diggle and Ribeiro (2007) discuss general principles for the design of monitoring networks and conclude that a sampling design exhibiting some spatial regularity with some clustering will balance the objectives of parameter estimation and prediction. It is also noted that optimising with respect to a Bayesian predictive distribution will account for uncertainty in parameter estimates.

Dobbie et al. (2008) provide a comprehensive overview of the literature regarding spatial design with particular emphasis on stream networks. Strobl and Robillard (2008) discuss general principles for designing water quality sampling schemes and Khalil and Ouarda (2009) review statistical approaches used in the design of water quality monitoring networks. Herlihy et al. (2000) use a spatially balanced, randomized approach to select monitoring site locations. Much of the work on sampling in stream networks focuses on finding the optimal design of a given size by minimizing some criterion, or maximising a utility function if a Bayesian approach is used. For example, Dixon et al. (1999) use simulated annealing to find optimal locations for monitoring sites on a river network by minimising a cost function based on the expected cost of obtaining information after a problem is detected and Fuentes et al. (2007) aim to find the subnetwork of a given size that maximises posterior predictive entropy, accomplished by retaining monitoring sites with high predictive uncertainty and eliminating sites with small uncertainty. Kao et al. (2008) build on this by including pollution loads in the optimisation procedure so that areas with higher probability of a pollution event are more likely to be sampled. Monitoring networks of several sizes are investigated. Chilundo et al. (2008) considers reducing the size of a monitoring network on the Limpopo River.

More recently, work has been carried out to design and/or reduce monitoring networks in streams, taking into account the branching structure and flow direction that characterise stream networks. Som et al. (2014) discuss optimal sampling schemes for river networks under a variety of criteria and state that the optimal monitoring network design depends on "the characteristics of the target spatial domain and intended inference". Inference scenarios considered are covariance parameter estimation, prediction with known/estimated covariance parameters and fixed effects estimation with known/estimated covariance parameters. Som et al. (2014) take an exhaustive approach to finding the optimal subset of 6 monitoring sites from a simulated network of 21 sites of by calculating the criteria for all possible subsets of a simulated network and choosing the subset which minimises each criterion of interest. This is followed by a complicated stratified selection method applied to larger synthetic networks. Falk et al. (2014) use a pseudo-Bayesian approach to find the optimal sampling design by maximising four different utility functions for synthetic and real data sets. This is followed by estimating the optimal subset of 22 monitoring sites from an existing network of 88 sampling locations for continuous, binary and count observations collected in Queensland, Australia.

The literature on monitoring network design and reduction conclude in general that it is crucial to consider the purpose of the monitoring network as part of the design process. Spatially regular designs are most useful for parameter estimation whereas clustered designs are more appropriate for prediction purposes, especially in heterogeneous areas.

Many of the examples of monitoring network reduction discussed here such as Som et al. (2014) and Falk et al. (2014) aim to find the optimal reduced network of a specified size and use complex search algorithms to achieve this. This chapter does not aim to find an optimal reduced monitoring network but rather to assess the ability of monitoring networks of varying size to estimate parameters and make predictions. This is achieved by sampling the existing monitoring network several times to create many subnetworks and summarizing parameter estimates and predictions from statistical models based on these subnetworks. Specifically, this simulation study aims to

- investigate how covariance function parameter estimates are affected by reducing the size of the monitoring network.

- assess the differences between predictions made using statistical models built from different sizes of monitoring networks.
- assess how uncertainty of predictions is affected by reducing the size of the monitoring network.

Subnetworks can be created using a variety of sampling schemes and the study presented in this chapter considers simple random sampling, weighted sampling and stratified sampling. Simple random sampling assigns equal probability to each monitoring site in the full network of being included in the subnetwork. This is easy to implement but can result in poor coverage of the network since no account is taken of specific network features. Instead, a weighted sampling approach might be preferable where weights are assigned to each monitoring site so that some monitoring sites have a higher probability of being included in the subnetwork. Diaz-Ramos et al. (1996) note that estimates based on a weighted sample will tend to have lower variance if the response of interest and the variable used for weighting are strongly positively correlated. This means that there is less uncertainty associated with estimates of model parameters or predictions if the variable used for weighting is informative about the response of interest. Alternatively, Dobbie et al. (2008) discuss using expert elicitation to determine weights if this information is available. River network structure can be included in the sampling scheme via stratified sampling. Two stratified sampling schemes are considered in this chapter: the first is based on Strahler number of stream segments (Strahler, 1957) and is called proportional sampling and the second is Neyman stratification (Neyman, 1934). Strahler number for stream order is given by allocating order=1 to stream segments that have a source point at the most upstream location on the segment. When two streams of the same order join at a confluence, the order is increased by 1 (see Figure 4.1). This means that the main stem of the network has the highest order number. Strata are defined by Strahler number in this chapter so that subnetworks reflect the structure of the river network. Proportional sampling means that subnetworks reflects the composition of the full network. For example, if 33% of monitoring sites have Strahler number 1 in the full network then 33% of the monitoring sites in the subnetwork will have Strahler number 1. Neyman sampling aims to maximise precision given a fixed sample size. This type of stratified sampling reflects variability in the data and higher proportions of the sample are drawn from strata with higher variability using

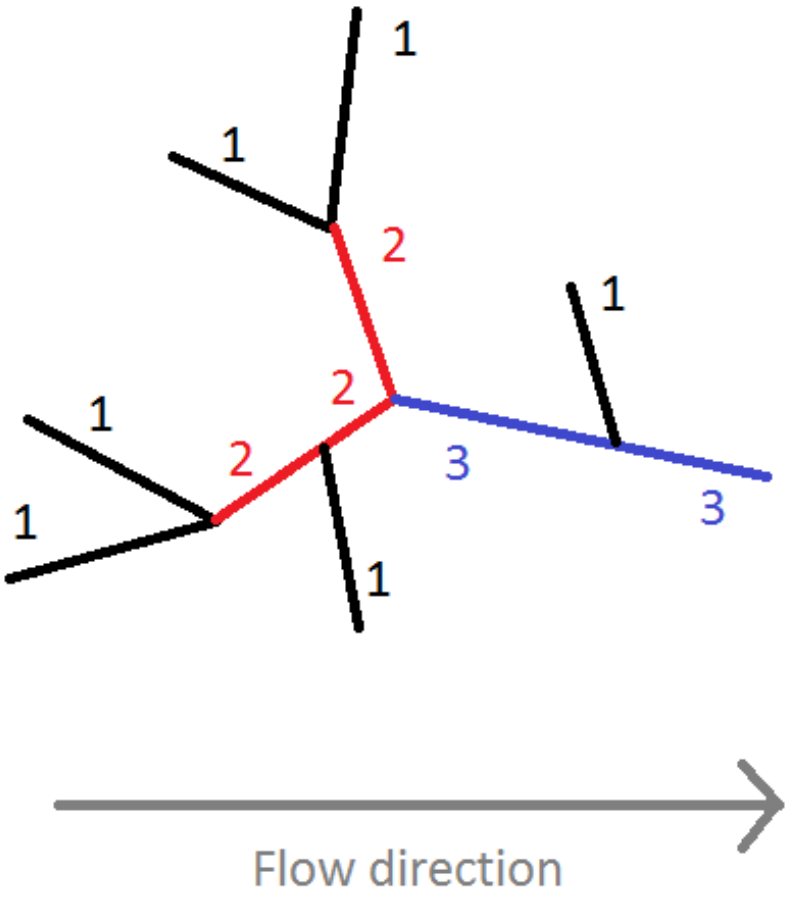


FIGURE 4.1: Graphical representation of Strahler stream order.

$$n_i = \frac{n * (N_i * S_i)}{\sum (N_i * S_i)},$$

where n_i is the sample size in stratum i, n is the total sample size of the subnetwork, N_i is the population size of stratum i and S_i is the standard deviation of stratum i.

Data from the Trent area, introduced in Chapter 3, are used to conduct this study and the response of interest is winL - winter log(TON) averaged over 2003-2010 (referred to as the 'later years' in Chapter 3). Assuming that estimated covariance parameters and predictions from the model based on all 687 monitoring sites are the truth, the aim of this study is to learn what information is lost when the number of monitoring sites is

Strahler	Full	Proportional					Neyman				
		90	**80**	**50**	**20**	**10**	**90**	**80**	**50**	**20**	**10**
1	379	341	303	190	76	38	366	325	203	82	41
2	166	150	133	83	33	17	138	123	77	31	15
3	92	83	74	46	18	9	82	73	46	18	9
4	32	29	26	16	7	3	28	25	15	6	3
5	18	16	14	9	4	2	5	4	3	1	1

TABLE 4.1: Number of monitoring sites to be selected by simple random sampling within each stratum. The numbers in bold indicate the percentage of sites retained in a subnetwork under each stratified sampling scheme and rows represent the five strata.

reduced. Various model outputs are compared between the 'full' model (based on all 687 monitoring sites) and 'reduced' models (based on subsets of the monitoring sites). Reduced models are referred to throughout this chapter as subnetwork$_k$ where k is the percentage of sites retained in the subnetwork (90%, 80%, 50%, 20% and 10%). For simple random sampling and weighted sampling this means that subnetworks consist of 619, 550, 344, 138 and 69 sites respectively. Table 4.1 shows the number of monitoring sites selected for each strata under the two stratified sampling schemes. The biggest differences between proportional and Neyman sampling can be seen in strata 1 and 5 where stratum 1 has more monitoring sites selected under Neyman sampling than under proportional sampling since winL is more variable in the most upstream stream segments than in the main stem. The weighed sampling scheme uses the proportion of arable land in the catchment area around each monitoring site as weights. This is similar to the sampling procedure in Kao et al. (2008) and ensures that areas more likely to have high pollution have a higher probability of being included in the subnetworks. A small number (0.01) was added to sites whose proportion of arable land was zero to ensure no sites were excluded. Monitoring sites with a high proportion of arable land were therefore more likely to be included in the subnetworks. The size of each subnetwork is the same as for the random samples. Proportion of arable land is highly correlated with winL and so according to Diaz-Ramos et al. (1996) should provide parameter estimates and predictions with lower uncertainty.

It is of interest to investigate the effect of reducing the number of monitoring sites on (1) estimated covariance function parameters and (2) predictions at unsampled locations and their associated uncertainties. The response of interest here is winL since log(TON)

is highest in winter months and it is therefore important to understand how reducing the size of the monitoring network affects inferences made about winL. Due to legal limits of log(TON) in stream networks in Europe (European Parliament, 2000) it is important to understand if the choice of sampling scheme might affect inferences made from the results. For example, if one sampling scheme is more likely to overestimate or underestimate winL compared to other sampling schemes. The simulation study is carried out as follows:

1. Select a subnetwork retaining a proportion k of sites where $k = 0.9, 0.8, 0.5, 0.2, 0.1$ under random, weighted or stratified sampling schemes and fit the following model to i monitoring sites in the subnetwork using the `SSN` package (ver Hoef et al., 2014) package in `R`:

$$y_i = \beta_0 + \beta_1 \text{Easting}_i + \beta_2 \text{Northing}_i + \eta_i$$

$$\eta_i = b_{1i} + b_{2i} + \varepsilon_i$$

$$b_1 \sim N(0, \Sigma_{Epanechnikov.Tail-up})$$

$$b_2 \sim N(0, \Sigma_{Gaussian.Euclidean})$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

2. The following model outputs are stored:

   - Covariance parameters (Tail-up range and partial sill, Euclidean range and partial sill, nugget)
   - Variance components
   - Cross validation predictions
   - Cross validation prediction error
   - Standard error of cross validation predictions
   - Predictions at unsampled locations (approx 11,000 locations)
   - Standard error for predictions at unsampled locations

3. Process repeated 500 times to investigate the effect of retaining 90%, 80%, 50%, 20% and 10% of monitoring sites.

where y_i = winL at site $i = 1, ..., kM$ ($M = 566$), β_0, β_1 and β_2 are coefficients estimated using REML, η_i are spatially correlated errors and b_{1i} and b_{2i} are the i^{th} diagonal elements from the Tailup and Euclidean distance covariance matrices $\Sigma_{Epanechnikov.Tail-up}$) and $\Sigma_{Gaussian.Euclidean}$).

4.1.1 Results: covariance function parameters

Following Falk et al. (2014) who assume that covariance function parameter estimates based on all monitoring sites are the true values, estimates from the full network are used here as a baseline for comparison with parameters estimated from subnetworks.

Figure 4.2 shows the quartiles of the Tail-up range parameter. The natural logarithm of the parameter is plotted as the interquartile range is much larger when 20% or 10% of monitoring sites are retained compared to retaining 90% or 80%. The parameter estimated from the subnetworks is similar to the estimate from the full network when more than 50% of monitoring sites are retained, except for the weighted sampling scheme. The weighted sampling scheme provides estimates quite different to the other sampling schemes when 50% or fewer monitoring sites are retained and has a much larger interquartile range than the other sampling schemes when k=80. Histograms of the Tail-up range parameter estimated from subnetworks show a bimodal distribution for this parameter.

Figure 4.3 shows the range parameter for the Euclidean component of the spatial covariance function. The values estimated from the subnetworks are close to that estimated from the full network on average when 50% or more monitoring sites are retained. The weighted sampling scheme gives lower estimates of the parameter value than the other sampling schemes for all values of k considered.

Figure 4.4 shows the partial sill parameter for the Tail-up component of the spatial covariance function. The values estimated from subnetworks are similar to that estimated from the full network for all k on average but the interquartile range becomes large when k=20 or k=10. The weighted sampling scheme estimates lower values of the Tail-up partial sill parameter compared to the other sampling schemes when 50% or more monitoring sites are retained.

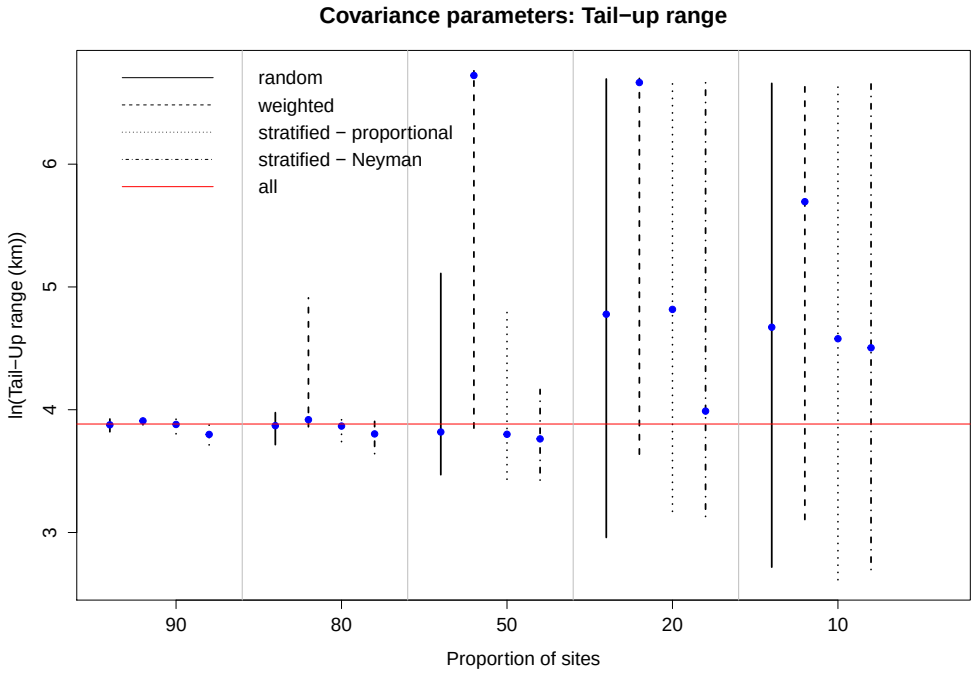


FIGURE 4.2: Interquartile range (lines) of the Tail-up range parameter estimated from subnetworks with the median (dot) highlighted in blue. The natural logarithm of the estimates is displayed. The red line indicates the parameter value estimated from the full network.

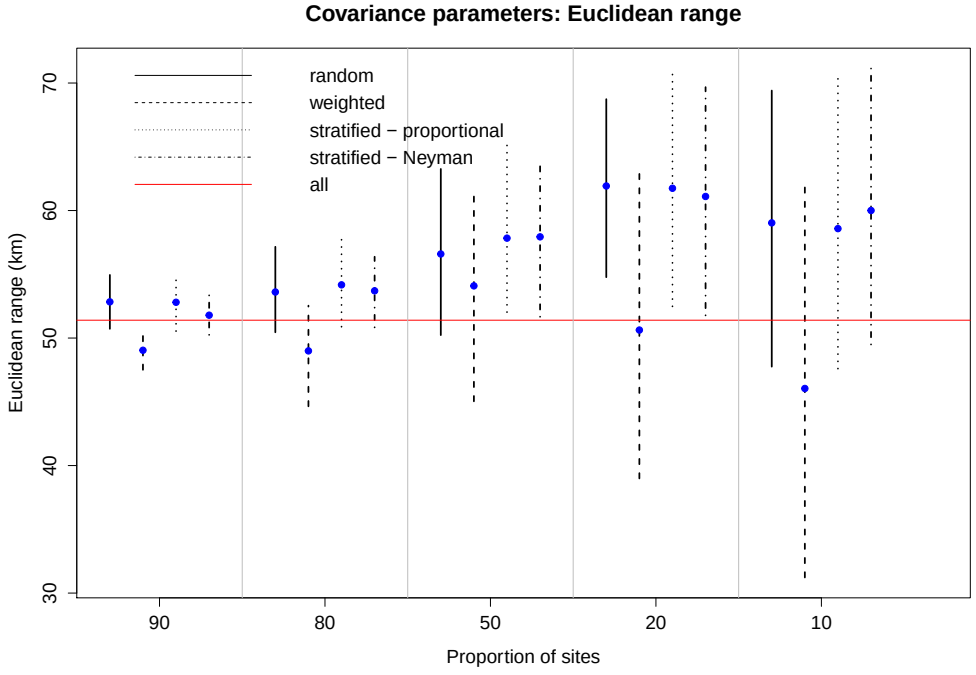


FIGURE 4.3: Interquartile range (lines) of the Euclidean range parameter estimated from subnetworks with the median (dot) highlighted in blue. The red line indicates the parameter value estimated from the full network.

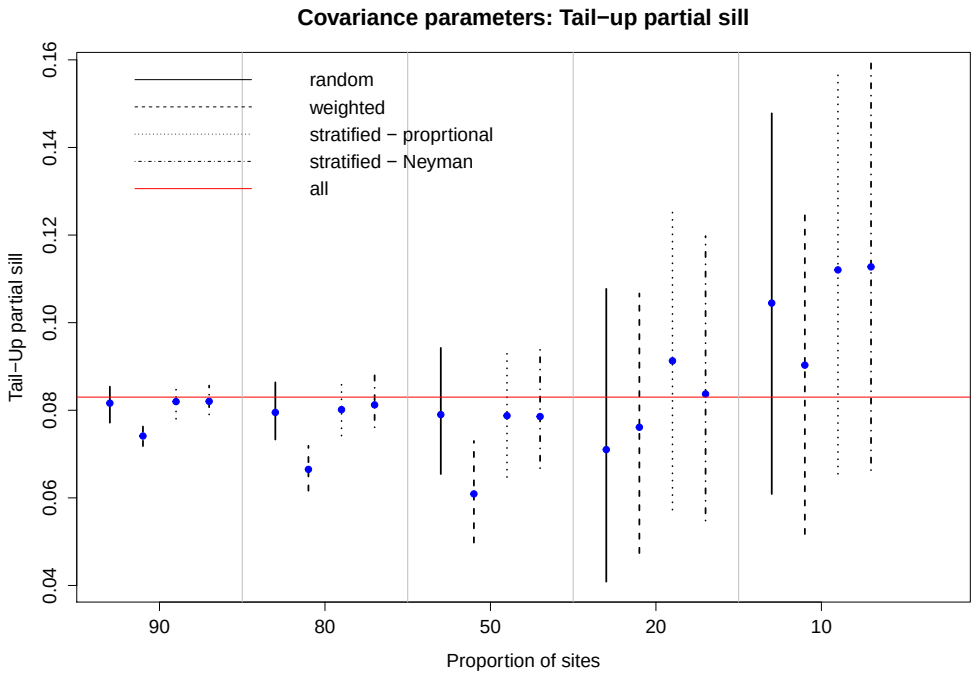


FIGURE 4.4: Interquartile range (lines) of the Tail-up partial sill parameter estimated from subnetworks with the median (dot) highlighted in blue. The red line indicates the parameter value estimated from the full network.

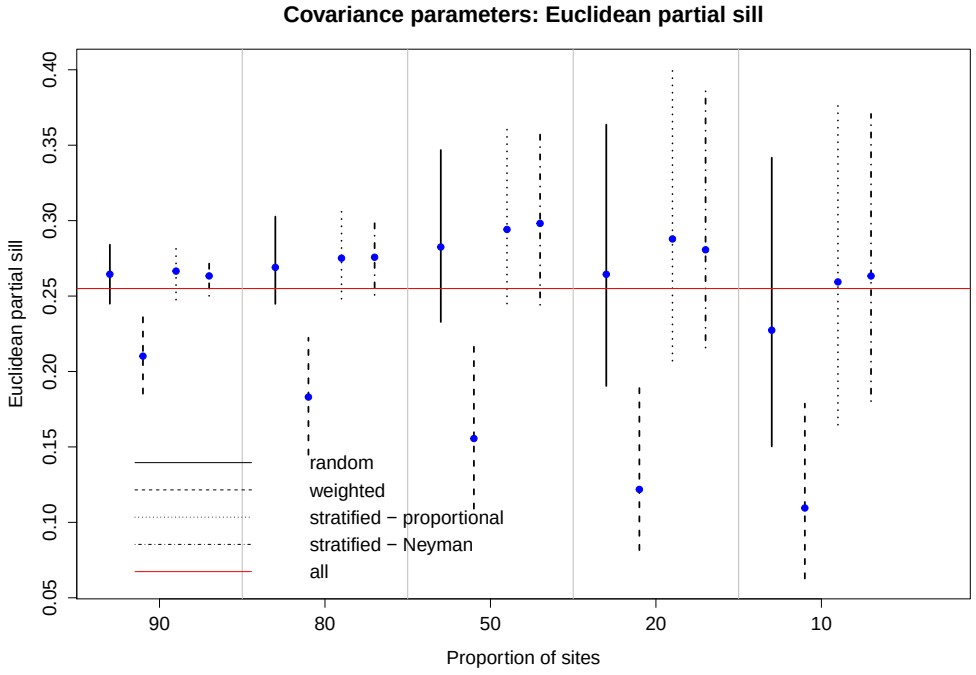


FIGURE 4.5: Interquartile range (lines) of the Euclidean partial sill parameter for subnetworks with the median (dot) highlighted in blue. The red line indicates the parameter value estimated from the full network.

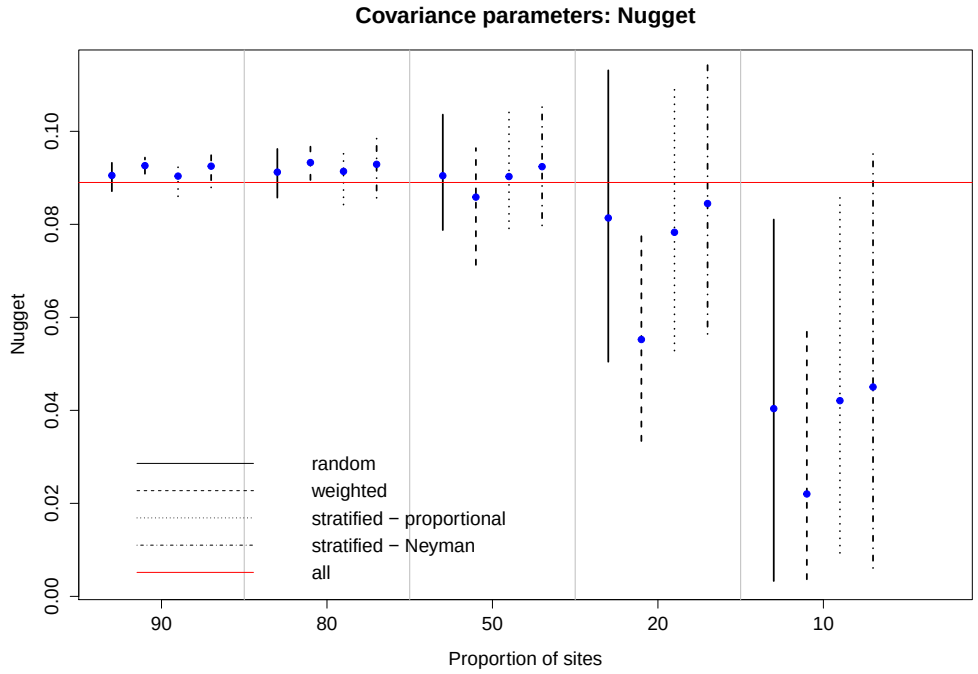


FIGURE 4.6: Interquartile range (lines) of the nugget parameter for subnetworks with the median (dot) highlighted in blue. The red line indicates the parameter value estimated from the full network.

Figure 4.5 shows the partial sill parameter for the Euclidean component of the spatial covariance function. The random and stratified sampling schemes estimate similar values to that estimated from the full network but the weighted sampling scheme estimates the parameter to be much lower than the other sampling schemes.

Figure 4.6 shows the nugget parameter for the spatial covariance function. The nugget values estimated from subnetworks are similar on average to the estimate from the full network when k=90, 80, 50 or 20 but the interquartile range is large for k=20. The weighted sampling scheme estimates the nugget parameter to be lower than the other sampling schemes when 20% or fewer monitoring sites are retained.

Figure 4.7 shows R-squared calculated for subnetworks. In subnetworks of all sizes the linear effects of Easting and Northing explain very little of the variability in winter log(TON). The contribution of R-squared estimated by the weighted sampling scheme is quite similar on average to the contribution estimated from the full network. The random and stratified sampling schemes estimate the contribution of R-squared to increase on average as k decreases.

Figures 4.8 and 4.9 show the contribution of the Tail-up and Euclidean components to

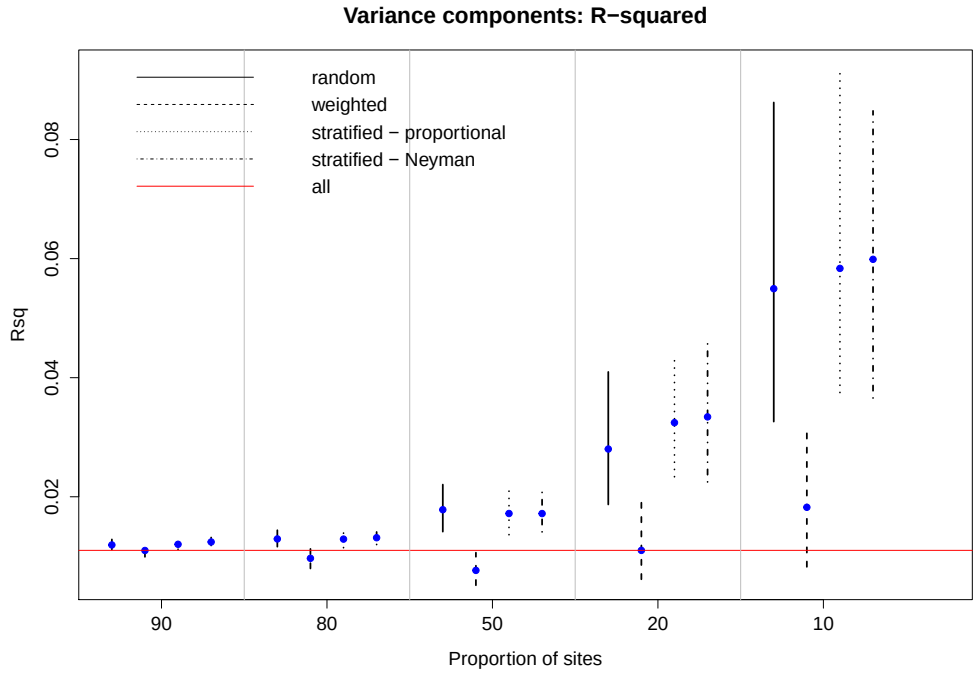


FIGURE 4.7: Interquartile range (lines) of the R-squared variance component for subnetwork with the median (dot) highlighted in blue. The red line indicates the estimated contribution of R-squared to total error variance when estimated from the full network.

total variance, respectively. The contribution of the Tail-up component to total variance estimated by subnetworks of all sizes is similar on average to the estimated contribution from the full network, although the interquartile range is large when k=20 or 10. The weighted sampling scheme however estimates the Tail-up component to have a greater contribution on average when 20% or fewer monitoring sites are retained, compared to the contribution estimated from the full dataset. The weighted sampling scheme estimates a lower contribution from the Euclidean component for all k compared to the contribution estimated from other sampling schemes and the full network. The random and stratified sampling schemes estimate the Euclidean contribution closer to that estimated from the full network for all k, with the interquartile range increasing as k decreases. The contribution of the nugget estimated from the random and stratified sampling schemes is similar to that estimated from the full network for k=90, 80 or 50. The weighted sampling scheme estimates the nugget contribution to be higher than that estimated from the full network when k=90, 80 or 50 but the contribution is similar to that estimated from the other sampling schemes when k=20 or 10.

Pairs plots of covariance function parameter estimates are given in Figures 4.11 and 4.12 to investigate relationships between covariance parameters. It seems that the

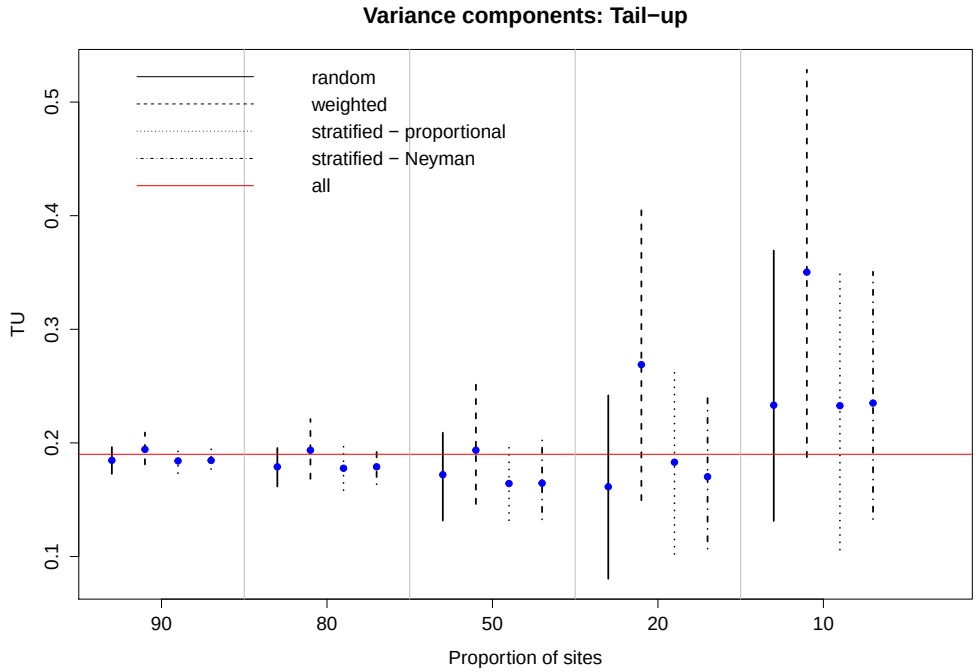


FIGURE 4.8: Interquartile range (lines) of the Tail-up variance component for subnetwork with the median (dot) highlighted in blue. The red line indicates the estimated contribution of the Tail-up component to total error variance when estimated from the full network.

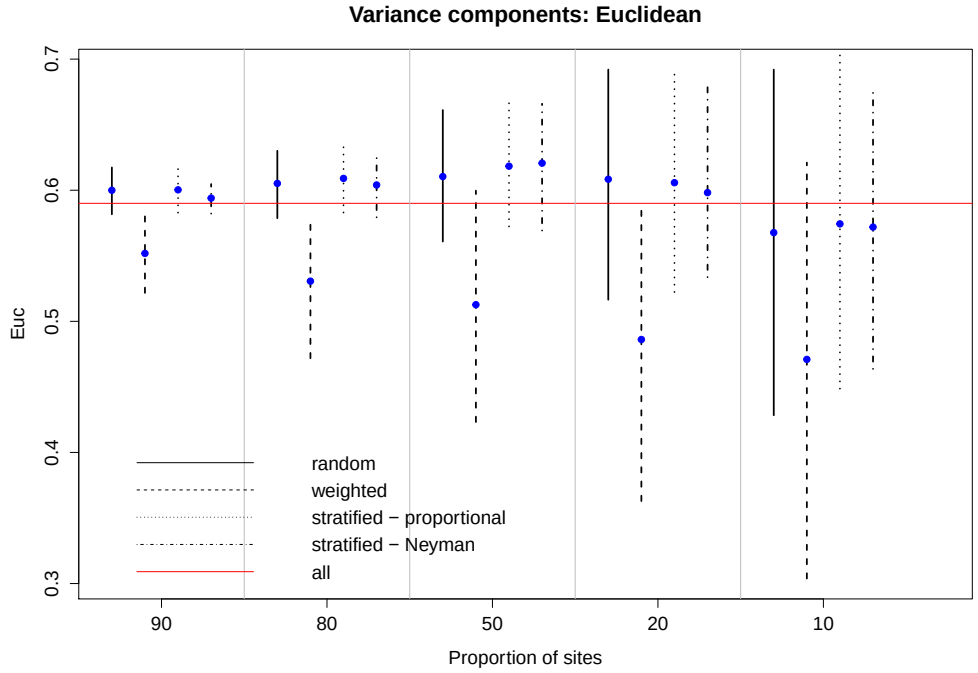


FIGURE 4.9: Interquartile range (lines) of the Euclidean variance component for subnetwork with the median (dot) highlighted in blue. The red line indicates the estimated contribution of the Euclidean component to total error variance when estimated from the full network.

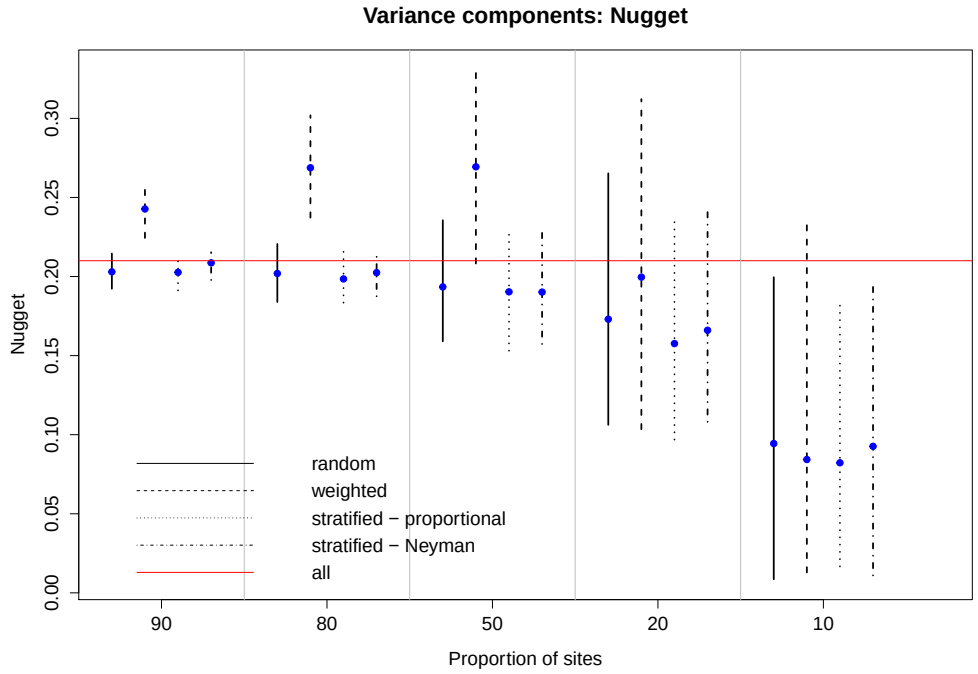


FIGURE 4.10: Interquartile range (lines) of the nugget variance component for subnetwork with the median (dot) highlighted in blue. The red line indicates the estimated contribution of the nugget component to total error variance when estimated from the full network.

nugget decreases as the Tail-up partial sill increases and that as the Euclidean range increases the Euclidean partial sill also increases. These relationships are much stronger for subnetwork$_{90}$ than in subnetwork$_{10}$.

4.1.1.1 Summary: covariance function parameters

Covariance parameters estimated from subnetworks have median values close to those estimated from the full network when 50% or more monitoring sites are retained, although the interquartile range of parameters is much greater when 50% or fewer monitoring sites are retained. The interquartile range of parameter estimates increases as the number of monitoring sites retained decreases. The parameter estimates from the weighted sampling scheme tended to be different from those estimated from the random and stratified sampling schemes.

The median value for variance components estimated from 500 subnetworks is close to the variance components estimated from the full network for all k considered and all sampling schemes with the exception that the weighted sampling scheme estimates the

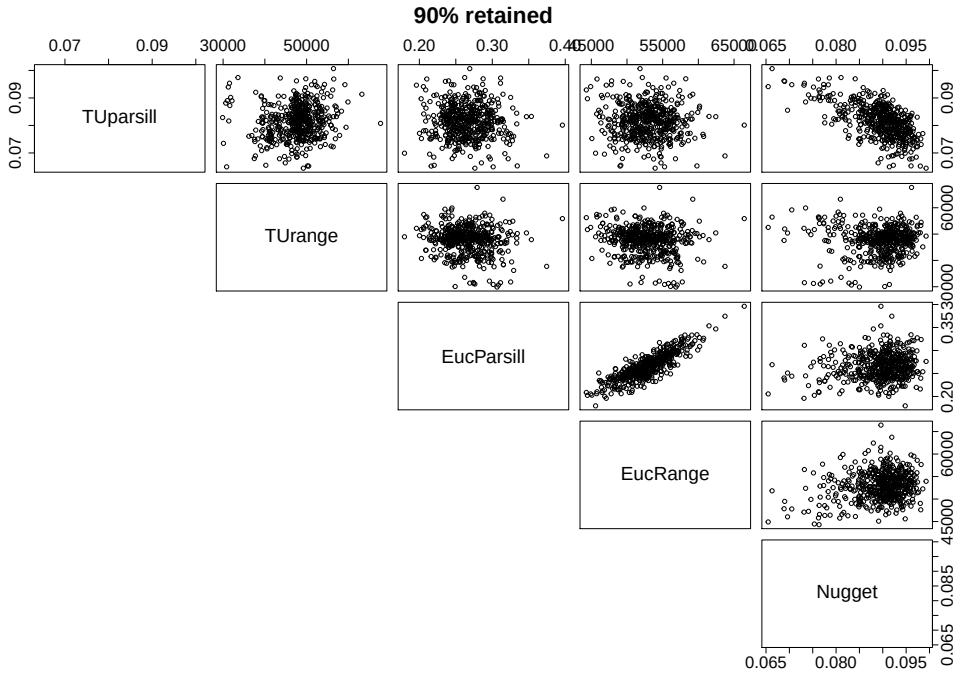


FIGURE 4.11: Covariance parameters for subnetwork$_{90}$ selected under random sampling scheme. Tail-up range parameter >80km have been excluded to improve the viewing.

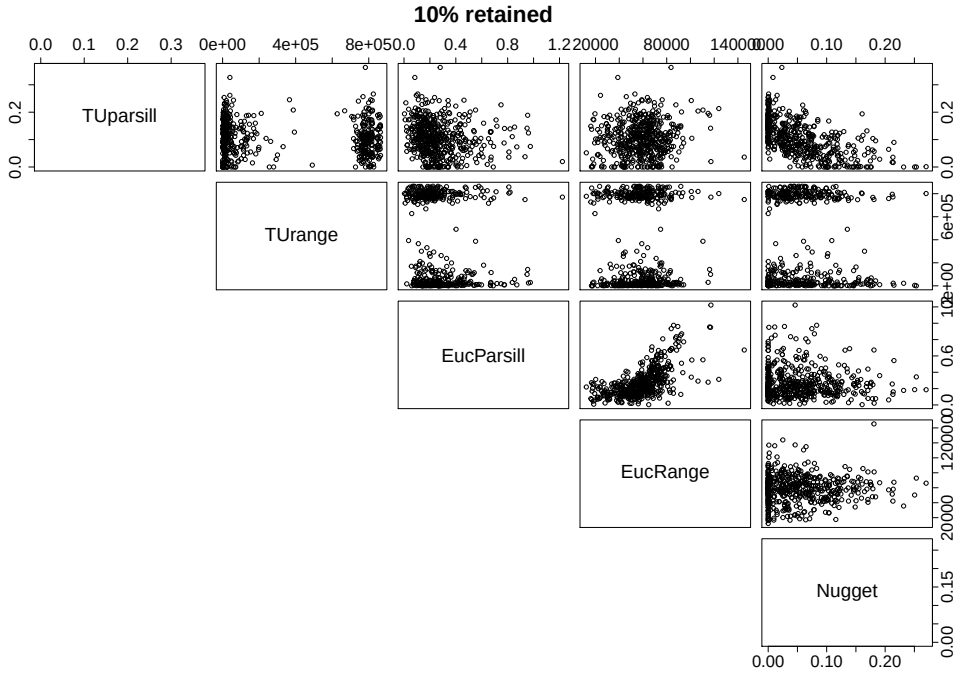


FIGURE 4.12: Covariance parameters for subnetwork$_{10}$ selected under random sampling scheme. Euclidean range parameter >200km and Euclidean partial sill parameter >3 have been excluded to improve the viewing.

contribution of the Euclidean component to be much lower than the other sampling schemes. This is balanced by an increased contribution from the nugget component. As with the covariance parameter estimates, the interquartile range of the variance components increases as k decreases.

4.1.2 Results: predictions

One aim of modelling river network data is to produce predictions at unobserved locations so it is of interest to try to quantify the 'information lost' from these predictions when the size of the monitoring network is reduced. Metrics used in the literature to compare predictions made from models based on different spatial interpolation techniques are used here to compare values predicted at unsampled locations between subnetworks. Robinson and Metternicht (2005) suggest a variety of measures to compare predictions from models based on kriging, inverse distance weighting (IDW) and splines and apply this to soil properties. Murphy et al. (2010) use a selection of these measures to compare predictions from models based on IDW, ordinary kriging and universal kriging with an application to water quality data. Wu et al. (2010) suggest comparing subnetworks by considering kriging predictions from models based on the full monitoring network as true values and comparing these to kriging predictions from models based on subnetworks, with the aim of minimising the root mean squared difference between these predictions. Ferreyra et al. (2002) also compare predictions from models based on full and subnetworks.

One metric commonly used to quantify the accuracy of predictions from a geostatistical model is root mean square prediction error (RMSPE) - a measure of the difference between true (observed) values and a value estimated from Leave One Out Cross Validation (LOOCV). Cross validation is a way of assessing how well a model can be generalized to a new data set and avoids the problem of 'redundant data' where the data set is split into a test set and a training set, meaning not all of the available data are used to build the model. In LOOCV the geostatistical model is fitted to all of the data points except one and the model is then used to predict the response at the excluded location. This is done for each of the N data points and RMSPE (4.1) is calculated to summarise the prediction error of the model, where $\hat{Y}(s_i)$ is the predicted value at location s_i and $Y(s_i)$

is the observed value at location s_i. Models with lower prediction error are more desirable. Chapter 1 also discussed Generalised Cross Validation, suitable for large data sets. LOOCV was used in this study however since the sample size did not make LOOCV computationally time consuming.

$$\text{RMSPE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(\hat{Y}(s_i) - Y(s_i)\right)^2} \tag{4.1}$$

Comparisons are also made here for different sizes of subnetworks and sampling schemes by considering prediction error (4.2). Prediction error is similar to RMSPE but the difference is calculated between predictions at unobserved locations made from the model based on the full network and predictions at unobserved locations from the model based on subnetworks (4.2) where N_{pred} is the number of prediction locations, $\hat{Y}_{full}(u_i)$ is the prediction at unobserved location u_i from the model based on the full network and $\hat{Y}_{subnetwork_k}(u_i)$ is the prediction at u_i from the model based on subnetwork).

$$\text{Prediction error} = \sqrt{\frac{1}{N_{pred}}\sum_{i=1}^{n}\left(\hat{Y}_{full}(u_i) - \hat{Y}_{subnetwork_k}(u_i)\right)^2} \tag{4.2}$$

The ratio of average kriging standard error (AKSE ratio) calculated for the model based on subnetworks and for the model based on the full network can be used to investigate the uncertainty of predictions at unobserved locations. AKSE can be calculated as in (4.3), where $\sigma^2(u_i)$ is the squared standard error of the kriging prediction at unobserved site u_i. AKSE ratio is therefore $\text{AKSE}_{subnetwork_k}/\text{AKSE}_{full}$. A ratio of 1 means that the uncertainty of predicted values based on the subnetwork is the same as the uncertainty of predicted values based on the full network and would suggest no information is lost by reducing the size of the monitoring network.

$$\text{AKSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sigma^2(u_i)} \tag{4.3}$$

The sampling schemes and subnetworks could be compared using other metrics than those considered here. For example, the criteria discussed in Som et al. (2014) or the utility functions used in Falk et al. (2014) could be compared between sampling schemes

and subnetworks depending on the intended inference, with the latter being of interest if a Bayesian approach is taken.

Figure 4.13 shows the interquartile range for RMSPE calculated from subnetworks selected using a variety of sampling schemes. As expected, RMSPE tends to increase as the size of subnetwork decreases. RMSPE is higher on average for subnetworks than for the full data set. Interestingly, RMSPE under the weighted sampling scheme remains fairly constant on average for all subnetworks and is generally lower than under the other sampling schemes. It is likely that subnetworks selected under the weighted sampling scheme will contain a larger proportion of monitoring sites with high values of winL than subnetworks selected by simple random or stratified sampling schemes. Models based on subnetworks with a high proportion of high values of winL will be better able to estimate these high values than models based on subnetworks with a small number of monitoring sites showing high winL values. Figure 4.14 shows the lower and upper quartiles of RMSPE and gives an impression of RMSPE for sizes of subnetworks other than those considered in this study. The lower and upper quartiles are quite similar for all sampling schemes when at least 80% of monitoring sites are retained but the variability in lower and upper quartiles is still quite small when fewer than 80% of monitoring sites are retained.

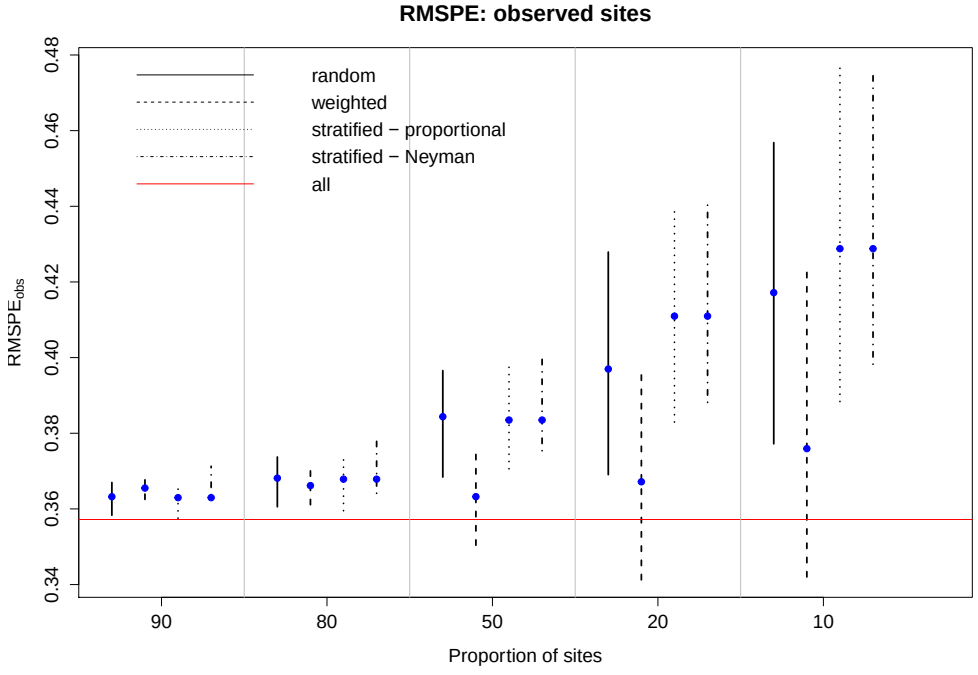


FIGURE 4.13: Interquartile range (lines) for RMSPE from LOOCV with the median (dot) highlighted in blue. The red line indicates RMSPE calculated from full data set.

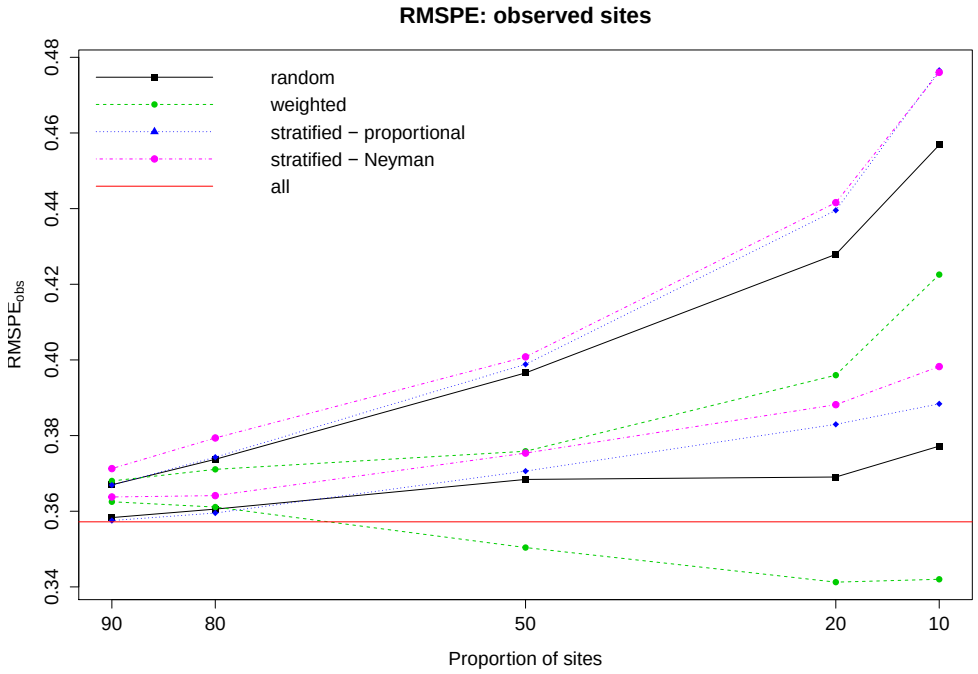


FIGURE 4.14: Lower and upper quartiles for RMSPE from LOOCV. The red line indicates RMSPE calculated from full data set.

Figure 4.15 shows that prediction error increases as the size of subnetwork decreases but subnetworks of all sizes have higher prediction error when selected using a weighted sampling scheme. Since models based on subnetworks from the weighted sampling scheme will estimate higher values of winL on average (due to the monitoring sites with high winL having a higher probability of inclusion in the subnetworks) predictions of winL at unobserved locations will be higher on average than predictions based on subnetworks selected by simple random or stratified sampling schemes. This means that the difference between predicted values at unobserved locations based on subnetworks from weighted sampling schemes and predictions made from models based on the full network will be greater on average than the difference between predictions based on subnetworks selected using simple random or stratified sampling schemes and the full network. The interquartile range for prediction error does not vary much as k decreases. Figure 4.16 shows the lower and upper quartiles of prediction error and gives an impression of these values for subnetworks of sizes other than those considered in this study. Prediction error appears to be quite similar between random and stratified sampling schemes but for very small subnetworks the stratified sampling schemes perform best.

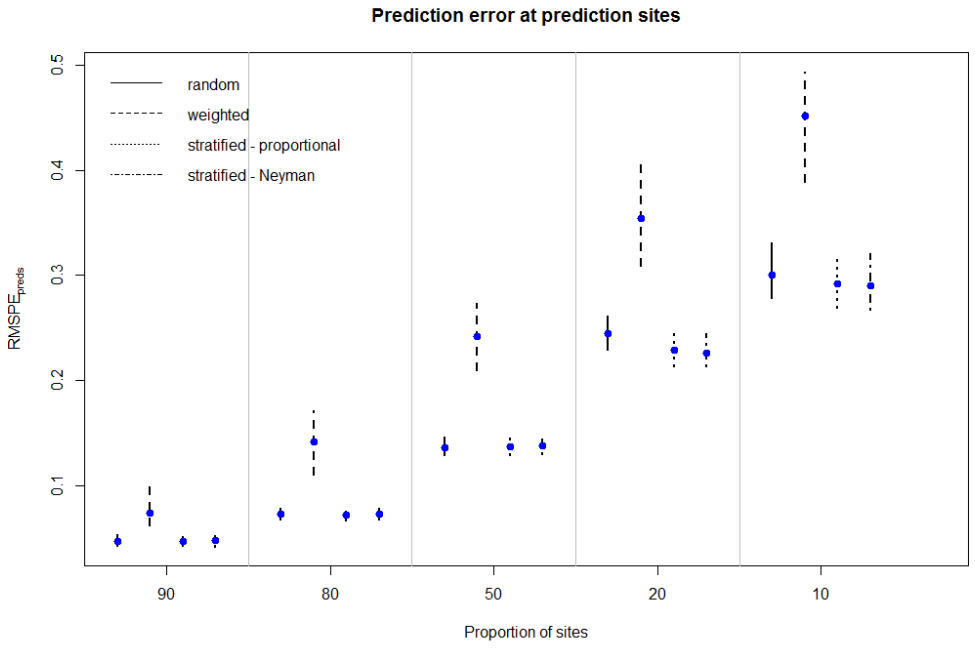


FIGURE 4.15: Interquartile range of prediction error at unobserved locations.

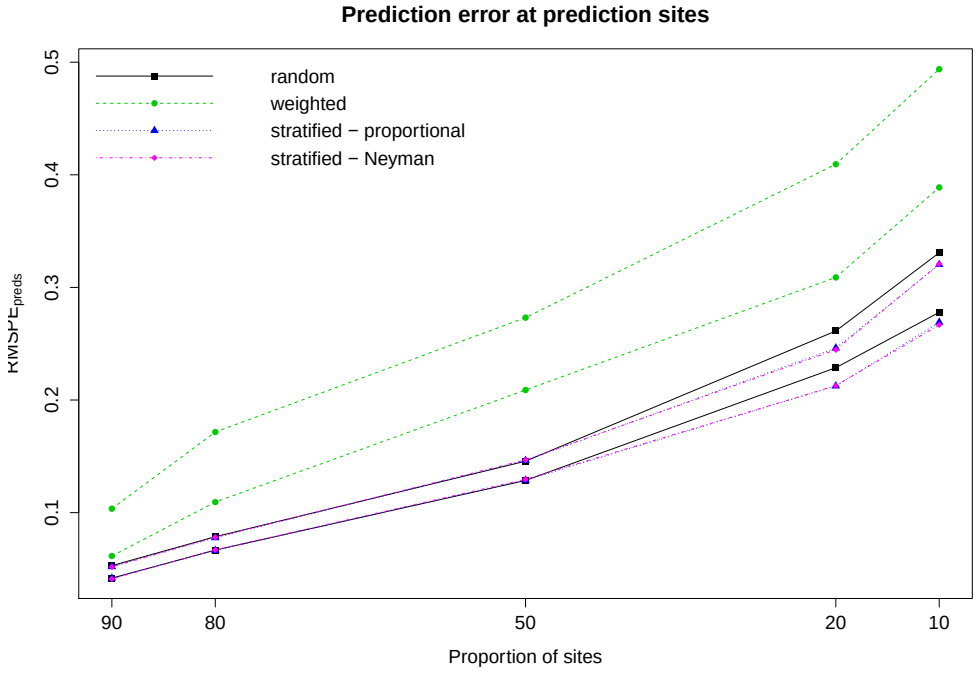


FIGURE 4.16: Lower and upper quartiles of prediction error at unobserved locations.

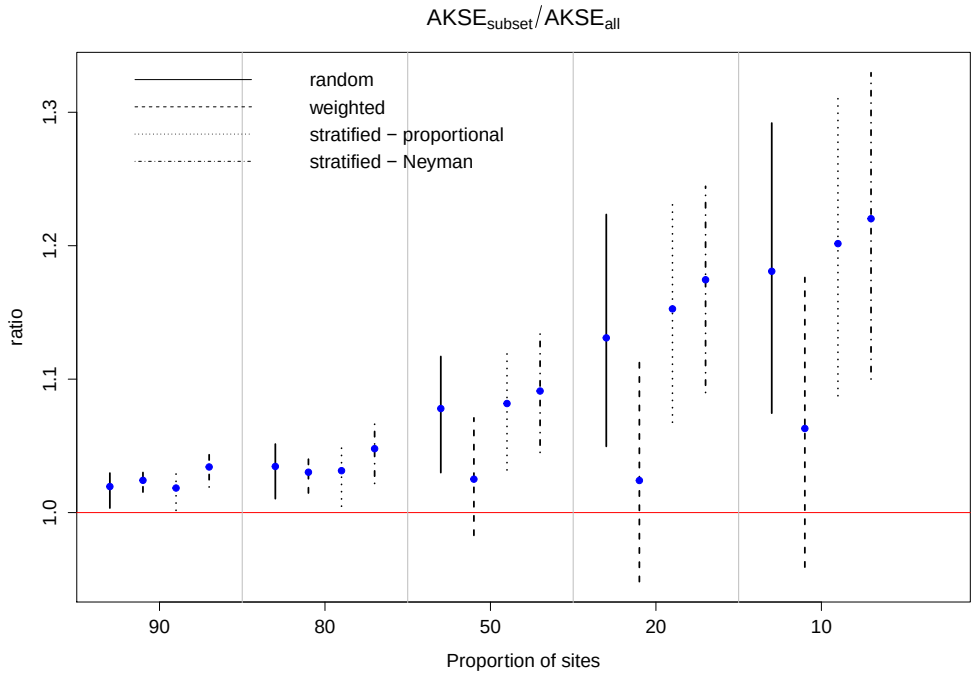


FIGURE 4.17: Ratio of AKSE from subnetworks to AKSE from full network. The horizontal red line is drawn at ratio=1, indicating AKSE in the subnetworks is equal to AKSE in the full network.

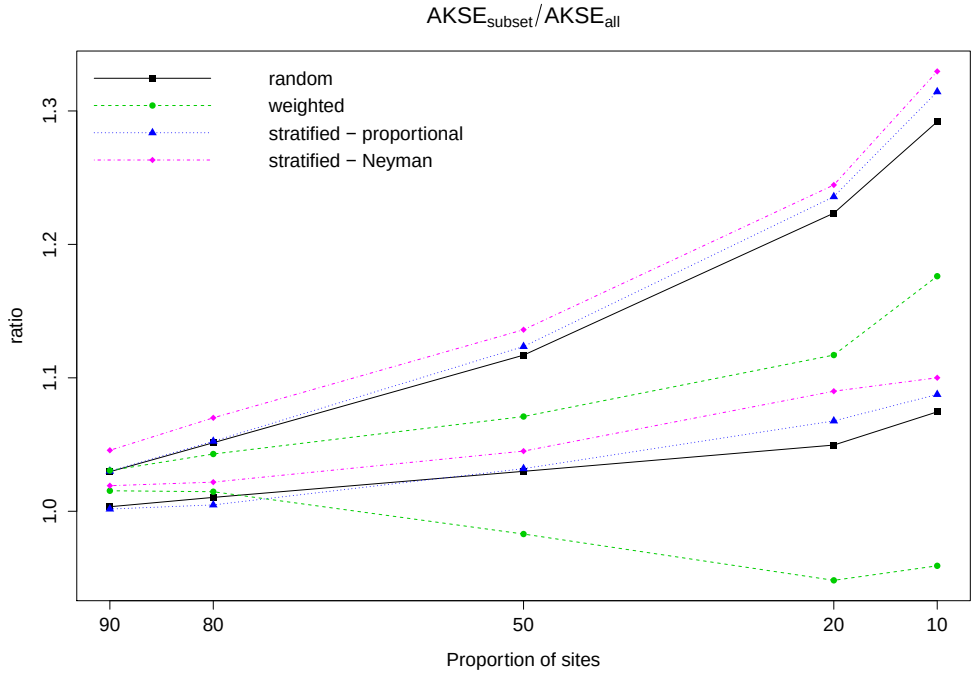


FIGURE 4.18: Ratio of AKSE from subnetworks to AKSE from full network.

Figure 4.17 shows the ratio of AKSE estimated from models based on subnetworks to AKSE from models based on the full network. AKSE ratio for subnetwork_{90} and subnetwork_{80} are centred close to 1 suggesting that AKSE for these subnetworks is, on average, similar to AKSE for the full network. AKSE is around 10% higher (with AKSE ratio centered around 1.1) for subnetwork_{50} and approximately 15% and 20% higher for subnetwork_{20} and subnetwork_{10} respectively. The AKSE ratio for a weighted subnetwork is lower than for the other sampling schemes. The AKSE ratio is smaller for the random sampling scheme compared to the stratified sampling scheme for small subnetworks. Figure 4.18 shows that the AKSE ratio is similar for all subnetworks and sampling schemes when more than 80% of monitoring sites are retained but when fewer than 80% are retained the weighted sampling scheme performs best since the AKSE ratio is more closely centred around 1.

4.1.2.1 Summary: predictions

RMSPE_{obs} increases as k decreases and the interquartile range also follows this pattern. This is seen for all subnetworks under the random and stratified sampling schemes but RMSPE_{obs} is quite stable on average for all k although the interquartile range increases as k decreases. RMPSE_{preds} also increases as k decreases under the random and stratified sampling schemes but the interquartile range is fairly stable. Subnetworks selected using the weighted sampling scheme have higher RMPSE_{preds} and a greater interquartile range compared to the other sampling schemes. The AKSE ratio showed that the uncertainty associated with RMPSE_{preds} increases as k increases for the random and stratified sampling schemes and the ratio is more stable under the weighted sampling scheme. The interquartile range for the AKSE ratio increases as k decreases for all sampling schemes. AKSE is less than 10% greater on average when at least 50% of monitoring sites are retained compared to AKSE estimated from the full network and when 20% or fewer monitoring sites are retained AKSE is 20% to 30% greater compared to the full network.

4.1.3 Summary of simulation study

This simulation study has shown that the parameter estimates from the weighted sampling scheme tended to be different from those estimated from the random and stratified

sampling schemes. The Tail-up range parameter estimated from subsets of the data selected using the weighted sampling scheme showed a bimodal distribution that became increasingly apparent as the proportion of monitoring sites retained decreased. It proved difficult to quantify the particular configuration of monitoring sites that produced the very large Tail-up range parameter but Figure 4.19 shows there is no strong relationship between RMSPE and the Tail-up range parameter. This means that even when the range is estimated to be greater than the longest stream distance between two monitoring sites in the network and thus implying that all monitoring sites are correlated, this does not negatively affect the predictive performance of the model.

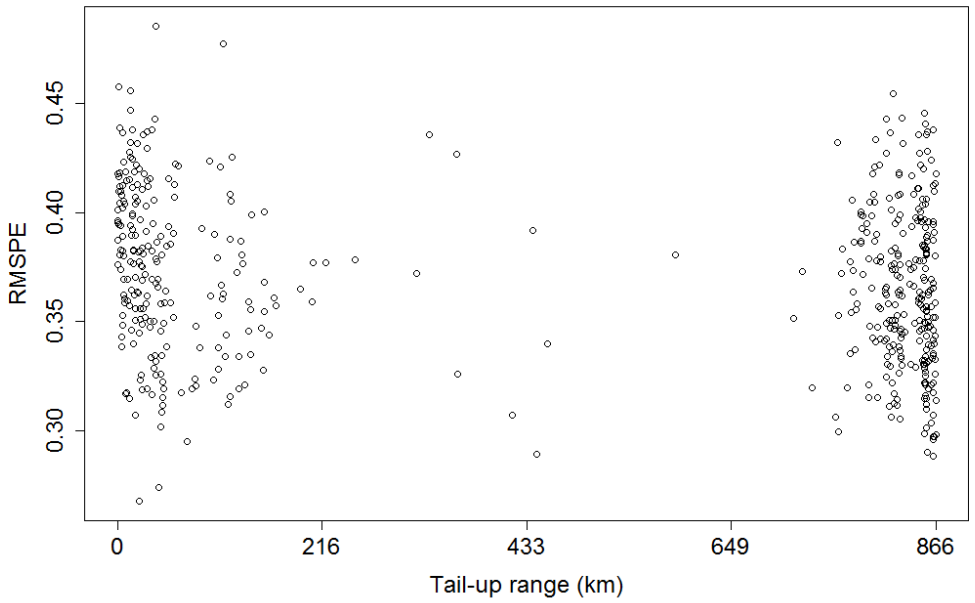


FIGURE 4.19: RMSPE against Tail-up range parameter estimated from subsets of the data selected using a weighted sampling scheme.

Further investigation of the monitoring network size could include repeating the simulation study with summer log(TON) values to assess the effect of reducing the number of monitoring sites when data are lower and highly variable in comparison to winter log(TON) which is higher and less variable. Murphy et al. (2010) show that RMSPE is higher in months where data are more variable. The simulations could be carried out for individual years instead of averages for several years and variability incorporated over time as in Ferreyra et al. (2002). Further work might also include selecting the optimal subnetwork of a particular size using the simulated annealing algorithm in Ferreyra et al.

(2002) combined with optimisation criteria in Som et al. (2014). An application of this can be found in Falk et al. (2014).

Going forward, the monitoring network in the Trent catchment area could be reduced by up to 50%, if an increase in uncertainty of predictions at unobserved locations of up to 10% is acceptable. Monitoring could be reduced by 90% if an increase in uncertainty of 20% on average and up to 30% were acceptable. The Trent area is a densely monitored area with 687 monitoring sites and so removing 50% of monitoring sites would mean 344 sites would be retained thus the Trent would still have a very dense monitoring network compared to other areas of England and Wales. The Trent is a large complex river network with a heterogeneous landscape and therefore will require a greater number of monitoring sites than a small homogeneous area. The conclusion that it would be possible to remove 50% of monitoring sites with little loss of information is not a general rule for LHA's of all sizes but is rather a conclusion applicable to a large, heavily monitored catchment area. In sparsely monitored LHA's it is unlikely that removing 50% of monitoring sites would result in as small a loss of information as in the Trent.

It is recommended, based on the sampling study, that care be taken when choosing an appropriate sampling strategy since weighted sampling leads to predictions of higher value if the variable used to weight the probability of a site being retained is positively correlated with the variable being predicted. The weighted sampling scheme can be thought of as reflecting a worse case scenario if there is an upper safety limit on the variable being predicted but predictions from models based on such a sampling scheme might lead to increased costs as a result of taking unnecessary action to reduce levels of the variable of interest. If the purpose of the analysis is to produce realistic predictions of a variable at unobserved locations then simple random sampling is recommended as a sampling strategy since this is straightforward to implement, is not affected by the choice of variable used for weighting or stratification and provides a smaller increase in the uncertainty of predictions than weighted or stratified sampling methods. Any redesign of the monitoring network should ultimately be based on requirements of the user as discussed in Diggle and Ribeiro (2007).

4.2 Covariates

The models fitted to the stream network data in Chapter 3 included only Easting and Northing as linear covariates and almost all of the spatial pattern is captured using a hybrid spatial covariance structure. In this section covariate data will be included in the model for winL and the effect of this on the covariance structure and predicted values will be investigated.

Rainfall, population density, flow, livestock counts and landcover data will be used as covariates, following the work in Bowman et al. (2010). Neill (1989) show a positive relationship between nitrate load and river flow while Neal et al. (2004) and Burt et al. (1988) discuss the relationship between rainfall and water quality. Pesce and Wunderlin (2000) conclude that water quality is worse (higher levels of stream chemistry variables) during the Argentinian dry season. Neal et al. (2006) discusses the impact of flow on nitrates while Paul and Meyer (2001) look at the effect of population and urbanisation on water quality. Hooda et al. (2000) propose several solutions to the increased nitrate levels in river water as a result of increased numbers of livestock and Gerber et al. (2007) specifically considers poultry. Robson and Neal (1997) note that agriculture, high population density and industrial sewage are sources of nitrates in river water in several English rivers, including the Trent.

The covariate data are available as spatial snapshots meaning they are not recorded over time. Variables were not all available at the same spatial scale and so were aggregated in two ways to have a common spatial scale for all variables: (1) rates per km^2 were calculated where the variable is standardized by the size of the reach catchment area (RCA - the land that drains directly to a particular stream segment), referred to as covariates$_{RCA}$ and (2) accumulated totals where the rate was multiplied by the area of the RCA in km^2 and the total then accumulated from monitoring site to source(s) to represent total contribution of the covariate over all of the land that drains to an individual monitoring site (referred to as covariates$_{acc}$). Examples of rainfall$_{RCA}$ and rainfall$_{acc}$ are shown in Figure 4.20. Covariate values were calculated first for each stream segment and then attributed to any monitoring sites within a segment, following the guidelines in Peterson (2011). Values for multiple sites within a single stream segment differ depending on how far down the stream segment the monitoring site is placed.

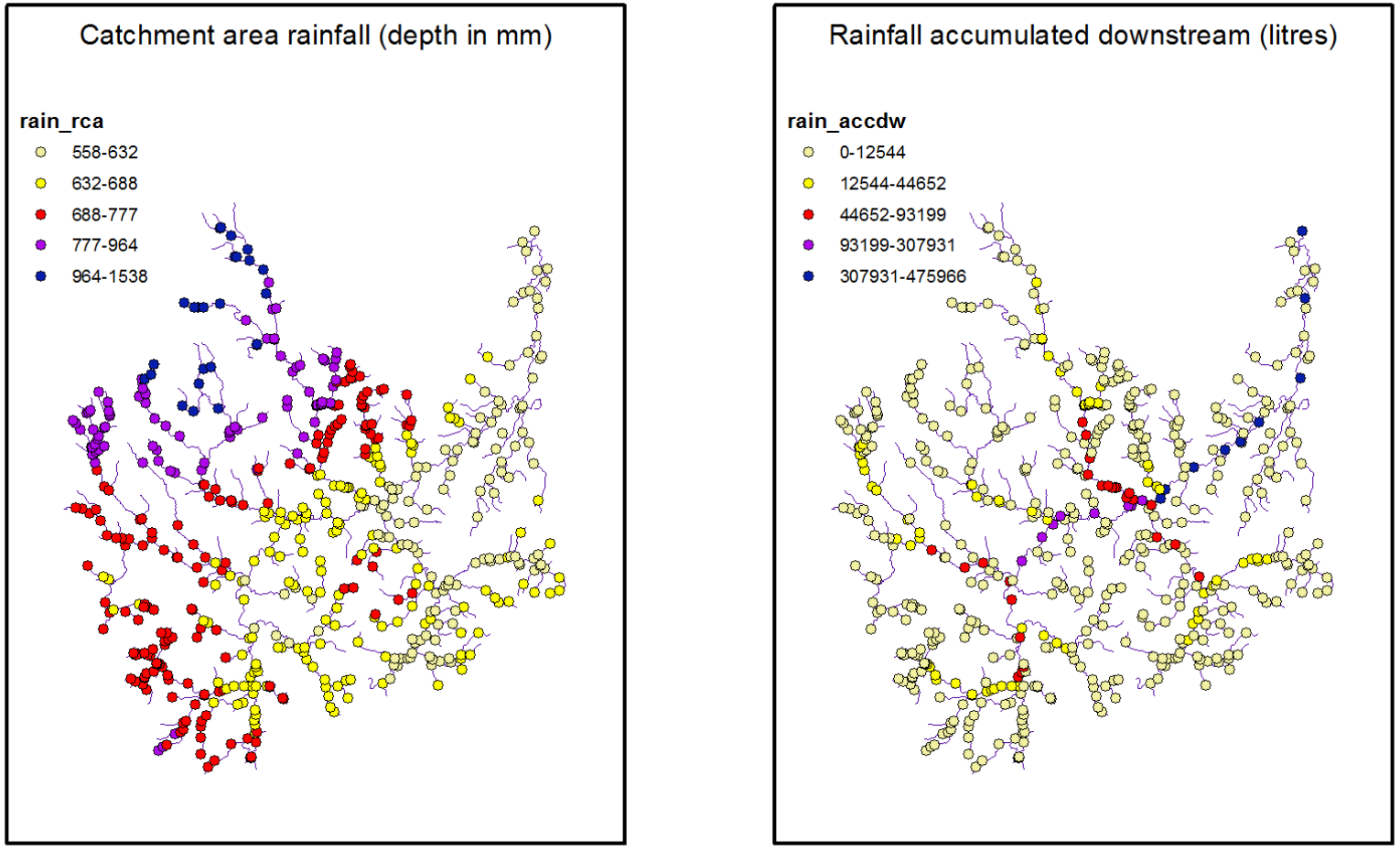


FIGURE 4.20: Rainfall estimated at RCA level from land that drains directly to the stream segment on which the monitoring site is located (left) and total volume rainfall accumulated from source point(s) to monitoring site.

The covariates are described in Section 4.2 and further details of the data processing steps can be found in Appendix A. Following this additive models were fitted to winL that include smooth functions of the covariate data and assume independence between monitoring sites. Additive models were fitted using the `mgcv` package in `R` described in Chapter 1. The predicted values at observed and unobserved locations along with their associated uncertainty will be compared to those from a model fitted to winL with no covariate data but assuming spatially correlated residuals. This allows a comparison of purely covariate vs purely covariance based models. Finally, spatial stream network (SSN) models will be fitted to the independent residuals from the additive models to investigate the nature of any remaining spatial structure and to assess whether the additive model could be improved by accounting for spatial correlation. At present the `SSN` package used to model spatial correlation cannot fit smooth functions of covariates and so a two stage approach is applied here where smooth functions of the covariates are modelled assuming independence after which the residuals are modelled using a spatial correlation function. The recent work in O'Donnell et al. (2014) opens up the possibility of fitting flexible regression models accounting for river network structure in the future, although the focus in their work is on incorporating river network structure into the deterministic part of the model rather than in the error structure as in ver Hoef et al. (2006).

Population

Population data were obtained from `http://www.ons.gov.uk/ons/index.html`. Three sources of information were used to calculate population density: (1) Population count of all ages for each local authority district (LAD) in England and Wales from mid 2012, (2) a shapefile containing boundary information for LAD's from 2012 and (3) area of each LAD in hectares (ha).

Rainfall

Rainfall data were provided by the EA in the form of long term average (1961 to 1990) annual depth in mm. In order to obtain an estimate of total rainfall contributing to a monitoring point it was necessary to calculate rainfall volume within each RCA and

accumulate rainfall volume between monitoring site and source point(s). Rainfall volume (litres) can be calculated as depth(mm)×area(m^2) × 0.001.

Livestock and Crops

Livestock data were provided by the EA and are taken from the ADAS 2010 agricultural census (further details of this dataset can be found at http://edina.ac.uk/agcensus/). The livestock data show the number of animals, broken down by species, in areas of 1km^2 . Table 4.2 shows all of the species included in the agricultural census. In this covariate study all of the livestock excluding chickens were counted together rather than having separate model terms for each species since large numbers of chickens can be accommodated in a small area and the chicken numbers greatly inflate the livestock numbers. Chickens per km^2 and total number of chickens contributing to a monitoring site were therefore calculated separately from other livestock.

Livestock	**Land use categories**
Total number of poultry	Salt water
Number of farmed deer	Arable and horticulture
Total number of goats	Bog
Total number of pigs	Improved grassland
Number of horses and ponies	Built up areas and gardens
Total number of sheep and lambs	Rough low productivity grassland
Total number of cattle and calves	Broad leaved, mixed and yew woodland
	Littoral sediment
	Coniferous woodland
	Inland rock
	Freshwater
	Neutral grassland
	Fen marsh and swamp
	Dwarf shrub heath
	Supra-littoral sediment
	Calcareous grassland
	Montane habitats
	Acid grassland

TABLE 4.2: Livestock from ADAS 2010 agricultural census and Land use categories from LCM2007.

Landcover

Land category information was provided by the EA and obtained from the the LCM2007 dataset (see Morton et al. (2011) for full details of this dataset). LCM2007 gives the proportion of area in small land parcels that is classified using 23 landcover categories, as well as the dominant classification in each land parcel. A single RCA contains many land

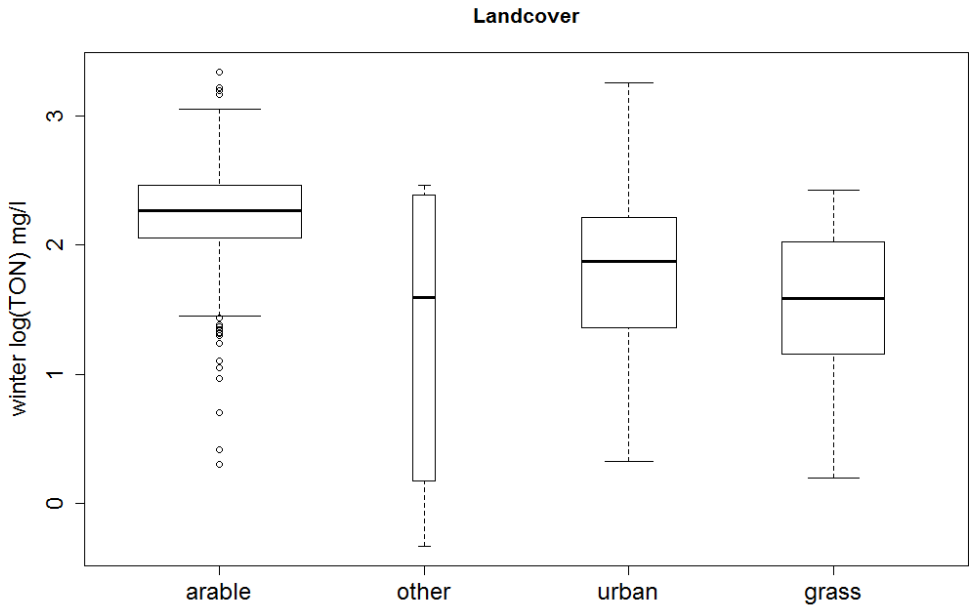


FIGURE 4.21: Landcover categories in the Trent area. The width of the boxes is proportional to the square root of the number of observations in each group.

parcels and so landcover category for an RCA was the category that covered the greatest proportion of land in the RCA. Landcover was not calculated for total contributing area.

Figure 4.21 shows the distribution of landcover categories assigned to monitoring sites in the Trent area. The width of the boxes is proportional to the square root of the number of observations in each group. winL is highest at sites classed as arable, followed by urban and grass. The category called 'other' contains only 7 observations covering 5 landcover categories (Bog; Broad leaved, mixed and yew woodland; Coniferous woodland; Dwarf shrub heath; Neutral grassland) with highly variable winL values.

Flow

Flow data were provided by the EA in the form of a long term average (1961-1990) measured in mean ml per day for each waterbody, a small proportion of the LHA, as defined by the EA. All monitoring sites and prediction locations are assigned the flow value of the waterbody in which they are located.

4.2.1 Modelling covariates

Four types of model were fitted to winL$_m$ where $m = 1, ..., 566$ to investigate the balance between including covariate data and spatial covariance structure in a statistical model. Models that contain covariates aggregated to RCA level (covariates$_{RCA}$) are used to investigate the effect of covariate information on the Euclidean distance component of the spatial correlation function. In addition to this, models that include covariates that are accumulated totals (covariates$_{acc}$) are used to investigate the effect of accumulated covariates on the Tail-up component of the spatial correlation function. A third type of model (covariates$_{mxd}$) combines covariates$_{RCA}$ and covariates$_{acc}$ and is used to investigate the effect of covariates on a hybrid spatial correlation structure. A model containing only location information (covariates$_{loc}$) is used as a baseline comparison for the models with covariate data.

RCA covariates

Figure 4.22 show average winter log(TON) 2003-2010 against RCA level covariates. A natural log transformation of the covariates is used to make the distribution of the data more symmetric. The plots suggest winL has a negative relationship with log(average annual rain depth (mm)) and a positive relationship with log(number of chickens per km^2).

A manual forward stepwise selection process was used to find the best combination of RCA level covariates to describe winL in the Trent area. Smooth functions were fitted between winL and each covariate shown in Figure 4.22 separately using the `mgcv` package in `R`. Smoothing parameters were automatically selected using REML since GCV performs less well when the data are correlated (Wood, 2006). The covariate whose model yielded the highest R^2 value was chosen as the best single RCA level covariate to describe winL. Each of the remaining RCA level covariates was added one at a time to this model and the pair of covariates with the highest R^2 value was selected as the best pair of covariates to describe winL. The model with two covariates was compared to the model with a single covariate using an approximate F test (Wood, 2006) and if $p<0.05$ the more complex model was chosen as best. The selection process was repeated until adding additional model terms did not significantly increase R^2. The final step in choosing the best subset of RCA level covariates was to test if smooth model terms

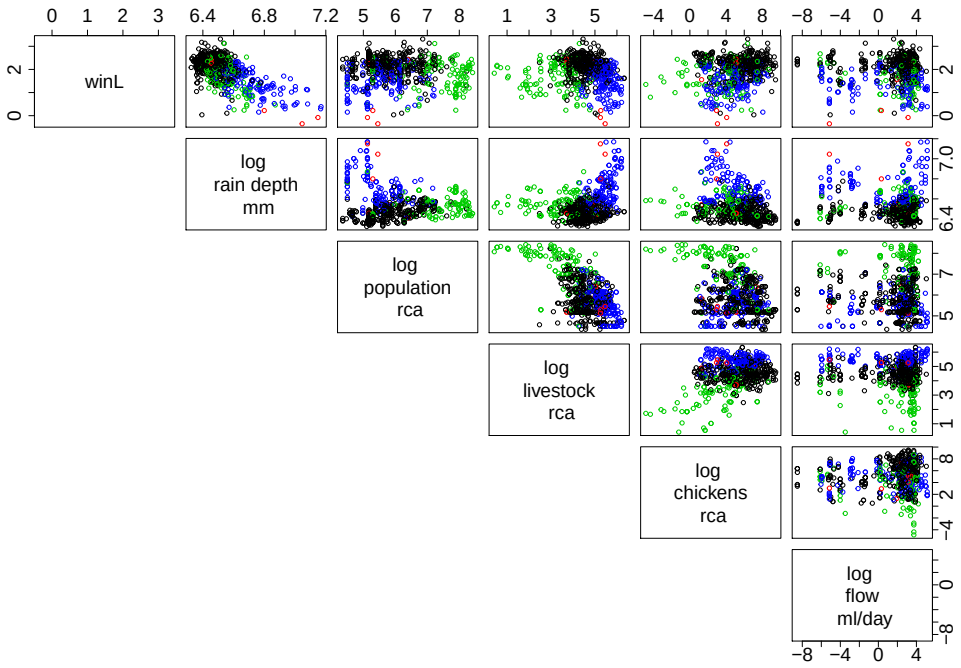


FIGURE 4.22: Plots of average winter log(TON) 2003-2010 (winL) against covariates aggregated to RCA level. "RCA" - values are rates per km^2 . Colours indicate land use category: arable (black), other (red), urban (blue), grass (green).

could be replaced by linear terms. Models with smooth terms were compared to models with linear terms using an approximate F test.

The single best RCA level covariate to describe winL was log(rain depth) and R^2 =45.7%. This increased to 49.4% by adding land use category as a factor variable. Additional covariates did not significantly improve this model and an approximate F test showed that the smooth function of rain could be replaced by a simpler linear term. The single worst performing covariate was chickens/km^2 where R^2 =7.8%. The final covariates$_{RCA}$ model is shown in (4.4) where β_0 is the intercept representing winL where the land use category is arable, β_1, β_2 and β_3 are adjustments to the mean where land use category is other, urban and grass respectively, β_4 is the slope parameter for log(rain depth) and $\varepsilon_m \overset{iid}{\sim} N(0, \sigma^2)$ is independent error.

A term for location, $f_1(\text{easting}_m, \text{northing}_m)$, was added (see 4.5) to account for any location effect not captured by the covariates and this increased R^2 to 59.6%. The location only model, covariates$_{loc}$ (4.6) was also fitted as a comparison and R^2 for this model was 58.2%. The best model for winL as a function of RCA level covariates is given in (4.5) and is a combination of covariates$_{RCA}$ and covariates$_{loc}$.

$$\begin{aligned} \text{winL}_m &= \beta_0 + \beta_1 \text{land}_{other_m} + \beta_2 \text{land}_{urban_m} + \beta_3 \text{land}_{grass_m} + \\ &\quad \beta_4 \text{log.rain.depth}_{\text{RCA}m} + \varepsilon_m \end{aligned} \tag{4.4}$$

$$\begin{aligned} \text{winL}_m &= \beta_5 + \beta_6 \text{land}_{other_m} + \beta_7 \text{land}_{urban_m} + \beta_8 \text{land}_{grass_m} + \\ &\quad \beta_9 \text{log.rain.depth}_{\text{RCA}m} + f_1(\text{Easting}_m, \text{Northing}_m) + \varepsilon_m \end{aligned} \tag{4.5}$$

$$\text{winL}_m = \beta_{10} + f_2(\text{Easting}_m, \text{Northing}_m) + \varepsilon_m \tag{4.6}$$

Accumulated covariates

Figure 4.23 shows winL plotted against accumulated covariates i.e. the covariate values are total values accumulated upstream from monitoring point to source point(s). As with the RCA level covariates, a natural logarithmic transformation has been applied. There appears to be a weak positive relationship between winL and log(chickens$_{acc}$) and all accumulated covariates are correlated with each other. The colours represent Strahler number with 1=headwaters and 5=main stem. Accumulated covariates are strongly related to position in the stream network so accumulated covariate values are smaller near source points and larger near outlet points.

The model selection procedure used for RCA level covariates was also used to select the best subset of accumulated covariates. The best single accumulated covariate was log(chickens$_{acc}$) with R^2 =7.1% and the worst was log(rain volume$_{acc}$) with R^2 =0.01%. The best subset of accumulated covariates is shown in (4.7) with R^2 =19.1% and $f_3()$ and $f_4()$ indicate smooth terms in the model. A term for location was added to this model giving (4.8) with R^2 =60.5% and $f_5()$, $f_6()$ and $f_7()$ are smooth terms. Approximate F tests showed that smooth terms could not be be replaced with linear terms ($p<0.05$).

$$\text{winL}_m = f_3(\text{log.chickens}_{\text{acc}_m}) + f_4(\text{log.livestock}_{\text{acc}_m}) \tag{4.7}$$

$$\begin{aligned} \text{winL}_m &= f_5(\text{Easting}_m, \text{Northing}_m) + f_6(\text{log.chickens}_{\text{acc}_m}) \\ &\quad + f_7(\text{log.livestock}_{\text{acc}_m}) \end{aligned} \tag{4.8}$$

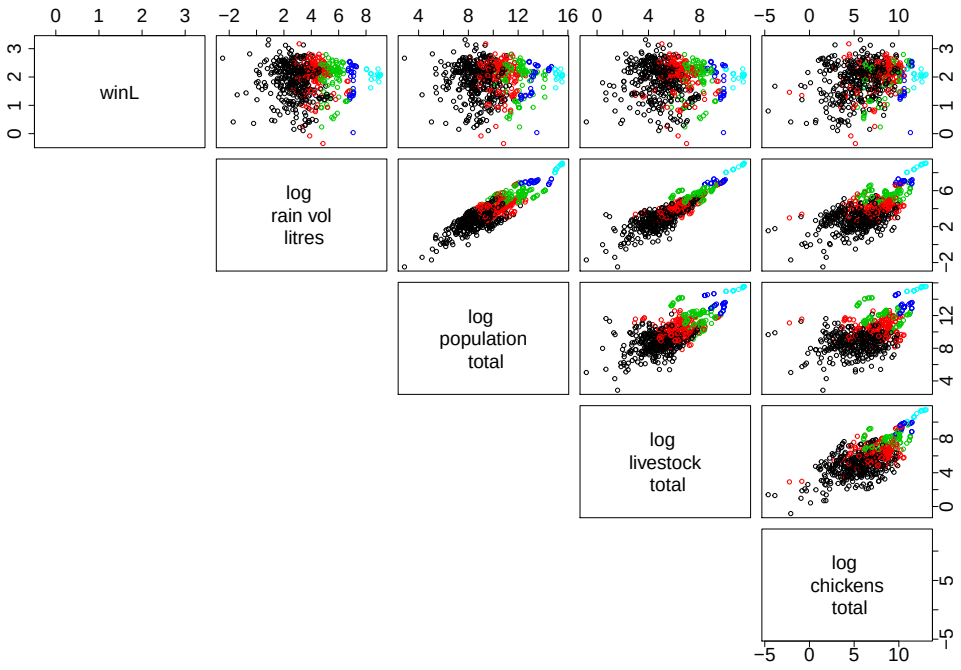


FIGURE 4.23: Plots of average winter log(TON) 2003-2010 (winL) against accumulated totals. Colours indicate Strahler number: 1(black), 2(red), 3(dark blue), 4(green) and 5(pale blue).

Mixed covariates

The final step in choosing a model to describe winL was to combine (4.5) and (4.7) to give a hybrid model combining covariates$_{RCA}$, covariates$_{loc}$ and covariates$_{acc}$ (4.9). This model has R^2 =60.7% and estimated values for model parameters $\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}$ can be found in (4.10) and for model parameters f_8, f_9, f_{10} in Figures 4.24 and 4.25.

$$\begin{aligned} \text{winL}_m &= \beta_{11} + \beta_{12}\text{land}_{other_m} + \beta_{13}\text{land}_{urban_m} + \beta_{14}\text{land}_{grass_m} + \\ &\quad \beta_{15}\text{log.rain.depth}_{\text{RCA}_m} + f_8(\text{Easting}_m, \text{Northing}_m) + f_9(\text{log.livestock}_{\text{acc}_m}) + \\ &\quad f_{10}(\text{log.chickens}_{\text{acc}_m}) + \varepsilon_m \end{aligned} \tag{4.9}$$

$$\begin{aligned} \text{winL}_m &= 12.23 - 0.49 \times \text{land}_{other_m} - 0.16 \times \text{land}_{urban_m} - 0.08 \times \text{land}_{grass_m} - \\ &\quad 1.6 \times \text{log.rain.depth}_{\text{RCA}_m} + f_8(\text{Easting}_m, \text{Northing}_m) + f_9(\text{log.livestock}_{\text{acc}_m}) + \\ &\quad f_{10}(\text{log.chickens}_{\text{acc}_m}) + \varepsilon_m \end{aligned} \tag{4.10}$$

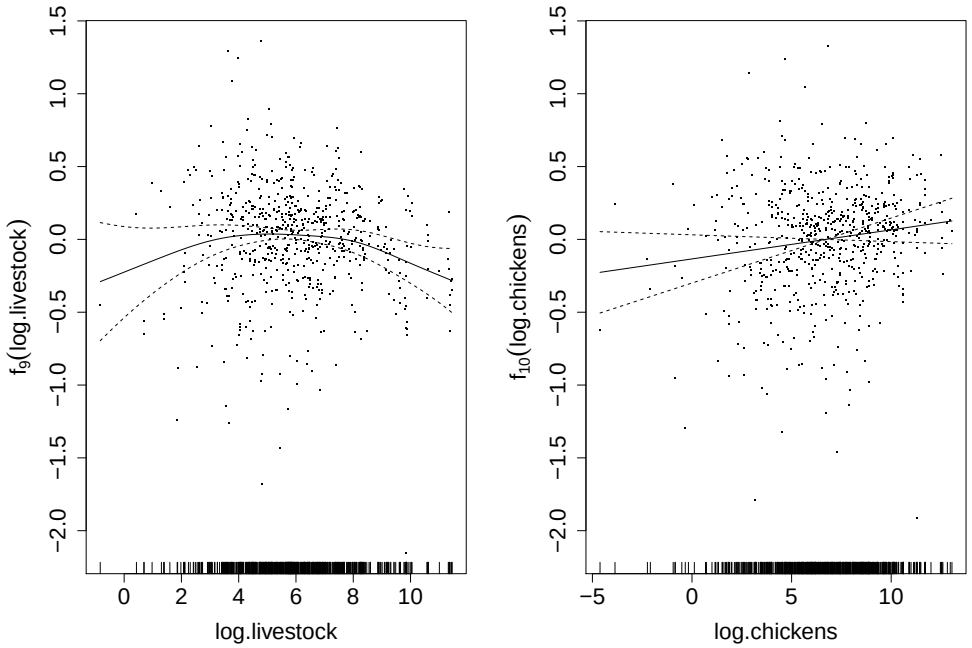


FIGURE 4.24: Plots of smooth functions from 4.10 with partial residuals (dots).

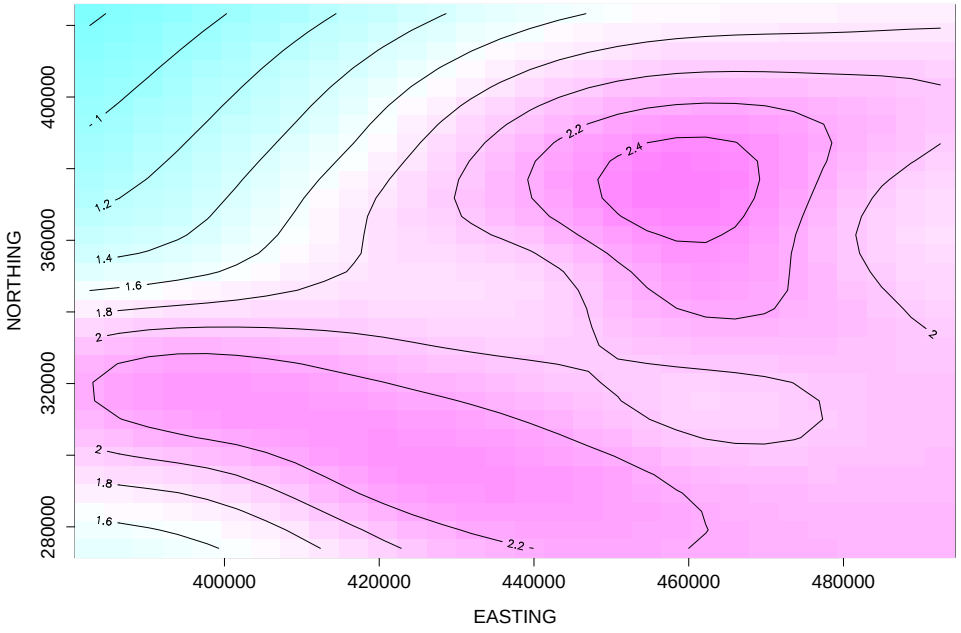


FIGURE 4.25: Plots of location term from 4.10. Blue indicates low values of winL and pink indicates higher values of winL.

A sensitivity analysis was carried out to ensure that automatic selection of number of basis functions when fitting the additive model using the `mgcv` package in `R` had chosen sensible smoothing parameters. Table 4.3 shows covariates$_{mxd}$ fitted with default values of k (the upper limit on the number of basis functions used to estimate the smooth function) and manually selected lower and higher values of k to give a range of k within which the model does not significantly change. The significance of the p-value does not change when k is increased or decreased and the *edf* do not vary much when k is changed, although *edf* for s(Easting, Northing) seems more variable than the other two smooth functions. R^2 also remains quite stable with varying values of k.

	Default R^2 : 60.7%			Low R^2 : 60.6%			High R^2 : 62.1%		
Smooth function	k	*edf*	p	k	*edf*	p	k	*edf*	p
s(Easting, Northing)	30	21·3	< 0·001	25	19·5	< 0·001	35	25·0	< 0·001
s(chickens)	10	1·0	0·11	7	1·0	0·12	13	1·0	0·12
s(livestock)	10	3·3	0·01	7	3·1	0·01	13	3·3	0·01

TABLE 4.3: Results of sensitivity analysis. k is the upper limit on the number of basis functions used to estimate the smooth function, *edf* is the effective degrees of freedom controlled by the degree of penalization selected by REML and p is the p-value for the smooth funtion from the model summary.

4.2.2 Modelling residual correlation

The residuals from (4.5), (4.6) and (4.7) and (4.9) were modelled using the `SSN` package in `R` to investigate the nature of any remaining spatial structure, after accounting for covariate information. Spatial correlation was modelled as in (4.11) where μ is the mean of the $m = 1, ..., 56$ residuals, η_{Euc_m} are correlated residuals where correlation is based on Euclidean distance, η_{TU_m} are correlated residuals where correlation is based on stream distance, η_{Euc+TU_m} are correlated residuals where correlation is based on a combination of stream and Euclidean distances and $\varepsilon_m \overset{\text{iid}}{\sim} N(0, \sigma^2)$ are independent errors. Euclidean distance based correlation (correlation$_{Euc}$) is modelled using the Gaussian Euclidean function and stream distance based correlation (correlation$_{TU}$) is modelled using the Epanechnikov Tail-up function, following the work presented in Chapter 3. Correlation based on a combination of Euclidean and stream distances as in 3.2.1 is referred to as correlation$_{hyb}$. There are 17 models to compare in total: 4 × covariate models (4.5), (4.6), (4.7) and (4.9) with independent residuals and the residuals from each of these have been modelled using the 3 correlation structures in (4.11). The SSN model with no

covariates and hybrid spatial correlation structure is also included to allow a comparison of a model with covariates and independent errors against a model with correlated errors but no covariates.

$$residuals_m = \mu + \left\{ \begin{array}{l} \eta_{Euc_m} \\ \eta_{TU_m} \\ \eta_{Euc+TU_m} \end{array} \right\} + \varepsilon_m \tag{4.11}$$

Covariance function parameter estimates (Table 4.4) and variance components (Table 4.5) were calculated for each of the 17 models. Inspection of these as well as boxplots of residuals, fitted values at observed locations, predicted values at unobserved locations and their associated standard errors makes it possible to investigate the effect of accounting for different types of spatial correlation in a variety of covariate models.

Figure 4.26 shows residuals from the four additive models (gam) along with residuals from modelling independent gam residuals with correlation$_{Euc}$ (Euc), correlation$_{TU}$ (Tail-up) and correlation$_{hyb}$ (hybrid). The residuals from the SSN model with no covariates and a hybrid spatial correlation structure (ssn.hybrid) is also shown as a comparison. It can be seen in the top left plot of Figure 4.26 that modelling the residuals from covariates$_{RCA}$ with a Euclidean spatial covariance function has little effect on the residuals but a variance ratio test shows that using the Tail-up correlation function leads to significant (p=0.01) reduction in the variability of the residuals. The top right plot shows that after accounting for accumulated covariates, modelling the residuals with correlation$_{Euc}$ leads to a significant reduction (p<0.001) in variability compared to assuming independence. Variability in the residuals from the covariates$_{acc}$ model is minimised using the correlation$_{hyb}$ function where the variance of the independent residuals is between 1.9 and 2.7 times greater than the variability of the residuals after accounting for correlation$_{hyb}$. The bottom left plot shows that variability of residuals from covariates$_{mxd}$ can be significantly (p=0.01) reduced by accounting for correlation$_{TU}$ and the bottom right plot shows that variability of residuals from covariates$_{loc}$ can also be significantly (p=0.01) reduced by accounting for correlation$_{TU}$.

Figure 4.26 shows that residual variance can be significantly reduced after accounting for covariate information by modelling the residuals with a suitable spatial correlation function. It can also be seen here that the ssn.hybrid model with no covariates gives residuals

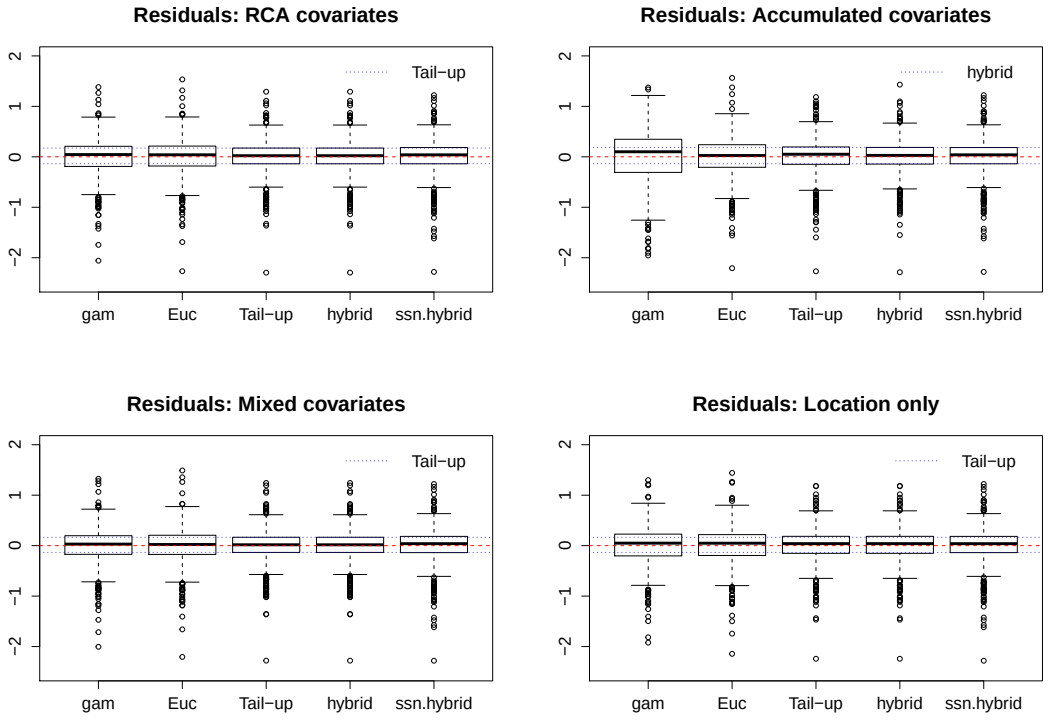


FIGURE 4.26: *gam* = residuals from additive models (covariates$_{RCA}$, covariates$_{acc}$, mixed covariates and location only), *Euc* = residuals after accounting for correlation based on Euclidean distance, *Tail-up* = residuals after accounting for correlation based on stream distance, *hybrid* = residuals after accounting for correlation based on Euclidean and stream distances, *ssn.hybrid* = residuals from the SSN model with hybrid covariance structure. Red dashed line is placed at 0. Blue dotted lines show the interquartile range for the residuals with the smallest IQR.

whose variance is of similar magnitude to models including covariate information.

Tables 4.4 and 4.5 can be used to further investigate the effect of modelling spatial correlation after accounting for covariate information. It can be seen by looking at the range parameter that in the Euc.RCA model there is correlation$_{Euc}$ only at small distances. In the Hyb.RCA model however the Euclidean range parameter is estimated to be greater than the largest pairwise distance between any 2 monitoring sites suggesting that once correlation$_{TU}$ is accounted for there is nothing left to be modelled by correlation$_{Euc}$. This can also be seen in the Euc.Mxd/Euc.Loc and Hyb.Mxd/Hyb.Loc models. In the .Acc models the Tail-up range parameter is greater than for the .RCA, .Mxd and .Loc models suggesting a tradeoff between accumulated covariates and the Tail-up correlation structure similar to but not as extreme as the tradeoff between RCA covariates and the Euclidean correlation structure. The Hyb.SSN model shows the covariance function parameters estimated for the SSN model with no covariates and correlation$_{hyb}$ and it can

be seen that the Tail-up partial sill and nugget are quite similar to the estimates from the models including covariate information. The Euclidean partial sill and Euclidean range are quite different from those estimated from models that include covariate information suggesting that spatial correlation is dominated by Euclidean distance. This can also be seen in the variance components in Table 4.5 where in the Hyb.SSN model 69% of the error variance is accounted for by the Euclidean component but in the covariate models with RCA covariates the Euclidean variance component is almost zero.

model	Tail-up sill	Tail-up range (km)	Euclidean sill	Euclidean range (km)	Nugget
Euc.RCA			0.02	**9**	0.12
TU.RCA	0.06	36			0.08
Hyb.RCA	0.06	36	0.0001	**667**	0.08
Euc.Acc			0.15	23	0.13
TU.Acc	0.18	55			0.08
Hyb.Acc	0.08	53	0.19	48	0.05
Euc.Mxd			0.02	**9**	0.12
TU.Mxd	0.06	39			0.08
Hyb.Mxd	0.06	39	0.0003	**671**	0.08
Euc.Loc			0.02	**9**	0.12
TU.Loc	0.07	35			0.08
Hyb.Loc	0.07	35	0.00003	**667**	0.08
Hyb.SSN	0.08	48	0.39	58	0.09

TABLE 4.4: Covariance function parameters for SSN models fitted to residuals from additive models. In the model column the first part of the model name refers to the correlation structure (Euc = Euclidean, TU = Tail-up, Hyb = hybrid) and the second part is the covariate model (RCA = covariates$_{RCA}$, Acc = covariates$_{acc}$, Mxd = covariates$_{mxd}$, Loc = covariates$_{loc}$, SSN = SSN model with no covariates).

Model	Tail-up	Euclidean	Nugget
Euc.RCA		0.12	0.88
TU.RCA	0.44		0.56
Hyb.RCA	0.439	**0.001**	0.56
Euc.Tot		0.55	0.45
TU.Tot	0.68		0.32
Hyb.Tot	0.23	0.54	0.23
Euc.Mxd		0.13	0.87
TU.Mxd	0.43		0.57
Hyb.Mxd	0.427	0.003	0.57
Euc.Loc		0.13	0.87
TU.Loc	0.44		0.56
Hyb.Loc	0.4371	0.0002	0.5627
Hyb.SSN	0.15	**0.69**	0.16

TABLE 4.5: Variance components for SSN models fitted to residuals from additive models.

Figure 4.27 shows predicted values of winL at observed locations for the 4 covariate models: (4.5), (4.6), 4.7 and 4.9 and the SSN model. It seems that although all of the models underestimate winL on average, covariates$_{acc}$ underestimates winL to a greater extent than the other models. The three covariate models that include RCA level covariates (4.5, 4.6 and 4.9) have predicted values similar to those from the SSN model, although the SSN model better captures the variability seen in the data (winL) than the other models. Figure 4.28 shows the standard errors for the predicted values at observed locations from the 4 covariate models and the SSN model. Covariates$_{acc}$ has the lowest standard errors but this is likely due to the lower variability seen in the predicted values in Figure 4.27. Standard errors are highest for the SSN model which is not unexpected since this model contains less information (in terms of covariates) than the other models.

Figure 4.29 shows predicted values of winL at unobserved locations. The covariates$_{RCA}$, covariates$_{mxd}$ and SSN models predict similar values on average with the covariates$_{mxd}$ model predicting slightly lower values on average and the covariates$_{acc}$ model predicting much lower values on average. The SSN model predicts values with a greater interquartile range than the other models but the covariates$_{RCA}$ and covariates$_{mxd}$ models predict more extreme low values than the other models. The covariates$_{acc}$ model predicts winL values with a much smaller spread than the other models, as was seen in Figure 4.27. Figure 4.30 shows the standard errors of winL predictions at unobserved locations. As with the standard error of predictions at observed locations in Figure 4.28, there is more uncertainty associated with predictions from the SSN model than the other models and standard errors are approximately 4 times greater in the SSN model than the other models.

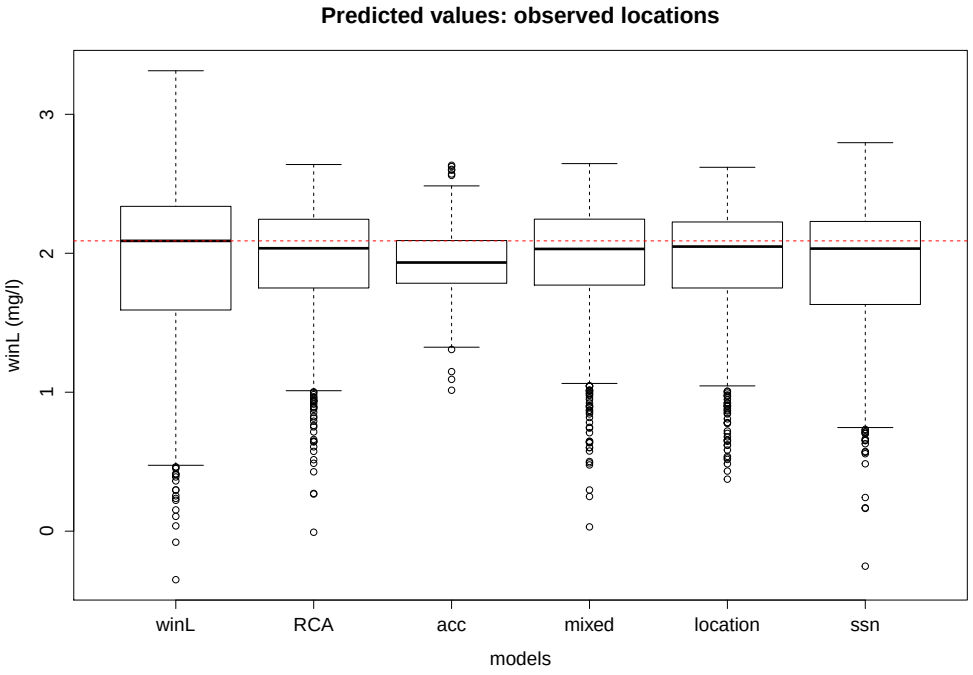


FIGURE 4.27: Predicted values at observed locations from covariate models and SSN model. Red dashed line is placed at median of winL = observed values.

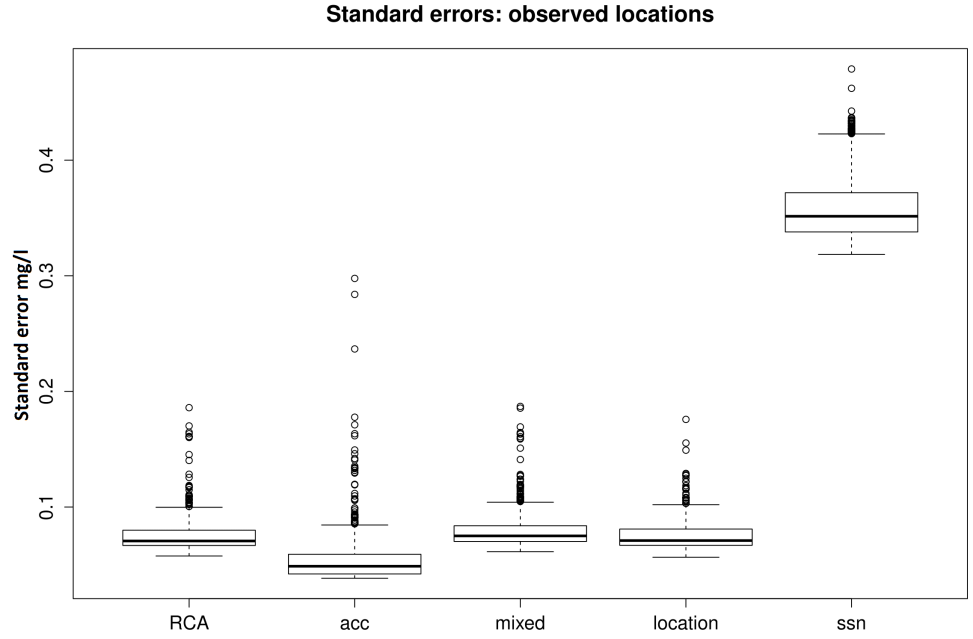


FIGURE 4.28: Standard errors of predicted values at observed locations from covariate models and SSN model.

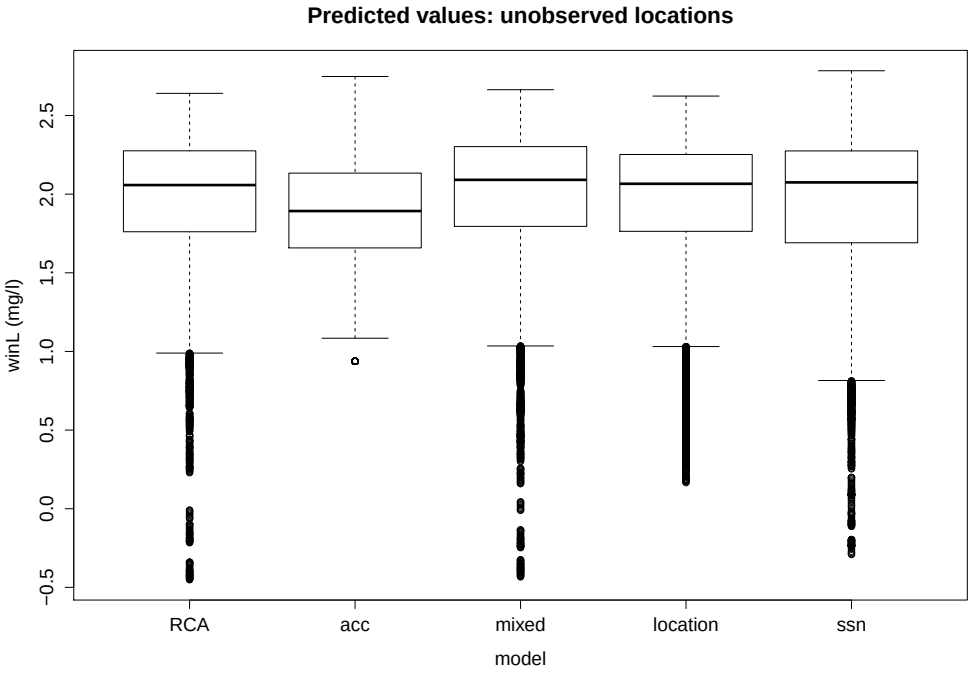


FIGURE 4.29: Predicted values at unobserved locations from covariate models and SSN model.

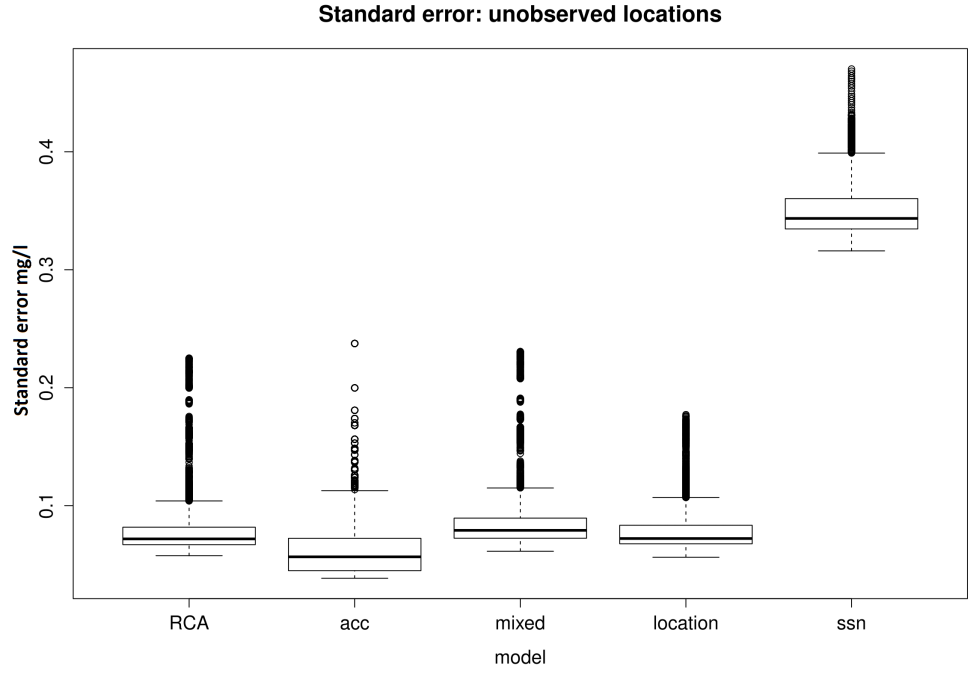


FIGURE 4.30: Standard errors of predicted values at observed locations from covariate models and SSN model.

4.2.3 Summary of covariate study

This covariate study has shown that the best subset of covariates to describe winL include a linear term for rain depth (mm), a shift in mean winL for land use category and smooth functions of location, accumulated livestock and accumulated chickens. This subset of covariates explains 60.7% of the variability of winL. A sensitivity analysis showed that this model was suitable for a range of k, the upper limit on number of basis functions used to model the covariate effect on the response.

The effect of the covariates considered in this study reflect other studies of nitrate levels in the literature. For example, the negative relationship between rainfall and nitrates was found in Pesce and Wunderlin (2000) and the positive relationship between agriculture or poultry farming and nitrates was also found in Neill (1989) and Gerber et al. (2007) respectively. The positive relationship between livestock and nitrates as discussed in Hooda et al. (2000) was not found in this study where a parabolic relationship was estimated.

Modelling correlation in the independent residuals from the covariate models showed that once RCA level covariates had been modelled a significant reduction in the variance of residuals could be achieved using a Tail-up (stream distance based) correlation function. Inspection of covariance function parameters and variance components showed that there is a tradeoff between RCA level covariates and the Euclidean correlation component as well as a weaker tradeoff between accumulated covariates and the Tail-up correlation component. A comparison between covariate models with independent residuals and the SSN model with no covariates and correlated residuals (where correlation was based on both Euclidean and stream distance) showed that predicted values at observed and unobserved locations were similar between the covariate and SSN models but the associated standard errors were around 4 times greater for the SSN model compared to the best fitting covariate model (covariates_{mxd}).

The covariate study provides possibilities for reducing the monitoring budget. For example, the data processing steps required to produce the hybrid spatial covariance structure (discussed in Chapter 3) are time intensive but this should be viewed as a one off cost since once any errors are removed from the shapefiles used to calculate stream distance and connectedness information, these can be stored and used as often as required. Covariate information on the other hand might require continuous monitoring over time,

adding to the costs of maintaining a monitoring network. Since models based only on the hybrid covariance structure, and no covariate information, provided predictions comparable to the model with several covariates, it might be possible for regulatory agencies to reduce dependence on covariate data for predicting nutrient levels at unsampled locations.

4.3 Comments

The simulation study shows that parameter estimates, predictions and their associated uncertainty are similar between the random and stratified sampling schemes, with the weighted sampling scheme often behaving differently from the others. This is likely because the subnetworks of all sizes selected under the weighted sampling scheme have a higher proportion of high values of winL and a lower proportion of low values of winL compared to subnetworks selected by random or stratified sampling schemes. It is recommended therefore that if a weighted sampling scheme is to be used then weighting variables should be chosen bearing in mind that this can strongly influence predictions at unobserved locations. There was little difference between random and stratified sampling schemes so it is recommended that unless there is a good reason for stratification, the random sampling scheme should be used due to the simple implementation of this scheme.

Going forward, it has been shown in this Chapter that the monitoring network in the Trent catchment area could be reduced by up to 50% while having little impact on predicted values and their associated uncertainty. The Trent area is a densely monitored area with 687 monitoring sites and so removing 50% of monitoring sites would mean 344 sites would be retained thus the Trent would still have a very dense monitoring network compared to other areas of England and Wales. Regulatory agencies could reap the benefit of reduced monitoring costs while maintaining confidence in predictions made at unobserved location using statistical models based on a complex spatial covariance structure.

The covariate study considered several variables known to influence water quality and land use category provided the most information for a single covariate. Models containing covariates and assuming an independent error structure were compared to models

with no covariates and a hybrid covariance structure (based on the conclusions in Chapter 3). It was shown that these two models were comparable in terms of predicted values at unobserved locations although the standard error of predictions was higher for the non-covariate model. It was also shown that there is value in modelling stream distance based spatial correlation once covariates have been accounted for but the inclusion of RCA level covariates almost cancels out Euclidean distance based spatial correlation. It can be concluded that there is value in modelling stream distance based correlation even when covariate information is available, but if no covariate information is available then good predictions can still be made if a more complex hybrid correlation function is used in the model, but with higher uncertainty.

Chapter 5

Estimating spatial and temporal patterns

Regulatory agencies will often be interested in both spatial and temporal patterns in the data. For example, dominant spatial patterns across time can be used to identify large scale changes in log(TON) across a river network. T-mode principal components analysis (PCA, introduced in Chapter 1) aims to find spatial patterns in the data and assess at what time points these spatial patterns occur. The presence of more than one dominant spatial pattern suggests a change in the spatial pattern over time. Alternatively, S-mode PCA aims to estimate dominant temporal patterns in the data and provides an indication of which sites, possibly grouped together, behave similarly over time. If common temporal patterns can be identified then this suggests redundancy in the monitoring network and strengthens the argument to reduced the number of monitoring sites, leading to cost savings for regulatory agencies. Adjusting statistical methods for known spatial relationships related to the structure of the river network was shown to increase the predictive capability of statistical models in Chapter 3 without a corresponding increase in monitoring sites or the inclusion of covariate data which might be difficult and expensive to collect. Incorporating information about the shape of the river network into statistical methodology might therefore improve the information extracted from statistical models without increasing monitoring costs. In this chapter, a novel development of PCA is proposed to take into account known structure in the data by applying row and column weights to the spatiotemporal data matrix. PCA was selected as it provides the advantages of dimension reduction, where a complicated dataset can be represented

as a few dominant spatial or temporal patterns and is computationally suitable for large datasets. PCA is therefore preferable to dynamic factor analysis, applied in Chapter 2, which became computationally inefficient as the number of monitoring sites and time points increased.

PCA can be used to evaluate dominant patterns in spatiotemporal datasets, and is especially useful for large data sets where the number of variables exceeds the number of observations such as are often found in genomics, imaging and climatology data sets (discussed in Chapter 1). PCA can be adapted for incomplete datasets or applied after imputation. PCA can also be run efficiently for very large datasets using iterative techniques described in Chapter 1. For these reasons PCA is a suitable statistical technique for assessing spatial and temporal patterns in environmental datasets. PCA makes use of correlation among variables to find structure in the data but does not explicitly make use of known structure, which in an environmental context could be spatial or temporal structure. This chapter will show how known structure can be incorporated within PCA.

First, the PCA methodology is adapted to introduce row and column weights, followed by the description of an asymmetric weight matrix reflecting river network topology. Next, the row and column weighted PCA is applied to data from the Trent catchment area (described in detail in Chapters 3 and 4) in both T- and S-mode. Finally, comparisons are made between the methodology developed in this chapter and existing work on weighted PCA. This chapter focusses particularly on the inclusion of an asymmetric weight matrix within PCA.

5.1 Adjusting PCA for known structure

This section will discuss various attempts in the literature to incorporate weights within PCA methodology and in particular, attention is given to methods proposed to take into account spatial or temporal structure in the data. Next some notation will be introduced to describe PCA methodology and adjustments to the standard methodology are proposed to take into account spatial and temporal structure. Following this, Section 5.1.2 describes the steps to follow to construct spatial weights reflecting flow direction and strength of relationship between monitoring sites on a river network.

There are examples in the literature of incorporating general weights into PCA to reflect known structure in the data. An early example can be found in Gabriel and Zamir (1979) who develop a low rank approximation of matrices using weighted least squares for any choice of weights, similar to the NIPALS iterative algorithm (Wold and Lyttkens, 1969) but introduce an initialization procedure of the algorithm to prevent convergence to a local rather than global minimum. a problem they found to occur in the original algorithm proposed by Wold and Lyttkens (1969). Tamuz et al. (2005) develop a similar algorithm to remove known linear systematic effects from photometric light curves with heterogeneous errors and note that this method is most suitable for data with high signal to noise ratio and highly variable errors. Pinto da Costa et al. (2011) apply a weighted PCA to microarray data where ranks are used rather than the raw data. PCA is then performed on a rank based correlation matrix rather than the standard Pearson correlation coefficient and the method is shown to be robust to outliers. Baldwin et al. (2009) describe a general weighting scheme to account for known structure in S-mode PCA and apply weighted PCA using a diagonal weight matrix while Allen et al. (2014) discuss a generalized matrix decomposition where any symmetric weight matrix can be incorporated into PCA using a weighted singular value decomposition. This generalized decomposition is also briefly mentioned in Abdi and Williams (2010) although no development of the methodology is given. In general, weights are incorporated into PCA to account for known structure in the data, to emphasise or de-emphasise the importance of particular variables and to make the method robust against outliers. The methods described in Baldwin et al. (2009) and Allen et al. (2014) are used in this Chapter to develop PCA methodology adjusted for spatial weights reflecting flow direction and strength of connectedness in a river network. These methods are adapted since they specifically consider spatiotemporal data, and Baldwin et al. (2009) in particular discuss S-mode PCA.

Weighting PCA to incorporate known spatial structure relies on the specification of weights showing how n spatial locations are related to each other. An $n \times n$ binary weight matrix has a 1 where two locations are 'neighbours' and a 0 if they are not. In the case of areal data two spatial units are neighbours if they share a border. For point data a location could have neighbours defined as all other locations within a specified distance or the ν nearest locations where ν must be specified in advance. The neighbourhood matrix can be non-binary if, for example, the matrix is standardised by

making each row sum to 1. If distance between locations is thought to be important then inverse distance might be used to downweight the influence of distant locations. The weight matrix can be symmetric or asymmetric if direction should be accounted for.

Attempts have been made to incorporate spatial information into PCA by combining PCA with Moran's I through the use of a neighbourhood matrix. See for example Wartenberg (1985), Thioulouse et al. (1995), Jombart et al. (2008) and Dray et al. (2008). Jombart et al. (2008) discuss in detail how spatial weights based on Moran's I can result in negative eigenvalues and interpret these as examples of local rather than global variability in the same way that negative values of Moran's I reflect local structure. In the context of genetics studies local structure is interpreted as repulsion where individuals with the same genetic structure deliberately repel each other. In the context of river networks it might be the case that negative eigenvalues could be interpreted similarly if the determinand of interest is species counts or presence/absence data but negative eigenvalues cannot be interpreted this way when stream chemistry variables are of interest. This combination of Moran's I with PCA is useful if the aim is to estimate principal components that are smooth in space. Frichot et al. (2012) take a different approach and use an inverse spatial correlation matrix as weights to remove spatial correlation. As a result, interesting spatial features originally masked by a smooth spatial pattern are clearly identified.

Harris et al. (2011) and Harris et al. (2015) introduce geographically weighted PCA where PCA is carried out at each areal unit (or monitoring site) separately based on its neighbours and accounts for spatial heterogeneity rather than autocorrelation which is explored using the PCA techniques based on Moran's I. Methods are developed to choose the number of neighbours and interpretation focuses on the proportion of variance explained by the first component at each site and which variables contribute most to the first component at each site. Cheng et al. (2011) discusses fuzzy masking PCA where a function of distance is used as weights to constrain PCA of image data to focus on geographical areas with particular geology. Loadings are calculated based on these samples but scores are calculated for all samples to try to identify areas of similar geological properties to those used to calculate the loadings but have not yet been identified as being similar. The idea here is to increase signal to noise ratio by reducing the influence of pixels that are not of interest. Guo et al. (2015) introduce two types of

spatial weights and aim to minimise the weighted PCA reconstruction error (rather than maximising the variance of the scores) for imaging data which is often high dimensional but with few observations. Local weights are used to incorporate spatial smoothness while global weights allow for the selective treatment of features of interest.

PCA can also be adjusted for temporal structure in the data. For example, Ku et al. (1995) develop dynamic PCA for statistical process control applications with temporally autocorrelated data and augment the data matrix with lagged variables. This is similar to singular spectrum analysis (SSA) and its multivariate equivalent (MSSA) which are used to incorporate temporal autocorrelation within PCA and an introduction can be found in Jolliffe and Uddin (2000). SSA and MSSA appear to be suitable only for analysing observations collected over time at a single monitoring site. Stahlschmidt et al. (2015) adapt PCA to multivariate spatio-temporal data where measurements are made on multiple variables at several time points and locations in space. They use a time average of the spatial covariance matrix and apply an eigendecomposition to this average to maximise the product of the variance of the scores and spatial autocorrelation but do not make it clear how the spatial weights are constructed. Skočaj et al. (2007) develop a weighted PCA for spatio-temporal image data and use the EM algorithm (described in Chapter 1) to minimise the weighted reconstruction error, in a similar manner to Allen et al. (2014). They also use temporal smoothing to improve data reconstruction from PCA based on incomplete data.

In this thesis, Chapter 3 discussed the Tail-up model (ver Hoef et al., 2006) and how a symmetric matrix of spatial weights based on features of a river network can be used to weight a stream distance based spatial covariance matrix. This Chapter proposes incorporating an asymmetric matrix of spatial weights reflecting river network structure into PCA methodology, with the aim of uncovering interesting spatial features masked by smooth transition along the river network, following the ideas in Frichot et al. (2012). First, methodology and notation are developed to describe how row and column weights can be incorporated within PCA methodology to account for relationships among observations and variables, respectively. Next, an asymmetric matrix of spatial weights is developed, based on the work in ver Hoef et al. (2006). Following this, the novel PCA adaptation is applied to data from the Trent catchment area and finally comparisons are made between the approach developed in this Chapter and existing weighted PCA methods, specifically those described in Allen et al. (2014).

5.1.1 Incorporating row and column weights into PCA

This section will propose a method for incorporating row and column weights reflecting structure among the observations or variables in a data matrix, and show how principal components and their loadings must be transformed so that they are on the correct scale.

Principal components analysis aims to successively maximise

$$\frac{1}{n-1}\mathbf{V}^\top\mathbf{X}^\top\mathbf{X}\mathbf{V},$$

which is the variance of k principal components $\mathbf{XV}$ calculated for $n \times p$ data matrix $\mathbf{X}$ and $p \times k$ loadings matrix $\mathbf{V}$ (note that the constant $\frac{1}{n-1}$ will be dropped from now on to simplify notation). Allen et al. (2014) show that if PCA is adjusted using $p \times p$ column weights matrix $\mathbf{\Omega}$ then this means that PCA is applied to the covariance matrix of

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{\Omega}$$

and the principal components $\mathbf{Z}$ are calculated as

$$\begin{aligned}\mathbf{Z} &= \mathbf{X}\mathbf{\Omega}\tilde{\mathbf{V}} \\ &= \tilde{\mathbf{X}}\tilde{\mathbf{V}}\end{aligned}$$

where $\tilde{\mathbf{V}}$ are the loadings from the decomposition of $\tilde{\mathrm{X}}$. In order to make comparisons between unweighted PCA (PCA_{uw}) and column weighted PCA_c the reconstruction error from PCA_c must be transformed so that the reconstruction error is relative to $\mathbf{X}$ rather than $\tilde{\mathbf{X}}$. Assume $\hat{\mathbf{X}}$ = data reconstructed from k retained principal components and $\mathbf{V}$ = $p \times k$ matrix containing the corresponding loadings and $\mathbf{X} = \hat{\mathbf{X}} + error$. The reconstruction error for PCA_{uw} can be defined as $\mathbf{X}\mathbf{V}_{k+1:p}\mathbf{V}_{k+1:p}^\top$ since

$$\begin{aligned}\mathbf{X} &= \mathbf{X}\mathbf{V}_{1:k}\mathbf{V}_{1:k}^\top \\ &+ \mathbf{X}\mathbf{V}_{k+1:p}\mathbf{V}_{k+1:p}^\top\end{aligned}$$

where $\mathbf{V}_{1:k}$ indicates the first k columns of the matrix containing loadings and $\mathbf{V}_{k+1:p}$ indicates the last $p-k$ columns of the loadings matrix. The reconstruction here refers to reconstructing the centered data. In order to fully reconstruct the data the column means would have to be added to $\hat{\mathbf{X}}$ but this has been omitted for simplicity. Since the column means are constants, they do not affect the sum of squared differences between $\mathbf{X}$ and $\hat{\mathbf{X}}$. The reconstruction error is contained in the $n \times p$ matrix $\mathbf{X}\mathbf{V}_{k+1:p}\mathbf{V}_{k+1:p}^{\top}$.

For PCA_c the reconstruction error can be defined as the second term on the right of (5.1),

$$\begin{aligned} \mathbf{X}\boldsymbol{\Omega} &= \mathbf{X}\boldsymbol{\Omega}\tilde{\mathbf{V}}_{1:k}\tilde{\mathbf{V}}_{1:k}^{\top} \\ &+ \mathbf{X}\boldsymbol{\Omega}\tilde{\mathbf{V}}_{k+1:p}\tilde{\mathbf{V}}_{k+1:p}^{\top} \end{aligned} \tag{5.1}$$

where the reconstruction error $\mathbf{X}\boldsymbol{\Omega}\tilde{\mathbf{V}}_{k+1:p}\tilde{\mathbf{V}}_{k+1:p}^{\top}$ is defined in terms of $\mathbf{X}\boldsymbol{\Omega}$. Post multiplying all terms in (5.1) by $\boldsymbol{\Omega}^{-1}$ gives

$$\begin{aligned} \mathbf{X}\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1} &= \mathbf{X}\boldsymbol{\Omega}\tilde{\mathbf{V}}_{1:k}\tilde{\mathbf{V}}_{1:k}^{\top}\boldsymbol{\Omega}^{-1} \\ &+ \mathbf{X}\boldsymbol{\Omega}\tilde{\mathbf{V}}_{k+1:p}\tilde{\mathbf{V}}_{k+1:p}^{\top}\boldsymbol{\Omega}^{-1} \\ \text{i.e.} & \\ \mathbf{X} &= \mathbf{X}\boldsymbol{\Omega}\tilde{\mathbf{V}}_{1:k}\tilde{\mathbf{V}}_{1:k}^{\top}\boldsymbol{\Omega}^{-1} \\ &+ \mathbf{X}\boldsymbol{\Omega}\tilde{\mathbf{V}}_{k+1:p}\tilde{\mathbf{V}}_{k+1:p}^{\top}\boldsymbol{\Omega}^{-1} \end{aligned} \tag{5.2}$$

The reconstruction error of PCA_c relative to X is now $\mathbf{X}\boldsymbol{\Omega}\tilde{\mathbf{V}}_{k+1:p}\tilde{\mathbf{V}}_{k+1:p}^{\top}\boldsymbol{\Omega}^{-1}$. Equation (5.2) also shows that the data are reconstructed as the product of $\tilde{\mathbf{X}}$ with transformed loadings $\tilde{\mathbf{V}}^{\top}\boldsymbol{\Omega}^{-1}$ or alternatively $\boldsymbol{\Omega}^{-1\top}\tilde{\mathbf{V}}$. It is also important to notice that the loadings in Equation (5.2) require transformation by the *transpose* inverse of the weight matrix used to calculate $\tilde{\mathbf{X}}$ to calculate the loadings relative to $\mathbf{X}$ rather than $\tilde{\mathbf{X}}$. These transformed loadings are called the solution to the generalized PCA problem in Allen et al. (2014). Baldwin et al. (2009) show that if PCA is adjusted for column weights so that $\tilde{\mathbf{X}} = \mathbf{X}\boldsymbol{\Omega}$ then the loadings $\mathbf{V}$ are $\boldsymbol{\Omega}^{-1}\tilde{\mathbf{V}}$ where $\boldsymbol{\Omega}$ is the matrix of column weights and $\tilde{\mathbf{V}}$ are the loadings calculated from the decomposition of $\tilde{\mathbf{X}}$. The column weight matrix applied in Baldwin et al. (2009) however is symmetric (in fact it is diagonal) so $\boldsymbol{\Omega}^{-1\top} = \boldsymbol{\Omega}^{-1}$ and it is not clear therefore that the loadings must be transformed using the transpose of the inverse column weight matrix.

Similarly, an $n \times n$ matrix $\mathbf{\Phi}$ of row weights can be defined and row weighted PCA, PCA_{rw} is applied to

$$\tilde{\mathbf{X}} = \mathbf{\Phi}\mathbf{X}.$$

$\mathbf{X}$ can be reconstructed as

$$\begin{aligned}
\mathbf{\Phi}\mathbf{X} &= \mathbf{\Phi}\mathbf{X}\tilde{\mathbf{V}}_{1:k}\tilde{\mathbf{V}}_{1:k}^{\top} \\
&+ \mathbf{\Phi}\mathbf{X}\tilde{\mathbf{V}}_{k+1:p}\tilde{\mathbf{V}}_{k+1:p}^{\top} \\
\text{i.e.} & \\
\mathbf{X} &= \mathbf{X}\tilde{\mathbf{V}}_{1:k}\tilde{\mathbf{V}}_{1:k}^{\top} \\
&+ \mathbf{X}\tilde{\mathbf{V}}_{k+1:p}\tilde{\mathbf{V}}_{k+1:p}^{\top}
\end{aligned}$$

The principal components $\mathbf{Z}$ require a back transformation as did the loadings for column weighted PCA giving

$$\begin{aligned}
\mathbf{Z} &= \mathbf{\Phi}^{-1}\tilde{\mathbf{X}}\tilde{\mathbf{V}} \\
&= \mathbf{\Phi}^{-1}\mathbf{\Phi}\mathbf{X}\tilde{\mathbf{V}} \\
&= \mathbf{X}\tilde{\mathbf{V}}.
\end{aligned}$$

Accounting for both row and column weights results in PCA_{rc} where the reconstruction error can be derived as

$$\begin{aligned}
\mathbf{\Phi}\mathbf{X}\mathbf{\Omega} &= \mathbf{\Phi}\mathbf{X}\mathbf{\Omega}\tilde{\mathbf{V}}_{1:k}\tilde{\mathbf{V}}_{1:k}^{\top} \\
&+ \mathbf{\Phi}\mathbf{X}\mathbf{\Omega}\tilde{\mathbf{V}}_{k+1:p}\tilde{\mathbf{V}}_{k+1:p}^{\top}
\end{aligned} \tag{5.3}$$

Pre multiplication of the terms in (5.3) by $\mathbf{\Phi}^{-1}$ and post multiplication by $\mathbf{\Omega}^{-1}$ gives

$$\begin{aligned}
\mathbf{\Phi}^{-1}\mathbf{\Phi}\mathbf{X}\mathbf{\Omega}\mathbf{\Omega}^{-1} &= \mathbf{\Phi}^{-1}\mathbf{\Phi}\mathbf{X}\mathbf{\Omega}\tilde{\mathbf{V}}_{1:k}\tilde{\mathbf{V}}_{1:k}^{\top}\mathbf{\Omega}^{-1} \\
&+ \mathbf{\Phi}^{-1}\mathbf{\Phi}\mathbf{X}\mathbf{\Omega}\tilde{\mathbf{V}}_{k+1:p}\tilde{\mathbf{V}}_{k+1:p}^{\top}\mathbf{\Omega}^{-1}
\end{aligned}$$

$$\begin{aligned}
\mathbf{X} &= \mathbf{X}\mathbf{\Omega}\tilde{\mathbf{V}}_{1:k}\tilde{\mathbf{V}}_{1:k}^{\top}\mathbf{\Omega}^{-1} \\
&+ \mathbf{X}\mathbf{\Omega}\tilde{\mathbf{V}}_{k+1:p}\tilde{\mathbf{V}}_{k+1:p}^{\top}\mathbf{\Omega}^{-1}
\end{aligned}$$

It has been shown here how to calculate the principal components and reconstruction error for PCA adjusted for row and column weights.

5.1.2 Defining spatial weights for river networks

Section 1.4.2 highlighted some examples of PCA applied to stream chemistry data recorded on river networks but none of these examples take into account the river network structure investigated in depth in the Chapters 3 and 4. It is intended here to define a weight matrix describing the flow direction and connectedness structure of a river network that can be incorporated into PCA methodology as either row or column weights, depending on the PCA mode of interest. Chapter 3 described an additive function based on area of land that drains to a stream segment and showed how this can be used to form spatial weights reflecting the influence of upstream sites on a downstream monitoring site. For the purposes of a weighted PCA the asymmetric weight matrix should be calculated as in Peterson et al. (2007) but not forced to symmetry. A non-zero value in the weight matrix means that two sites are flow connected and the magnitude of the value indicates the strength of influence of the upstream site on the downstream site. A zero means that two monitoring sites are not flow connected.

A simple example will now be given to illustrate how the asymmetric weight matrix is used to create weighted data (for singular value decomposition) or a symmetric covariance matrix (for eigen decomposition). Figure 5.1 shows a simple river network with three stream segments and three monitoring sites as well as the proportional influence (PI) of sites where $\text{PI} \in [0, 1]$.

The matrix of spatial weights $\mathbf{\Omega}$ shown in (5.5) is the element-wise square root of the proportional influence matrix (5.4).

$$\text{PI} = \begin{pmatrix} 1 & 0 & a \\ 0 & 1 & b \\ 0 & 0 & 1 \end{pmatrix} \qquad (5.4) \qquad\qquad \mathbf{\Omega} = \begin{pmatrix} 1 & 0 & a^{\frac{1}{2}} \\ 0 & 1 & b^{\frac{1}{2}} \\ 0 & 0 & 1 \end{pmatrix} \qquad (5.5)$$

The data matrix $\mathbf{X}$ for the river network in Figure 5.1, assuming there are observations at two time points is

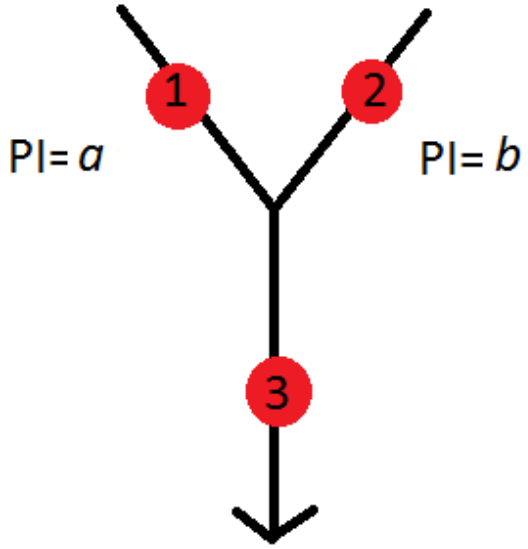


FIGURE 5.1: Diagram of a simple river network with three monitoring sites (red circles). Arrow represents direction of flow. PI=proportional influence.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \end{bmatrix},$$

where each column corresponds to a different monitoring site. The weighted data matrix is

$$\tilde{\mathbf{X}} = \mathbf{X}\boldsymbol{\Omega}$$

and the symmetric weighted covariance matrix is

$$\operatorname{cov}\left(\tilde{\mathbf{X}}\right) = \frac{1}{n-1}(\mathbf{X}\boldsymbol{\Omega})^{\top}\mathbf{X}\boldsymbol{\Omega}.$$

Written out in full (with $\frac{1}{n-1}$ omitted) this gives (5.6). The diagonal elements show that the variance at the most downstream site in Figure 5.1, site 3, is the sum of the variance at this site and the weighted variances at flow connected sites upstream from site 3.

$$\begin{pmatrix} Var(S_1) & Cov(S_1,S_2) & a^{\frac{1}{2}}Var(S_1) \\ & & + \; b^{\frac{1}{2}}Cov(S_1,S_2) \\ & & + \; Cov(S_1,S_3) \\ \\ Cov(S_1,S_2) & Var(S_2) & a^{\frac{1}{2}}Cov(S_1,S_2) \\ & & + \; b^{\frac{1}{2}}Var(S_2) \\ & & + \; Cov(S_2,S_3) \\ \\ a^{\frac{1}{2}}Var(S_1) & a^{\frac{1}{2}}Cov(S_1,S_2) & a^{\frac{1}{4}}Var(S_1) \\ + \; b^{\frac{1}{2}}Cov(S_1,S_2) & + \; b^{\frac{1}{2}}Var(S_2) & + \; b^{\frac{1}{4}}Var(S_2) \\ + \; Cov(S_1,S_3) & + \; Cov(S_2,S_3) & + \; Var(S_3) \\ & & + \; 2*b^{\frac{1}{2}}a^{\frac{1}{2}}Cov(S_1,S_2) \\ & & + \; 2*a^{\frac{1}{2}}Cov(S_1,S_3) \\ & & + \; 2*b^{\frac{1}{2}}Cov(S_2,S_3) \end{pmatrix} \tag{5.6}$$

Flow direction can only be represented correctly using an asymmetric weight matrix and data matrix $\mathbf{X}$ of the correct orientation. For S-mode PCA $\mathbf{X}$ is arranged so that each column represents a monitoring site and each row represents the ordered time points. $\mathbf{\Omega}$ must be constructed so that rows represent upstream sites and columns represent downstream sites. This preserves the flow direction of the river network so that water flows *from* rows *to* columns. For T-mode PCA the rows in $\mathbf{X}$ are monitoring sites and the columns are time points, the transpose of the data matrix used for S-mode PCA. The asymmetric weight matrix used in S-mode PCA must also be transposed so that row weights matrix $\mathbf{\Phi}$ has upstream sites in the columns and downstream sites in the rows so that water flows *from* columns *to* rows. A symmetric weight matrix such as that used in Peterson et al. (2007) would result in the variance at a single site being a linear combination of the variances at all connected sites in upstream and downstream directions and it does not make sense that the variance at a monitoring site would be affected by the variance at sites downstream.

This form of weight matrix reflecting flow direction and connectedness in a river network is a novel modification to the PCA methodology since no examples have been found in the literature describing the need for an asymmetric weight matrix of specific orientation to account for river network topology within the PCA framework.

5.2 Application to the Trent catchment area

This section will demonstrate the application of the proposed adaptation to PCA to data from the Trent catchment area. First, T-mode PCA is applied and adjusted for spatial correlation among observations. Next, S-mode PCA is applied and adjusted for spatial correlation among the variables and temporal correlation among the observations. The T-mode application is intended to be a simple example focusing on spatial correlation, while the S-mode example is intended to demonstrate how the proposed adaptation of PCA can be applied to a more complex data set with both spatial and temporal correlation.

5.2.1 T-mode PCA

Chapter 1 explained that PCA must be applied to a complete dataset with no missing values. In order to apply T-mode PCA to data from the Trent catchment area a subset of the 566 sites in the dominant network of the Trent was selected to provide a complete dataset. The response of interest is winter log(TON), investigated in detail in Chapter 3 and Chapter 4. Howden et al. (2011) suggests a minimum of 12 years of data are required to estimate a true trend that is not confounded with short term variability but Withers and Nadarajah (2015) showed that if data are aggregated to annual means then a minimum of five years of data are required to achieve only a 2% loss in efficiency of the trend estimator. Although it is not intended here to estimate the slope of a linear trend, these guidelines provide sensible suggestions for choosing the length of complete time series required for inclusion in the subset. This resulted in 481 monitoring sites being selected to have 13 consecutive and complete annual winter log(TON) values for the years 1995-2007. The columns of the 481×13 data matrix were centered by subtracting the column means and PCA was performed using singular value decomposition of the data matrix (equivalent to eigen decomposition of the covariance matrix) since all variables are recorded in the same units and therefore are on the same scale.

The diagonal elements of the 13×13 covariance matrix represent variance in log(TON) at 13 time points. The asymmetric row weights matrix $\mathbf{\Phi}$ was created for the 481 monitoring sites so that streams were represented as flowing from columns to rows. Wartenberg (1985) notes in the Appendix that a generalized PCA can incorporate correlation in

a similar way to generalized least squares (GLS) by using the inverse of the temporal correlation matrix to weight the covariance matrix upon which PCA is applied. This is also discussed at greater length in Allen et al. (2014) and the idea is that it is appropriate to apply standard PCA to the weighted data, followed by a transformation of the principal components or loadings, depending on whether row or column weights are applied, respectively. As a result of this the inverse of the spatial weights matrix is used here to weight the observations. In fact, it is the matrix square root of the inverse of the spatial weights matrix that is used, following the recommendations in Baldwin et al. (2009). This means that PCA is applied using the singular value decomposition on

$$\tilde{\mathbf{X}} = \mathbf{\Phi X}. \tag{5.7}$$

First, an unweighted T-mode PCA was performed and this was followed by a spatially weighted T-mode PCA. Results for both analyses are given followed by a discussion comparing the results from unweighted and spatially weighted PCA.

The first two columns in Table 5.1 show the loadings for the the first two principal components (PC's) from unweighted T-mode PCA. The first two PC's account for 89% and 3% of the variance in the data respectively. The loadings for the first component are all of the same sign and of similar magnitude and therefore this PC represents the average spatial pattern over all years. The second component represents a contrast between 1995-1997 and the years 2001-2001, 2005-2007 so 3% of the variability in the data is related to the size of this difference. Since the second principal component accounts for only 3% of the variance in the data however it can be concluded that one principal component is sufficient to describe the spatial pattern of winter log(TON) in the Trent catchment area.

Next, T-mode PCA was adjusted for river network structure as in (5.7). The first two components now account for 85% and 4% of the variance respectively and so two components are now required to explain the same proportion of variance as the first unweighted principal component. The loadings for the first two flow weighted components are given in the last two columns of Table 5.1. There are some small differences in the magnitude between unweighted and flow weighted loadings, but the interpretation of the loadings

Year	UW_1	UW_2	FW_1	FW_2
1995	-0.27	**-0.25**	-0.27	**-0.31**
1996	-0.24	**-0.57**	-0.25	**-0.58**
1997	-0.28	**-0.48**	-0.27	**-0.46**
1998	-0.29	-0.16	-0.29	-0.13
1999	-0.28	-0.11	-0.28	-0.08
2000	-0.30	**0.07**	-0.30	0.08
2001	-0.28	**0.22**	-0.28	**0.23**
2002	-0.26	0.26	-0.26	**0.27**
2003	-0.26	0.09	-0.26	0.07
2004	-0.29	0.08	-0.29	0.12
2005	-0.28	**0.32**	-0.29	**0.27**
2006	-0.29	**0.25**	-0.29	**0.20**
2007	-0.26	**0.21**	-0.26	**0.26**

TABLE 5.1: Loadings for unweighted (UW) and flow weighted (FW) T-mode PCA.

is the same as for the unweighted T-mode PCA: the first flow weighted principal component reflects variance around the mean spatial pattern over all years and the second flow weighted component is a contrast between early and later years.

Figure 5.2 shows the biplot for the first two unweighted principal components (top left) and a closer view of the principal component scores (top right). Sites (blue numbers) in the same quadrant as arrow heads have high log(TON) in the years indicated by the red arrow labels so sites in the top and bottom right quadrants have low log(TON) in all years while sites in the top and bottom left quadrants have high log(TON) in all years. The arrows representing the loadings show that the first component is an average of the spatial patterns for all years while the second component is a contrast between 1995-1999 and 2000-2007. The bottom left plot in Figure 5.2 shows the biplot for the first two flow weighted principal components. The years 2000-2007 are a little more closely grouped together than for the unweighted PCA and 1995 is more similar to 1996/1997 than 1998/1999 for the unweighted PCA. Inspection of the principal component scores in the bottom right of Figure 5.2 shows that there is almost no difference between the scores for the unweighted and flow weighted principal components. In fact, Figure 5.3 shows that differences in the principal component scores between unweighted and flow weighted PCA are most evident for the principal components that explain the smallest proportion of the variance in the data and these correspond to the error structure in the data. Since the row weights have been constructed to adjust for structure in the errors

it is reasonable that the biggest differences in the structure of the principal components and their loadings are found here.

Figures 5.4(a) and 5.4(b) show the principal component scores for the first two unweighted components. As with the biplots, a large negative score indicates high values of log(TON) and a large positive score indicates low values of log(TON) in the groups of years indicated by the corresponding loadings. The spatial pattern for the first unweighted component looks to be identical to the spatial pattern for the first flow weighted component and the pattern of scores for both of the second components are also very similar. The biggest difference in scores between the unweighted and spatially weighted analyses can be found in the components with the smallest contribution to toal variance. The pattern on the scores maps for the first components resembles the pattern of land use categories described in Chapter 4 where log(TON) is lower in urban areas and higher in agricultural areas.

Data were reconstructed from unweighted and flow weighted PCA using the first principal component from each analysis. Reconstruction error was calculated as the sum of squared differences between the centered data $\mathbf{X}$ and the reconstructed centered data $\hat{\mathbf{X}}$ using

$$\mathrm{Tr}\left[(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^{\top}\right], \tag{5.8}$$

where Tr[] is the trace of the matrix. The reconstruction error for T-mode PCA where data were reconstructed from the first principal component is 277.7 and 278.5 for the unweighted and flow weighted analyses respectively. Reconstruction error is marginally worse for flow weighted PCA compared to unweighted PCA but since the first principal components explained 89% and 84% of the variance in the data for unweighted PCA and flow weighted PCA respectively, it is reasonable that the reconstruction error is very similar between the two analyses.

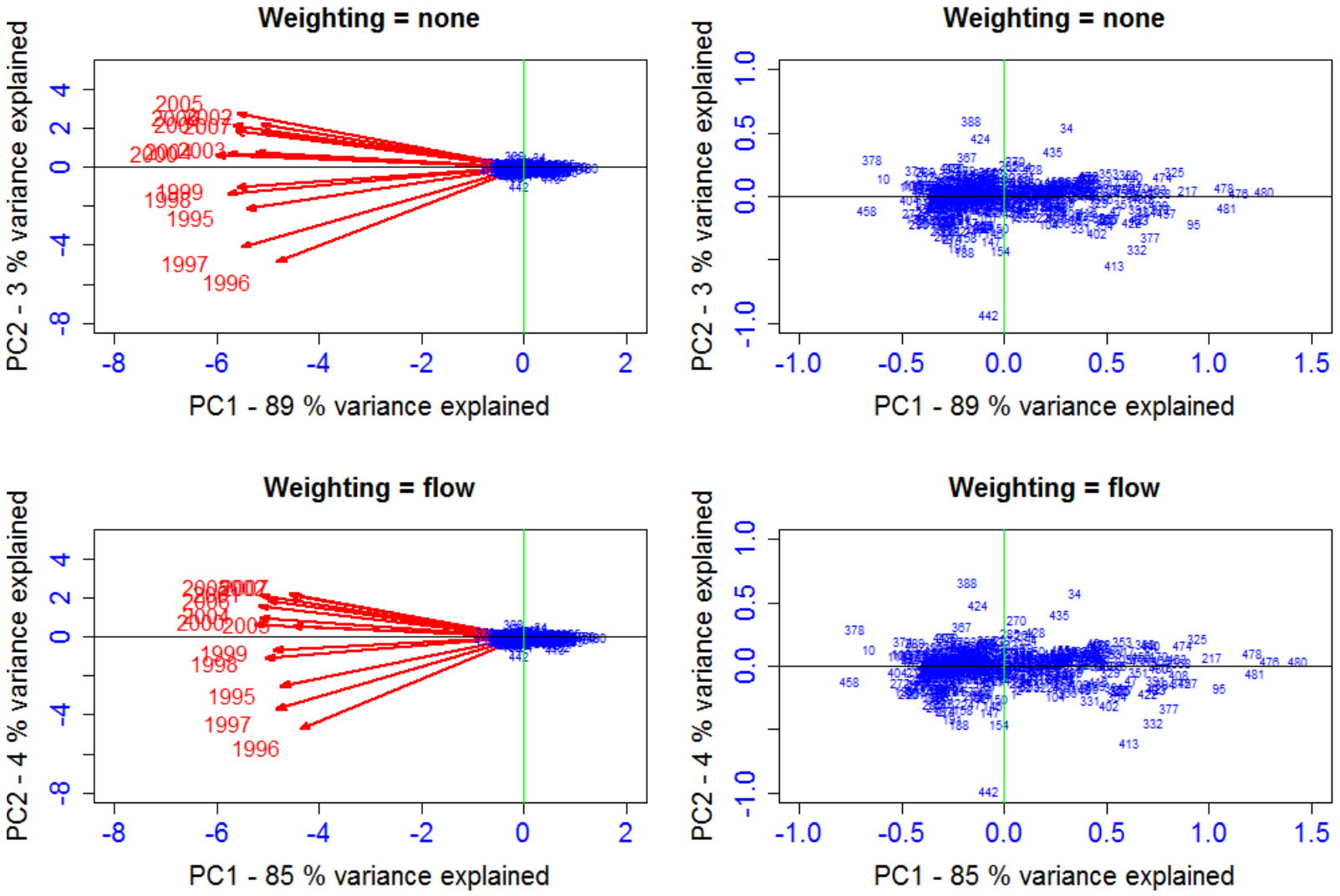


FIGURE 5.2: Biplots for the first two unweighted and flow weighted T-mode principal components. Red arrows show sign and magnitude of loadings. Blue numbers are principal component scores.

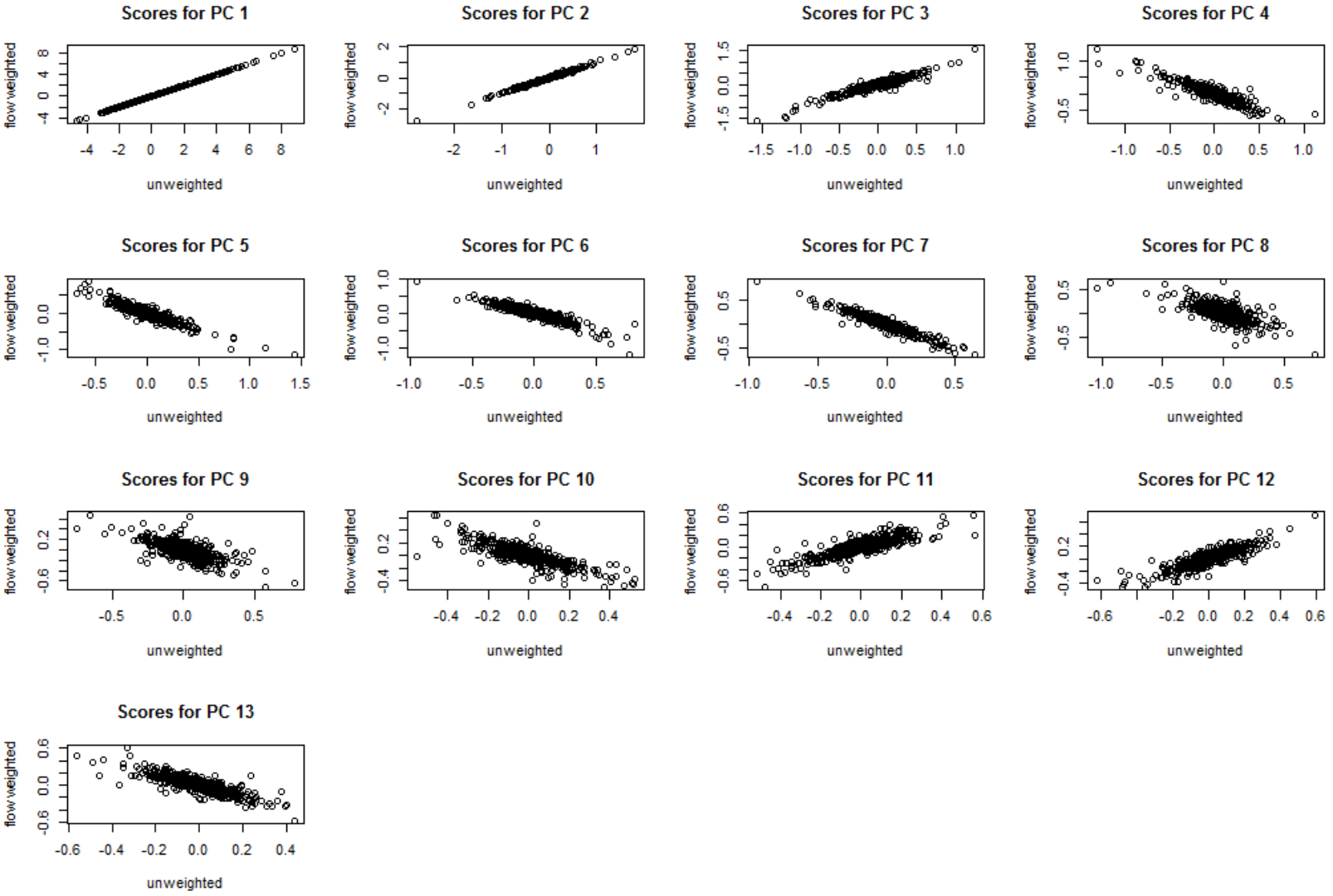
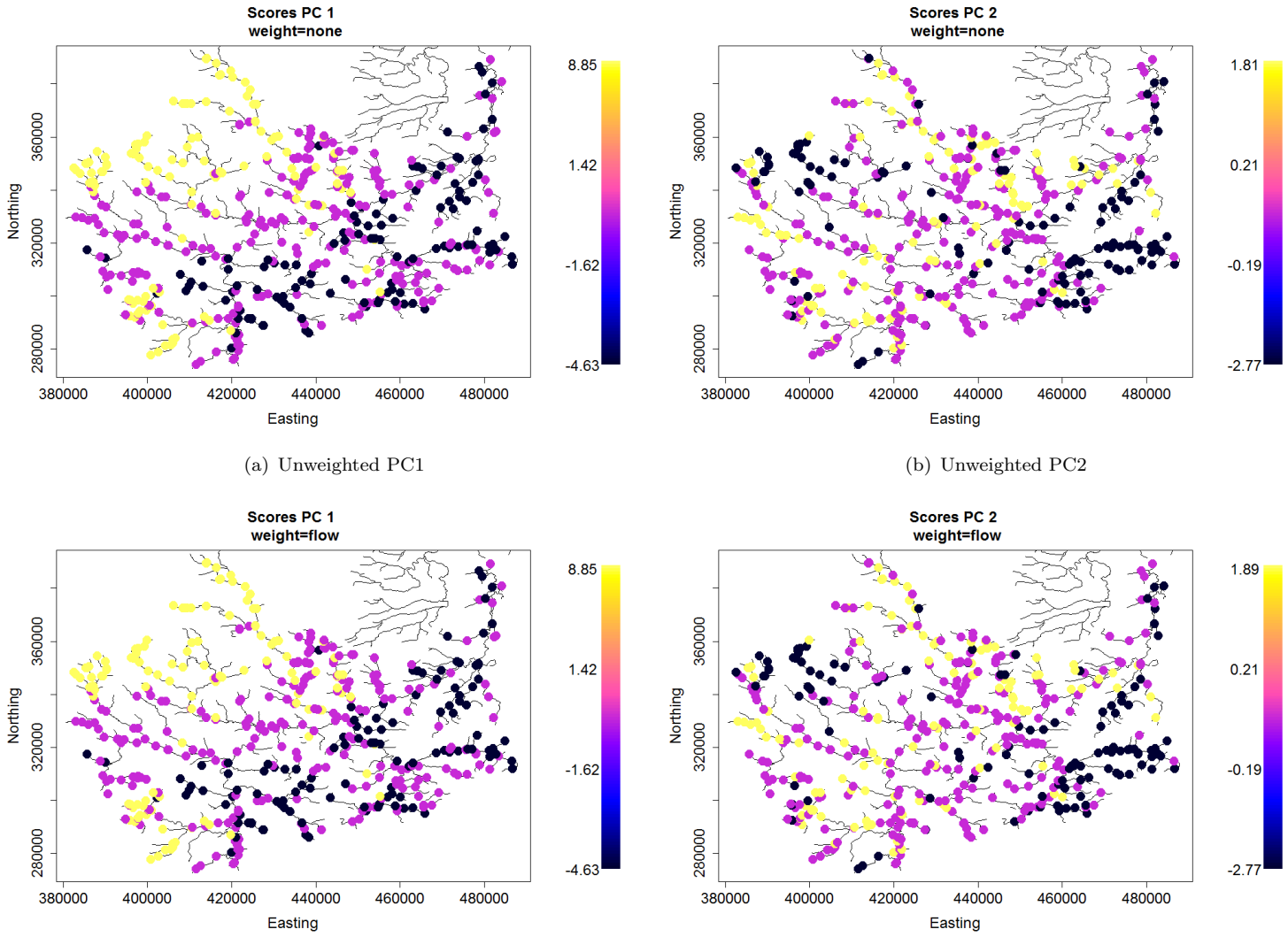


FIGURE 5.3: Principal component scores for unweighted and flow weighted T-mode PCA.



(a) Unweighted PC1

(b) Unweighted PC2

(c) Flow weighted PC1

(d) Flow weighted PC2

FIGURE 5.4: Scores for two unweighted and flow weighed T-mode principal components. Black dots are scores in the lower quartile and yellow dots are scores in the upper quartile.

5.2.1.1 Summary: T-mode PCA

The application of T-mode PCA showed that the spatial pattern of log(TON) is dominated by a single principal component suggesting that the spatial pattern has remained quite stable over time. The principal components maps showed that log(TON) levels largely reflect land use, discussed in Chapter 4. Monitoring sites in the North-West area of the catchment are influenced by high population density and less by agriculture so tend to have lower levels of log(TON) than in the South-East of the catchment which is largely agricultural and has the highest levels of log(TON). This pattern has changed little during winter 1995-2007 and this makes sense since the effect of changing land use does not happen quickly (Burt et al., 2008). The second T-mode principal component explained only a small amount of variance in the data but suggested a contrast between early and later years with monitoring sites in the South-East experiencing a greater decrease in log(TON) than sites in other areas. T-mode PCA has shown that the spatial pattern of log(TON) has remained stable over time but it would be interesting to apply this to a longer time period, including years before and after the implementation of the Water Framework Directive and Nitrates Directive.

5.2.2 S-mode PCA

S-mode PCA was applied to monthly log(TON) for 21 years (1990-2010), following on from the application of T-mode PCA in the previous section. After aggregating the available data to monthly mean log(TON) values there were still approximately 30% of missing values in the data. The imputation method described in Josse and Husson (2012) and implemented in the `R` package `missMDA` was used to complete the data so that S-mode PCA, adjusted for spatial and temporal structure, could be applied. Missing values are imputed using an iterative PCA algorithm. The PCA is carried out using singular value decomposition so is suitable in the situation where there are more variables (columns) than observations (rows) in the data. The algorithm can be initialized by substituting missing values with column means or a random value. For the work in this thesis column means of the original, uncentered data were used as the starting values. The algorithm proceeds as follows:

1. Put initial values in place of missing values.

2. Columns of data are mean centered.

3. PCA is carried out on the "full" centered data set.

4. Missing values are then replaced in the original data set with the values reconstructed values from a specified number of principal components.

5. Column means are recalculated.

6. Steps 2-5 repeated until convergence.

The algorithm is regularized to prevent overfitting by replacing the reconstruction of the data in step 4 with a "shrunk" reconstruction step based on the proposals in Tipping and Bishop (1999). By preventing overfitting this means that components that just contain noise are not used when reconstructing the data. It also means that when there is little correlation in the available data or if there are a lot ("a lot" is not defined) of missing values then the algorithm will tend to impute column means where the means are calculated from the available data.

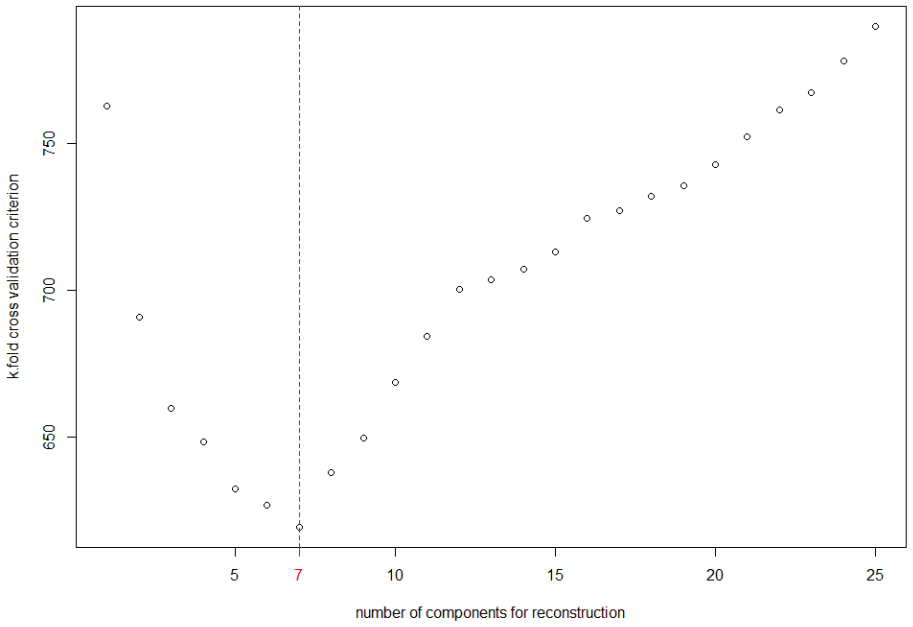


FIGURE 5.5: Results from K-fold cross validation.

The number of principal components used to reconstruct the data in step 4 must be specified in advance and the `missMDA` package includes a function to estimate this. K-fold cross validation was used to estimate the number of principal components to retain

for data reconstruction and this involved 100 simulations where 5% of the observed non-missing data were withheld at random in each simulation. Once the data were reconstructed in step 4, the sum of squared differences was calculated for the withheld points and the number of principal components to retain was selected based on the number that minimised the sum. Figure 5.5 shows the K-fold criterion calculated for 1-15 retained principal components and the criterion is minimised at 7 components. Figure 5.6 shows a plot of the mean monthly log(TON) values across all sites calculated from the observed values (red solid line) and the mean monthly values across all sites calculated from the dataset completed using 7 principal components (black solid line). The two lines are clearly different meaning that the imputation technique used has not simply substituted missing values with column means calculated from observed values. It is also good to note that the imputed values are not strongly influenced by the few very low observed values in the middle of the time period. Little and Rubin (1987) explain that if missing values are replaced with imputed values then it is necessary to assess uncertainty caused by imputation. Josse and Husson (2012) provide a multiple imputation algorithm to create several simulated datasets that can be used to assess uncertainty. Figure 5.7 shows the direction of the first two principal components when PCA is applied to the completed dataset as X and Y axes and the direction of the first two principal components when PCA is applied to the simulated datasets used to assess uncertainty. If there are many arrows of similar length pointing in the direction of the axes then this implies that the uncertainty due to imputation is small. Due to time constraints 300 simulated data sets were constructed.

The previous section described the construction of spatial weights for PCA and explained that using the matrix square root of the inverse of the weight matrix was an appropriate way to adjust PCA for spatial structure based on river network topology. This is in agreement with the discussions in Wartenberg (1985), Baldwin et al. (2009) and Allen et al. (2014). The column weight matrix $\mathbf{\Omega}$ is the transpose of the row weight matrix $\mathbf{\Phi}$ used for T-mode PCA so that water flows from upstream sites in rows to downstream sites in columns. The column weight matrix used for S-mode PCA is therefore $\mathbf{\Omega}^{-\frac{1}{2}}$.

S-mode PCA can be adjusted for temporal structure using row weight matrix $\mathbf{\Phi}^{-\frac{1}{2}}$ where $\mathbf{\Phi}$ is an $n \times n$ symmetric matrix containing the elements $\rho^{|i-j|}$ where ρ is the strength of correlation between observations at time points $1, ..., n-1$ and $2, ..., n$ and

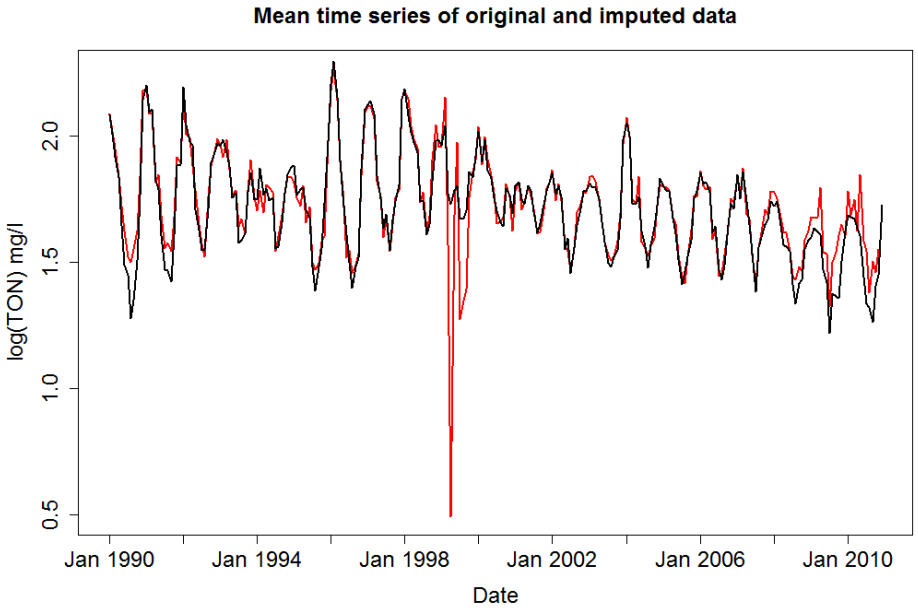
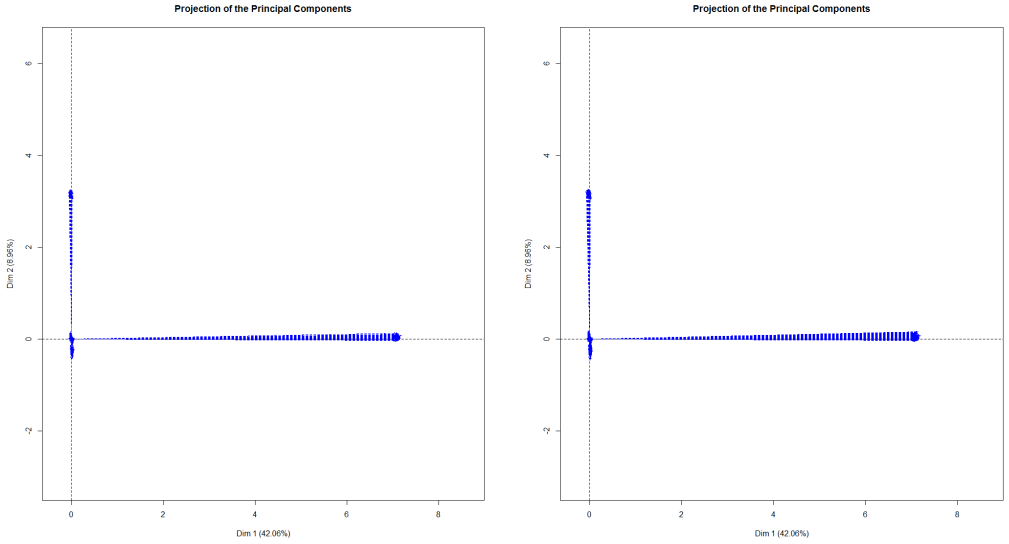


FIGURE 5.6: Mean log(TON) across 566 sites calculated from observed values (solid red line) and data set completed using imputation (solid black line).



(a) (b)

FIGURE 5.7: Direction of the first two principal components estimated for 100 simulated datasets (a) and 200 simulated datasets (b).

$i = 1, .., n; j = 1, ..., n$ and $\mathbf{\Phi}$ is constructed as in (5.9). Chapter 1 explained that in many environmental examples an AR(1) correlation structure is sufficient to model temporal correlation and balances the need to model correlation with a simple model.

$$\mathbf{\Phi} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \\ \rho & 1 & \rho & \rho^2 & \rho^3 & \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \cdots \\ \rho^3 & \rho^2 & \rho & 1 & \rho & \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \\ & & \vdots & & & \ddots \end{pmatrix} \tag{5.9}$$

The value for ρ in the Trent catchment area was estimated by fitting an additive model (5.10) to each of 566 monitoring sites separately, where r indexes month and t indexes year (1990-2010). A cyclic cubic regression spline was used to estimate smooth term s_1 and a cubic regression spline was used to estimate smooth term s_2. The parameters k_1 and k_2 (in this case the dimension of the basis) were selected using generalized cross validation (GCV) and recorded. Additive models were then fitted to each monitoring site again but this time specifying k_1 and k_2 as 4 and 8 respectively - the upper quartile of the 566 values estimated by GCV (after rounding). These values were chosen so that the flexible fitted line for each monitoring site was only undersmoothed for a small proportion of sites. Residuals were calculated for each site and correlation calculated between the residuals at time points t and $t-1$ using the `Hmisc` package (Harrell Jr and with contributions from Charles Dupont and many others., 2014) in `R` so that correlation was calculated only between complete pairs. The median correlation value was 0.27 with an interquartile range of 0.2-0.35 and so $\rho = 0.27$ was used to construct $\mathbf{\Phi}$.

$$log(\text{TON}_{rt}) = s_1(r, k_1) + s_2(t, k_2) + \varepsilon_{rt} \tag{5.10}$$

First, an unweighted S-mode PCA (PCA_{uw}) was applied to the 566 monitoring sites in the dominant network in the Trent catchment area, followed by a flow weighted PCA where column weights were applied to the data reflecting spatial structure in the river network (PCA_f). Finally a row and column weighted PCA was applied ($\text{PCA}_{f\rho}$) to adjust PCA for temporal and spatial structure in the data.

Table 5.2 gives the results from applying S-mode PCA_{uw}, as well as PCA_f and $\text{PCA}_{f\rho}$. For PCA_{uw} the first component explains 42% of the variance in the data. Adjusting for spatial structure means this reduces to 38% and adjusting for spatial and temporal

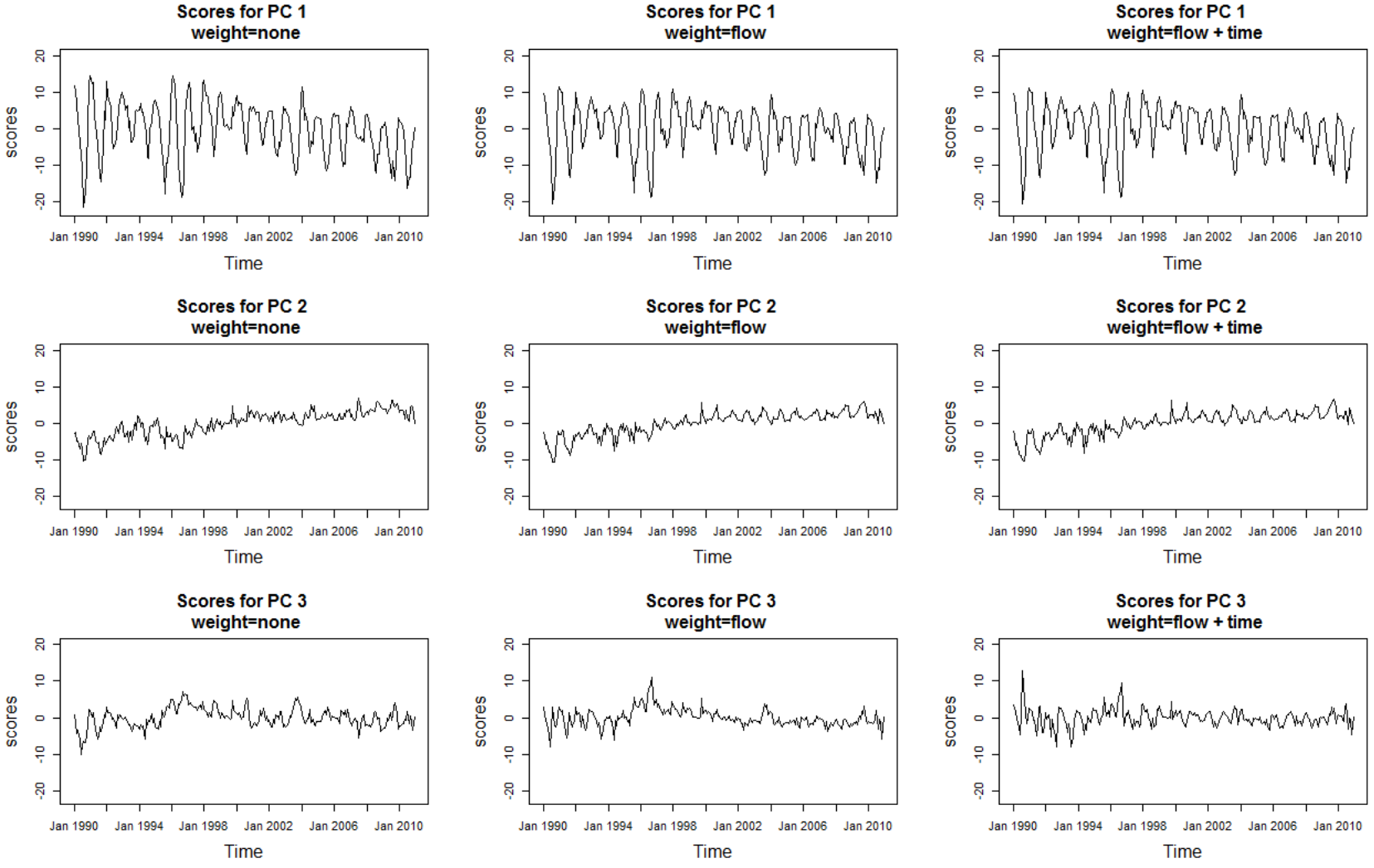
structure means that the first PC accounts for 31% of the variance in the data. The first three components (var_3 in Table 5.2) for PCA_{uw}, PCA_f and $\text{PCA}_{f\rho}$ account for 57%, 52% and 43% of the variance respectively. In order to explain at least 70% of the variance (var_k in Table 5.2) PCA_{uw} requires 8 components and k increases to 12 when PCA is adjusted for spatial structure and k increases further to 23 when PCA is adjusted for both spatial and temporal structure. As a result of this, the reconstruction error calculated as the sum of squared differences between the centered data $\mathbf{X}$ and the data reconstructed using k retained principal components $\hat{\mathbf{X}}$ (5.11) (SSD_k in Table 5.2) decreases by 8% when PCA is adjusted for spatial structure and by 24% when PCA is adjusted for both spatial and temporal structure.

PCA	PC1	PC2	PC3	var_3	k	var_k	SSD_k
PCA_{uw}	42%	9%	6%	57%	8	70.8%	9069
PCA_f	38%	9%	5%	52%	12	70.5%	8354
$\text{PCA}_{f\rho}$	31%	7%	5%	43%	23	70.1%	6910

TABLE 5.2: Results from PCA_{uw}, PCA_f and $\text{PCA}_{f\rho}$. var_3 is the amount of variance explained by the first three principal components, k is the number of principal components retained to explain at least 70% of the variance of the data, var_k is the amount of variance explained by k principal components, SSD_k is the reconstruction error from k principal components.

Time plots of the scores for the first three components from PCA_{uw}, PCA_f and $\text{PCA}_{f\rho}$ are in Figure 5.8. Figure 5.8(a) shows that the first principal component from PCA_{uw} estimates a temporal pattern with greater variability in early years compared to later years and a decrease in log(TON) from around 1999. The second principal component shows greater variability in the early part of the time period under consideration and a small increase in log(TON) in later years while the third component shows a peak in log(TON) around 1997 followed by a shallow decline in log(TON). Figure5.8(b) shows that the temporal pattern of log(TON) is slightly less variable around 1996 than for the first unweighted principal component and the second flow weighted principal component highlights greater variability in the early years with an increase in log(TON) between 1990 and 2000. The third flow weighted principal component shows a spike in log(TON) around 1997, similar to the third unweighted principal components. For $\text{PCA}_{f\rho}$ shown in Figure 5.8(c) the scores for the first principal component are less variable in the first half of the time period than for PCA_{uw} but the third $\text{PCA}_{f\rho}$ principal component seems to capture the extra variability.

Figure 5.9 shows mean log(TON) for monthly observations ± a small multiple of the PC's in Figure 5.8 to show the effect of the temporal patterns estimated using $SPCA_{uw}$, $SPCA_c$ and $SPCA_{rc}$ on the mean temporal pattern, borrowing an idea from functional PCA (Ramsay and Silverman, 2006). The plots are limited to the first half of the time period so as to make it possible to see any differences between the results weighted and unweighted analyses more clearly. The plots in the left panel show that there is almost no difference between the first principal component for $SPCA_{uw}$, $SPCA_c$ and $SPCA_{rc}$ except for the variability around the peak at January 1996 where $SPCA_{uw}$ estimates greater variability around that point than the weighted analyses. For the second PC, log(TON) is less variable around January 1995 for $SPCA_{uw}$ and more variable around January 1996 than $SPCA_c$ and $SPCA_{rc}$. Greater differences are seen between the three analyses in the third principal component, in particular the variability around the troughs in January 1990, 1991 and 1992 is quite different between $SPCA_{uw}$, $SPCA_c$ and $SPCA_{rc}$. S-mode PCA has shown that PC1 reflects variation around the mean such that the periodic signal is dampened down, indicated by the crossover pattern when a small amount of the principal component is added to or subtracted from the mean temporal pattern. PC2 reflects a shift in mean value while PC3 for $SPCA_{uw}$ and $SPCA_c$ correspond to a small time lag, indicated by the crossover pattern of adding or subtracting a small amount of the PC to the mean pattern and variability occurs around the slope rather than the turning points



(a) (b) (c)

FIGURE 5.8: Scores plots for first three principal components from PCA_{uw}, PCA_{f} and $\text{PCA}_{f\rho}$.

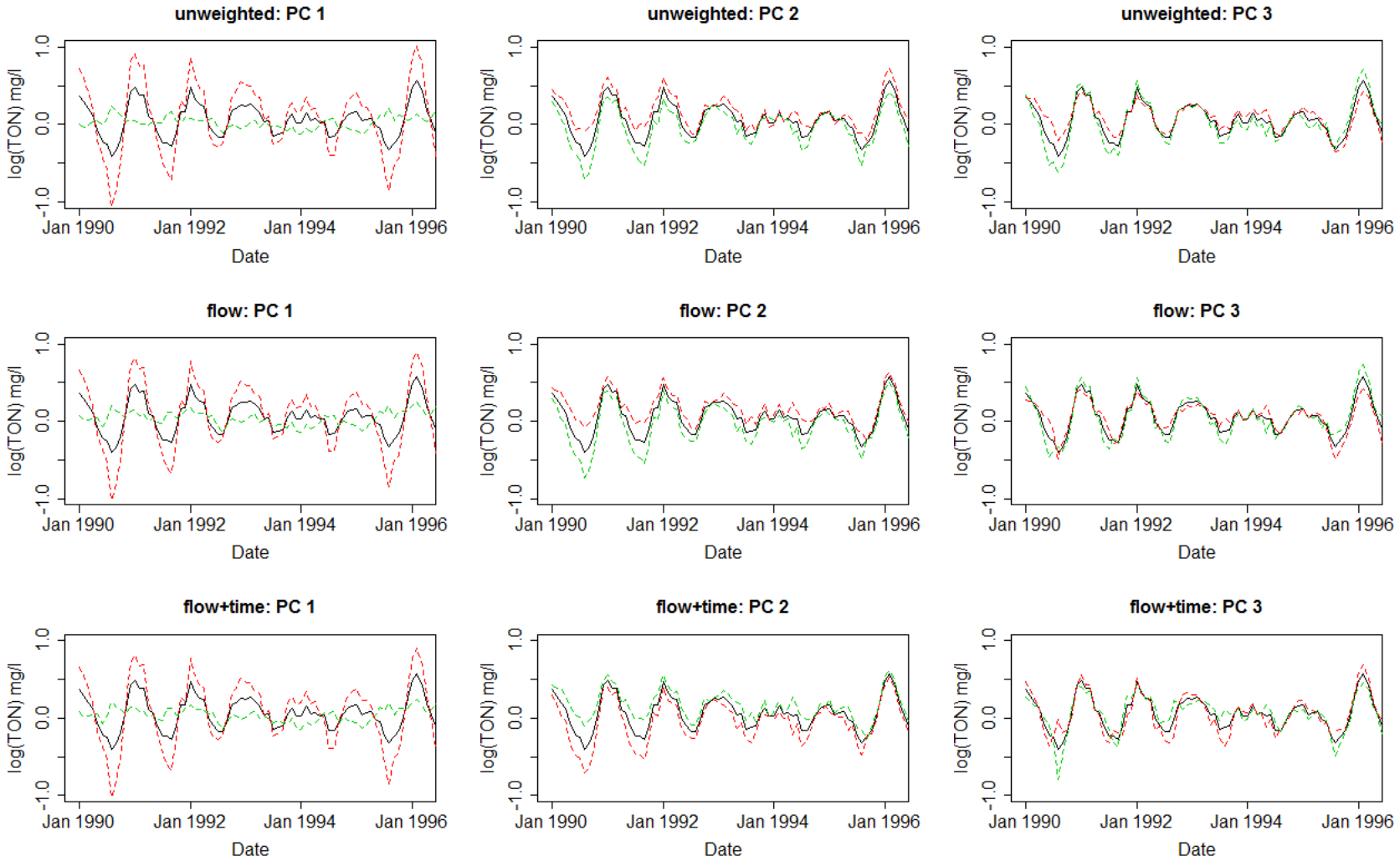


FIGURE 5.9: Mean log(TON) for 566 monitoring sites (solid black line) with + (green dashed line) and - (red dashed line) a small proportion of the principal component scores from PCA_{uw}, PCA_{f} and $\text{PCA}_{f\rho}$.

In S-mode PCA maps of the loadings can be used to show which monitoring sites exhibit similar temporal patterns. Figure 5.10 shows maps of the loadings for PCA_{uw} (left), PCA_{f} (middle) and $\text{PCA}_{f\rho}$ (right) for the first three principal components (top to bottom). Black points are monitoring sites whose loadings fall in the lower quartile of the distribution of loadings from the three analyses while yellow points are monitoring sites with loadings in the upper quartile. The distribution of loadings is very similar for the first principal component between the three weighting schemes and greater differences can be found for the third principal component. It might also be useful to consider the loadings for the first three components together as in Figure 5.11 where the loading for the first principal component is shown in the twelve o'clock position, moving round clockwise to the second and third principal components.

Adjusting for spatial structure by using PCA_{f} reduced reconstruction error by 8% compared to PCA_{uw} and adjusting for both spatial and temporal structure reduced the error by 24%. Figure 5.12 shows the sum of squared differences for each site for each of the three weighting schemes. Adjusting for river network structure reduced reconstruction error at the two sites with the highest reconstruction error under PCA_{uw} by 80% and adjusting for spatial and temporal structure reduced reconstruction error at these two sites by 90%.

For S-mode PCA, accounting for spatial structure in the data meant that the percentage of variance explained by the first principal component in PCA_{f} and $\text{PCA}_{f\rho}$ was smaller than for PCA_{uw}. An explanation is that using inverse weights means that the data have been decorrelated as shown in Figure 5.13 where inverse weighting makes correlation closer to zero.

By inverse weighting the data some of the correlation structure is removed, making the columns of $\mathbf{X}$ less dependent. In PCA if all p columns (variables) are fully independent then each component would only explain $1/p\%$ total variance in the data. Allen et al. (2014) state that using weights designed to decorrelate the data means that SVD with equally weighted errors is appropriate and that while the singular values (the square root of the eigenvalues of the covariance matrix) are the singular values of the decorrelated data, the spatial and temporal structure is multiplied back into the principal components as in Section 5.1.

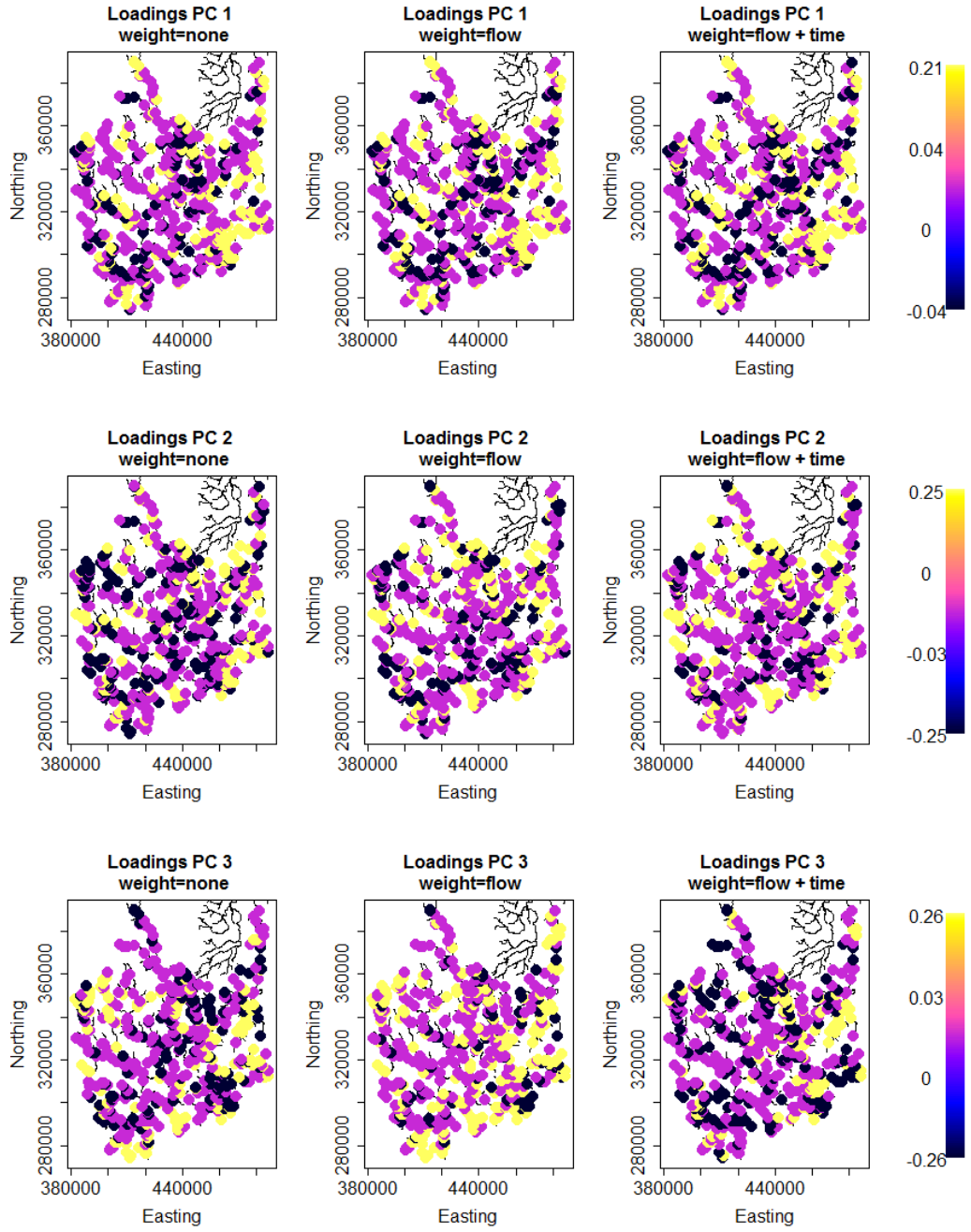


FIGURE 5.10: Loadings for first three principal components from PCA_{uw}, PCA_{f} and $\text{PCA}_{f\rho}$.

5.2.2.1 Summary: S-mode PCA

S-mode PCA has been used to identify dominant temporal patterns in the Trent catchment area. Glyph maps of loadings were used to show monitoring sites that behave similarly over time and were useful for displaying results for more than one principal component. The glyph maps suggested that adjusting PCA for spatial and temporal

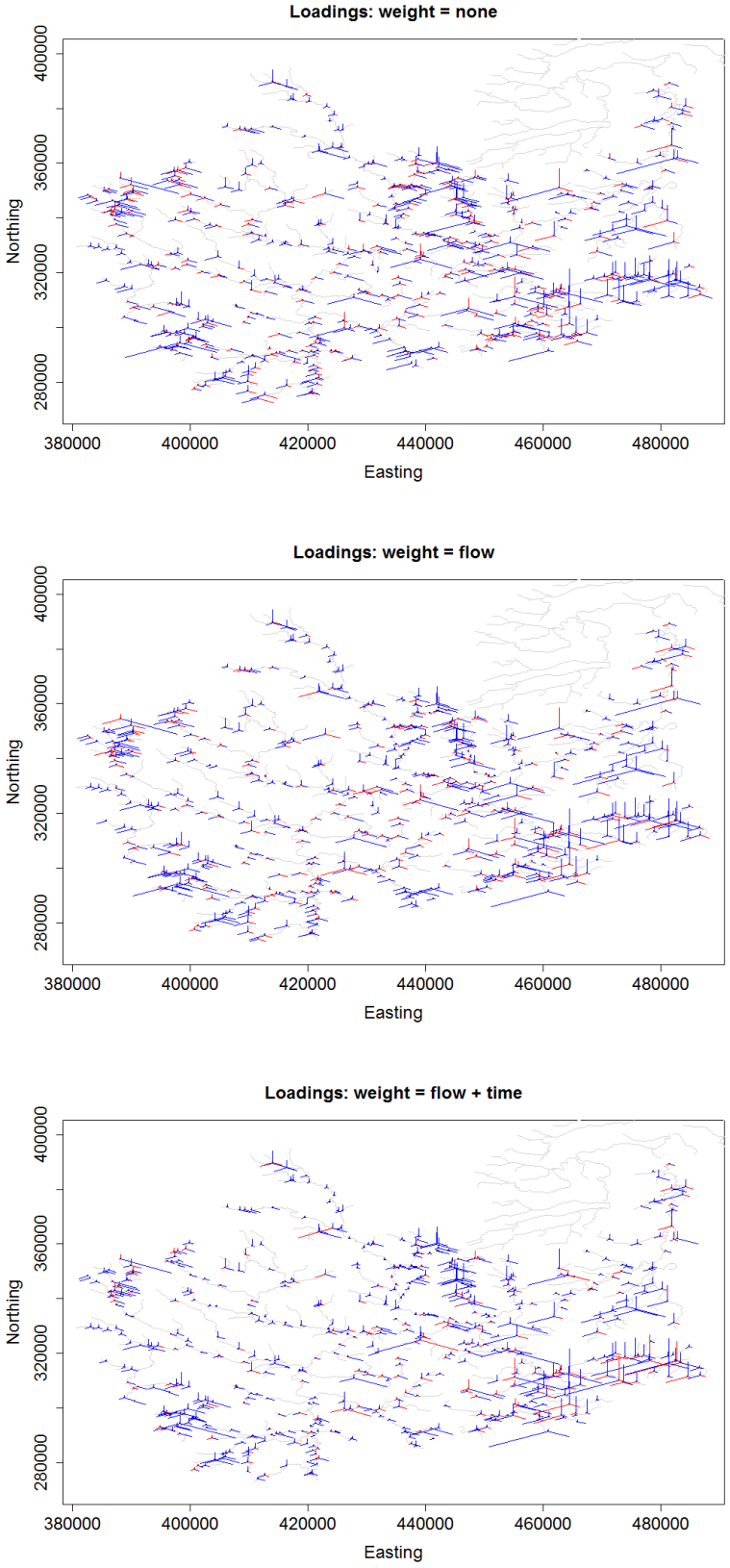


FIGURE 5.11: Glyph plots with Loadings for first three principal components from PCA_{uw}, PCA_{f} and $\text{PCA}_{f\rho}$. Red indicates negative values and blue indicates positive values. Length of line indicates magnitude of loading relative to others.

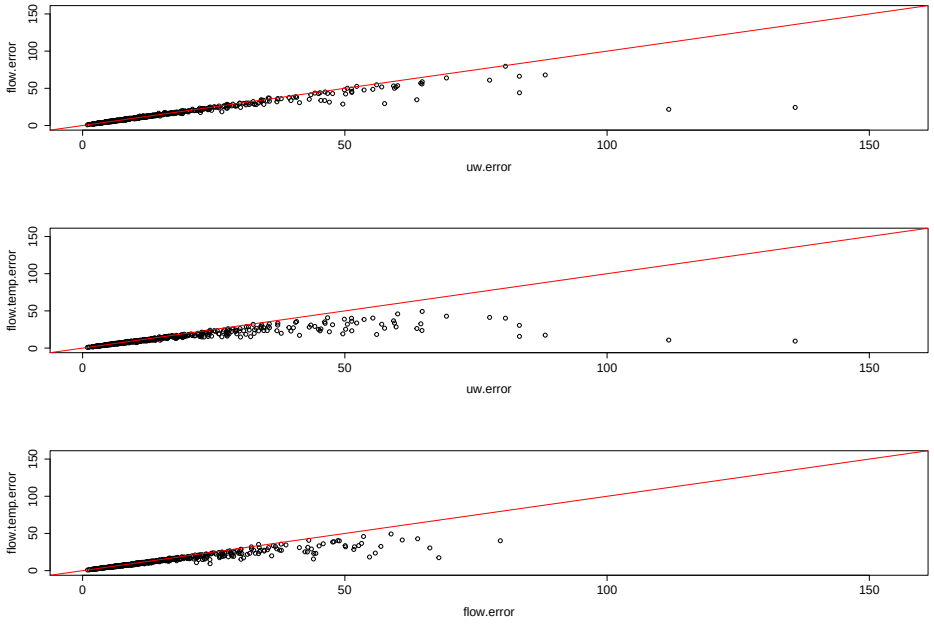
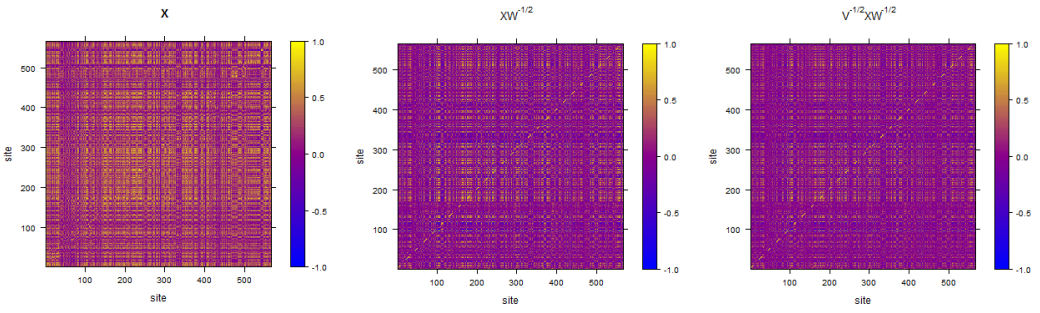


FIGURE 5.12: Points represent sum of squared differences between X and $\hat{X}$ for each of 566 monitoring sites. Red line is $x = y$. uw.error is from PCA_{uw}, flow.error is from PCA_{f} and flow.temp.error is from $\text{PCA}_{f\rho}$.



(a) (b) (c)

FIGURE 5.13: Correlation between time series for 566 sites. (a) correlation for $\mathbf{X}$, (b) correlation for $\mathbf{X\Omega}^{-\frac{1}{2}}$, (c) correlation for $\mathbf{\Phi}^{-\frac{1}{2}}\mathbf{X\Omega}^{-\frac{1}{2}}$

structure means that clearer distinctions can be made between the dominant and less dominant monitoring sites (although in the application to the Trent catchment area these differences were quite subtle). Time series of the principal components were used to illustrate the temporal patterns identified, and plots of the mean time series from the catchment area $\pm$ each of the principal components were used to show how the dominant time series related to the mean time series. Figure 5.9 showed that the differences in the temporal pattern described by the first principal component between PCA_{uw}, PCA_{f} and $\text{PCA}_{f\rho}$ are quite subtle. In general, using column weights to adjust PCA for known

spatial structure resulted in a less variable temporal pattern than when spatial structure is not accounted for. The second and third principal components reflect deviations from the dominant temporal pattern in all S-mode analyses.

The MRI example in Allen et al. (2014) showed that adjusting PCA for spatial and temporal structure meant that the dominant signal in the data could be clearly separated from structured noise. The separation between signal and noise is not quite so clear in the application to data from the Trent and there are several reasons why this could have happened. It might be that the dominant temporal pattern in the Trent catchment area was already quite clearly distinguished from noise meaning that PCA_f and $\text{PCA}_{f\rho}$ had only subtle differences from PCA_{uw}. The simulation study in Allen et al. (2014) investigated the effect of adjusting PCA for spatial and temporal correlation of various strengths and it would be interesting therefore to simulate data on a river network with varying strengths of flow connected spatial correlation to investigate the effect this would have on the results of T- and S-mode PCA. Another possible reason for such subtle differences between the unweighted and weighted PCA is the spatial structure in the data is dominated by land use, related to the Euclidean distance based component of the spatial covariance function as shown in Chapters 3 and 4. Since the spatial weight matrix developed in this chapter is based on the stream distance flow connected structure in the data it is reasonable to assume that adjusting PCA for this structure will have a smaller effect on the results that if PCA were adjusted for Euclidean distance based spatial structure. A hybrid weight matrix combining both Euclidean distance and stream distance based spatial structure might provide even further insight. This could form the basis for future development of the methodology.

5.2.3 Comparison to existing method

This section will discuss the methods developed in Allen et al. (2014) who use symmetric weight matrices to adjust for structure in the rows and columns of the data matrix, and show how the proposed adaptation of PCA using an asymmetric weight matrix, developed in this Chapter, gives comparable results.

Allen et al. (2014) who propose a generalized singular value decomposition (GSVD) that takes into account structure in the rows and columns of the data. This GSVD finds the best low-rank approximation with respect to the QR-norm which weights errors

unequally rather than minimimising the sum of squared differences (also known as the Frobenius norm) between the data and data reconstructed from principal components which weights errors equally. For flow weighted T-mode PCA, the principal component scores calculated using asymmetric row weight matrix $\mathbf{\Phi}^{-\frac{1}{2}}$ are 1.65×the principal component scores calculated using the methods in Allen et al. (2014) if (using terminology from Allen et al. (2014)) row weight matrix Q is set to

$$\mathbf{Q} = \mathbf{\Phi}^{-\frac{1}{2}\top}\mathbf{\Phi}^{-\frac{1}{2}},$$

since T-mode PCA with row weights representing flow direction is equivalent to decomposing

$$\left(\mathbf{\Phi}^{-\frac{1}{2}}\mathbf{X}\right)^{\top}\left(\mathbf{\Phi}^{-\frac{1}{2}}\mathbf{X}\right) = \mathbf{X}^{\top}\mathbf{\Phi}^{-\frac{1}{2}\top}\mathbf{\Phi}^{-\frac{1}{2}}\mathbf{X} = \mathbf{X}^{\top}\mathbf{Q}\mathbf{X}.$$

The row weight matrix used in Allen et al. (2014) is therefore symmetric whereas the method presented in this chapter was developed so that an asymmetric weight matrix reflecting flow direction could be incorporated into PCA methodology. GSVD aims to minimise reconstruction error in the QR-norm rather than the Frobenius norm by calculating

$$\mathrm{Tr}\left[\mathbf{Q}(\mathbf{X}-\hat{\mathbf{X}})(\mathbf{X}-\hat{\mathbf{X}})^{\top}\right]. \tag{5.11}$$

Table 5.3 shows the reconstruction error with respect to the Frobenius norm and the QR-norm. As discussed earlier, reconstruction error in the Frobenius norm is marginally worse for the flow weighted PCA but this is to be expected since the first principal component explained a smaller percentage of the variance in the data than the unweighted component. In the QR-norm however the flow weighted PCA has lower reconstruction error even though the first flow weighted principal component explains less of the variance in the data than the unweighted component.

For S-mode PCA with spatially weighted columns, setting column weight matrix R in Allen et al. (2014) to

Norm	Unweighted	Flow weighted
Frobenius	277.7	278.5
QR	89.6	89.4

TABLE 5.3: Reconstruction error for unweighted and flow weighted T-mode PCA.

$$\left(\mathbf{\Omega}^{\frac{1}{2}\top}\mathbf{\Omega}^{\frac{1}{2}}\right)^{-1}$$

gives loadings such that these loadings are equal to $0.607\times$ the transformed loadings $\mathbf{\Omega}^{\frac{1}{2}\top}\tilde{\mathbf{V}}$ in (5.2) but the reconstruction errors in both the Frobenius and QR-norm are the same for the asymmetric column weighted PCA developed in this chapter and the symmetric column weighted PCA in Allen et al. (2014). For S-mode PCA with row and column weights the QR-norm is calculated as

$$\text{Tr}\left[\mathbf{Q}(\mathbf{X}-\hat{\mathbf{X}})\mathbf{R}(\mathbf{X}-\hat{\mathbf{X}})^{\top}\right],$$

where $\mathbf{Q}$ are the row weights adjusting PCA for temporal structure:

$$\mathbf{Q} = \mathbf{\Phi}^{-\frac{1}{2}\top}\mathbf{\Phi}^{-\frac{1}{2}} = \mathbf{\Phi}^{-1}$$

since $\mathbf{\Phi}^{-\frac{1}{2}}$ is symmetric. $\mathbf{R}$ are the column weights adjusting PCA for spatial structure:

$$\mathbf{R} = \left(\mathbf{\Omega}^{\frac{1}{2}\top}\mathbf{\Omega}^{\frac{1}{2}}\right)^{-1}.$$

Note that $\mathbf{R}$ is symmetric even though $\mathbf{\Omega}^{-\frac{1}{2}}$ is not. The structure of $\mathbf{R}$ is determined by Allen et al. (2014) who show that $\mathbf{R} = \tilde{\mathbf{R}}\tilde{\mathbf{R}}^{\top}$ and the column weighted data matrix $\tilde{\mathbf{X}}$ is $\mathbf{X}\tilde{\mathbf{R}}$. In order that $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{\Omega}^{-\frac{1}{2}}$,

$$\tilde{\mathbf{R}}\tilde{\mathbf{R}}^{\top} = \mathbf{\Omega}^{-\frac{1}{2}}\mathbf{\Omega}^{-\frac{1}{2}\top} = \left(\mathbf{\Omega}^{\frac{1}{2}\top}\mathbf{\Omega}^{\frac{1}{2}}\right)^{-1}.$$

Table 5.4 shows the Frobenius norm and QR-norm calculated for PCA_{uw}, PCA_{f} and $\text{PCA}_{f\rho}$. In the Frobenius norm $\text{PCA}_{f\rho}$ has the smallest reconstruction error but this is based on data reconstructed from 23 principal components compared to 8 principal components for PCA_{uw}. In the QR-norm based on k principal components $\text{PCA}_{f\rho}$ also has the smallest reconstruction error but interestingly $\text{PCA}_{f\rho}$ also has the smallest reconstruction error when data are reconstructed from only 8 principal components to make the result comparable with PCA_{uw}.

PCA	k	var_k	Frob_k	QR_k	QR_8
PCA_{uw}	8	70.8%	9069	3365	
PCA_{f}	12	70.5%	8354	2967	3309
$\text{PCA}_{f\rho}$	23	70.1%	6910	1457	2070

TABLE 5.4: Results from PCA_{uw}, PCA_{f} and $\text{PCA}_{f\rho}$. k is the number of principal components retained to explain at least 70% of the variance of the data, var_k is the amount of variance explained by k principal components, Frob_k is the reconstruction error in the Frobenius norm (sum of squared differences) from k principal components, QR_k is the reconstruction error in the QR-norm with k principal components and QR_8 is the reconstruction error in the QR-norm with 8 principal components.

5.3 Comments

This chapter has developed a novel adaptation to standard PCA methodology to incorporate an asymmetric weight matrix reflecting river network topology. This removes known sources of variability in the data such as variability between pairs of monitoring sites located on opposite sides of a confluence point and might make it possible to more clearly identify common temporal or spatial patterns in the data (see Frichot et al. (2012) for an example of new spatial patterns revealed after adjusting PCA for spatial correlation). Improving the estimation of common temporal patterns in the data could provide regulatory agencies with evidence of duplicated information in the river network, thus strengthening the argument to reduce the number of monitoring sites leading to cost savings.

Application of weighted PCA in T- and S-mode has revealed a stable spatial pattern over 13 years of annual winter log(TON) and the dominant temporal pattern for 21 years of monthly log(TON). If a dataset collected over a longer time period were available then adjusting T-mode PCA for spatial structure might more clearly reveal different spatial patterns of log(TON) over time since variability in the data would more closely reflect

changes in land use rather than being a mixture of variability based on land use and river network structure. The differences between weighted and unweighted PCA in both T-mode and S-mode were quite subtle and one explanation for this is that spatial and temporal correlation was only weak to moderate in strength for the data from the Trent catchment area. Future work therefore could include simulating data with higher levels of spatial and temporal correlation to investigate the effect of adjusting PCA for spatial correlation due to river network structure.

The results from the analyses in this chapter were shown to be comparable with existing weighted PCA methodology but the method developed in this chapter specifically deals with an asymmetric weight matrix designed to reflect the direction of river flow and strength of relationship between flow connected monitoring sites. Existing literature on weighted PCA methods does not discuss the inclusion of an asymmetric weight matrix. It was shown that the orientation of the spatial weight matrix in relation to the data matrix is crucial so that the direction of flow is accounted for properly. This method has potential to improve identification of dominant spatial and temporal patterns in a spatiotemporal dataset. A particular advantage is that structured noise is separated from the main signal in the data and so the methodology developed in this chapter is especially useful for noisy data such as high frequency data.

Chapter 6

Conclusions and further work

The overall aim of this thesis has been to adapt and extend existing statistical methods that can be used to estimate common temporal patterns in water quality data, taking into account river network structure. Environmental policies such as the Water Framework Directive (European Parliament, 2000) require that member states monitor the temporal evolution of water quality in rivers as well as other water bodies such as lakes and estuaries. With approximately 7000 monitoring sites located on rivers in England and Wales it is difficult to interpret the temporal pattern by looking at monitoring sites individually. At the other extreme, looking at the temporal pattern for a large hydrological area by combining the data from several monitoring sites means it is only possible to consider a single temporal pattern for each area which might mask other important temporal patterns. Statistical methods such as principal components analysis can be used to identify the dominant pattern, or patterns if more than one exists, in a spatiotemporal dataset. This thesis has adapted PCA to take into account the spatial structure of river networks with the aim of providing additional insight into the estimated common temporal pattern of nitrates. The techniques developed in this thesis could be applied to any spatiotemporal example where there is temporal and direction dependent spatial structure.

6.1 Comparing temporal trends and seasonal patterns

Chapter 2 investigated statistical methods that can be used to compare smooth curves representing the temporal trends and seasonal patterns of log(TON) for 59 LHA's in England and Wales. First, the 59 temporal trends were compared using dynamic factor analysis to find common patterns shared among the 59 LHA's. It was found that the 59 curves were best modelled using a single common trend and a diagonal and equal covariance matrix suggesting that variance is similar for all LHA's. A diagonal covariance matrix seems reasonable since LHA's are constructed to be independent of neighbouring areas. The common temporal pattern estimated indicated a maximum log(TON) around 1996 and a smaller peak around 2006 with an overall decline in log(TON) since 1995. The seasonal pattern for the 59 LHA's could also be modelled using a single common trend with high values in winter months and low values in the summer and as with the temporal trends, the covariance structure was estimated to be diagonal and equal.

DFA has many advantages as a statistical method to estimate dominant patterns among several time series. Rather than estimating the average of the 59 temporal trends and seasonal patterns, DFA can estimate several common trends, providing additional insight into the data. Spatial structure can be incorporated into DFA through the covariance matrix where off-diagonal elements capture spatial relationships not explained by the common trends (or explanatory variables if available) and recently spatial forms of DFA have been developed by Lopes et al. (2008), Lopes et al. (2011) and Strickland et al. (2011). Missing data can also be easily handled within DFA by the EM algorithm. This does rely however on data being recorded at regular intervals and at the same time points for all monitoring sites. Irregular sampling was solved in Chapter 2 by evaluating the temporal trend, estimated for each LHA using an additive model, at the same regularly spaced time points. A further drawback of DFA is the high computational cost as the number of time points and time series increases. Brodgar software (`http://www.brodgar.com/`) can only handle up to 30 time series and although the `MARSS` package (Holmes et al., 2012) in `R` was able to model 59 time series, the DFA model with unconstrained covariance matrix and a single common trend took 4 days for parameters to converge on a standard desktop computer. The application of DFA suggests that other, less computationally demanding, dimension reduction techniques will prove useful when estimating common temporal patterns for water quality data.

The comparison of temporal trend and seasonal pattern curves was continued by focusing on a single LHA (LHA61) and estimating the temporal trend and seasonal pattern using four statistical methods: additive models fitted using the `mgcv` package in `R` (GAM), additive models fitted using the `inla` package (INLA), functional data analysis (FDA) and dynamic factor analysis (DFA). Each of these methods is suitable for modelling temporal trends and seasonal patterns in spatiotemporal data and incorporate spatial information in different ways. The DFA results were interesting in that the best models for temporal trend and seasonal pattern respectively had two common trends suggesting that a single mean curve might not be the best way to describe the temporal pattern of log(TON) in this LHA. The shape of temporal trends and seasonal patterns estimated from four statistical methods were compared to assess differences between the curves. In order to compare the shape of the curves they were first normalised so that values were in the range $[0, 1]$. This was necessary because the y axis scale for the DFA estimates is not on the scale of the original data and normalising the x and y axes means that curves can be compared based on shape. The comparison was made using the DBF statistic, originally developed for use in genomics applications. This tests if the distance between mean curves from two or more groups is significantly different from zero where distance can be defined using any suitable measure. In this thesis a measure of visual distance was used to reflect how differences between curves might be perceived by eye and takes into account both horizontal and vertical distances. The results of the DBF test showed that the curves were significantly different.

A novel comparison approach based on curvature was developed to identify the nature of significant differences between curves estimated from four statistical methods. Curvature was calculated for normalised curves and confidence intervals for curvature were developed based on the distributions of functions of random variables. Curvature and confidence intervals were plotted and overlapping confidence intervals indicated sections of the curves where shape, in terms of curvature, was not significantly different. For temporal trend the GAM, FDA and DFA trend 1 were most similar with turning points generally occurring at the same time. DFA trend 2 suggested turning points occurring earlier in the time period and the INLA estimate of temporal trend could be described as a mixture of the two DFA trends. For the seasonal pattern the estimates were quite similar in terms of magnitude of curvature at turning points for the four statistical methods

being compared although the GAM estimate suggested maximum and minimum values of log(TON) occurring earlier in the year than the other methods while the INLA estimate suggested these occurred later in the year than other statistical methods.

Analysts will often use their preferred statistical technique to estimate temporal trends and seasonal patterns rather than consider the different interpretations that might result from the application of a variety of statistical techniques. The novel comparison of shapes of curves estimated from four statistical methods shows that different statistical methods can lead to different estimates of temporal trend and seasonal pattern. The FDA approach where curves were estimated at each monitoring site and averaged is less suitable than the other methods compared since this approach appeared to flatten the the curve, with lower curvature values than other methods. The DFA model had two common trends suggesting that the mean curve alone was not the best description of the temporal pattern of nitrates in LHA61 and the INLA curves appeared to be a mixture of the two DFA curves. The GAM curves were smoother than the INLA curves but the implementation is much simpler for GAM than INLA. The GAM curves, unlike those estimated using INLA, did not include spatial covariance which might account for differences between the GAM and INLA estimates. Going forward, it is recommended that analysts use DFA to investigate if more than one curve is required to represent the temporal pattern of water quality data, for several monitoring sites. If the purpose of the analysis is to estimate the single curve that best describes the temporal behaviour of nitrates over several monitoring sites then either GAM or INLA models are suitable. Where possible, it is recommended that information about spatial dependence should be incorporated into the model.

This novel comparison based on curvature could be used by regulatory agencies to compare small numbers of temporal trends or seasonal patterns and provide additional insight into the nature of differences or similarities of curves. For example, the seasonal pattern could be compared between monitoring sites or hydrological areas to determine if the period of non-zero curvature around the annual minimum log(TON) value is statistically different between sites or areas which might indicate the presences of different seasonal dynamics at different sites or areas.

6.2 Investigating spatial covariance structure

Chapters 3 and 4 investigated the spatial pattern of log(TON) at a single time point. In Chapter 3 several statistical models were fitted to seasonally averaged log(TON) for 1990-2010, 1990-2000 and 2003-2010 (referred to as 'all', 'early' and 'late' years respectively) to assess suitable spatial covariance structures for observations in the Trent catchment area, and whether the choice of covariance structure was the same throughout the year and over a 21 year period. The stability of the covariance structure was also assessed for different river networks within the same catchment area. Four forms of covariance structure were considered: non-spatial where monitoring sites were assumed independent, Euclidean distance based covariance (Euclidean models), stream distance based covariance (Tail-up models) and hybrid covariance where Euclidean distance and stream distance were combined. The stream distance based covariance functions are weighted to maintain stationarity of the variance and reflect the relative influence of upstream monitoring sites on downstream sites.

In Chapter 3 several models were fitted and the best model was selected using root mean square prediction error (RMSPE) since one of the main aims of geostatistical modelling is to predict values at unobserved locations. This showed, for the dominant river network in the Trent catchment area, that a hybrid covariance structure was most suitable and in particular the hybrid model with Epanechnikov Tail-up and Gaussian Euclidean covariance functions was the best combination for almost all seasonal log(TON) averaged over all, early and late years. There was very little difference in RMSPE between the hybrid models with Gaussian Euclidean covariance combined with any Tail-up model but the hybrid models clearly performed better than all other models, indicating that the choice of covariance structure is more important than the particular covariance function used i.e. once the hybrid covariance structure has been selected there is little to distinguish between particular choices of Tail-up or Euclidean models. Geostatistical models were also fitted to the second largest network in the Trent and the best model also had a hybrid covariance structure although the Exponential Euclidean model was preferred to other Euclidean models. Models were then fitted to multiple networks within the same LHA to investigate if multiple networks should be accounted for in the mean or covariance structure as a fixed or random effect. It was shown that it was not necessary to account for multiple networks although this might have been due to the fact that the

LHA used in this study was dominated by a single large network and differences between networks might have been too small to be considered significant. Finally, spatial relationships among several LHA's were considered and it was shown that it was not necessary to adjust the model for this level of spatial correlation. An explanation for this is that dominant land use was the same in all of the LHA's considered in Chapter 3 meaning that log(TON) can be expected to be similar in terms of mean and variance across multiple networks and LHA's. It is recommended that in future two types of models should be fitted to river network data: one with and one without adjustments for multiple networks or LHA's. The best model should be selected using RMSPE where the aim of the analysis is to make predictions at unobserved locations. Regulatory agencies could benefit from cost savings by improving the predictive performance of statistical models through the inclusion of a more complex covariance structure, removing the need to augment the monitoring network with additional monitoring sites.

Chapter 4 continued the investigation of the spatial covariance function for data collected on river networks through the design and implementation of two novel studies. The first study (sampling study) considered the size of the monitoring network and how understanding of the spatial covariance function and predictions from geostatistical models would be affected by reducing the number of monitoring sites. The second study (covariate study) was designed to compare covariate models - statistical models assuming spatial independence where spatial relationships are captured using covariate information - with covariance models where spatial dependence was captured only through the spatial covariance function.

The sampling study looked at the effect of retaining 90%, 80%, 50%, 20% and 10% of the 687 monitoring sites in the Trent catchment area (LHA28) under four sampling schemes: simple random sampling (random), weighted sampling where weights were the proportion of land that drains to a monitoring site classed as 'arable', proportional stratification sampling and Neyman stratification sampling. Geostatistical models with a hybrid spatial covariance function, combining Euclidean and stream distance, were fitted to 500 sampled subsets of the data and information about covariance function parameter estimates and predictions of winter log(TON) were recorded. The results showed that the range of covariance function parameter estimates increased as the proportion of monitoring sites retained was reduced and estimates from the weighted sampling scheme were different from the other three sampling schemes. The Tail-up range parameter was

particularly affected by the weighted sampling scheme and had a bimodal distribution which was not found for any other parameters. Although it proved difficult to quantify the particular arrangement of monitoring sites that caused this, it was shown that the predictive capability of the models was not affected by the unusual distribution of this parameter.

RMSPE from leave one out cross validation (LOOCV) at observed sites increased as the proportion of monitoring sites retained decreased. This is to be expected but interestingly, the increase in RMSPE was smaller for data sampled using the weighted sampling scheme when 50% or fewer monitoring sites were retained. Prediction error at unobserved sites also increased when the proportion of monitoring sites retained was reduced but prediction error was much higher for samples from the weighted sampling scheme. An explanation for this is that the weighted sampling scheme is more likely to select monitoring sites with high values of log(TON) and so high values will be better predicted by models based on subsets of the data selected using the weighted sampling scheme than the other sampling methods. Correspondingly, prediction error at unobserved locations will be higher under the weighted sampling scheme since models build from these subsets of the data will predict log(TON) to be higher than when the model is built from monitoring sites with a mixture of high and low log(TON) values. The average kriging standard error (AKSE) ratio, the ratio of AKSE calculated for subsets of the data to AKSE for the model based on all monitoring sites, also increased as the number of monitoring sites retained decrease. The increase in AKSE ratio was lowest for the weighted sampling scheme and highest for subsets selected using Neyman stratification. Subsets retaining only 10% of monitoring sites and selected using simple random sampling had AKSE that was, on average, less than 10% greater than when AKSE was calculated based on 687 monitoring sites.

The Trent is a complex, densely monitored river network and regulatory agencies could lower costs by reducing the size of large monitoring networks. This novel sampling study has shown that monitoring could be reduced in the Trent by 50% with an increase in uncertainty of predictions at unobserved locations of up to 10%. Monitoring could be reduced by 90% if an increase in uncertainty of 20% on average and up to 30% were acceptable. In the absence of prior expert knowledge, a sampling strategy could be used to select sites to retain in a reduced monitoring network. It is recommended, based on

the sampling study, that care be taken when choosing an appropriate sampling strategy since weighted sampling leads to predictions of higher value if the variable used to weight the probability of a site being retained is positively correlated with the variable being predicted. The weighted sampling scheme can be thought of as reflecting a worse case scenario if there is an upper safety limit on the variable being predicted but predictions from models based on such a sampling scheme might lead to increased costs as a result of taking unnecessary action to reduce levels of the variable of interest. If the purpose of the analysis is to produce realistic predictions of a variable at unobserved locations then simple random sampling is recommended as a sampling strategy since this is straightforward to implement, is not affected by the choice of variable used for weighting or stratification and provides a smaller increase in the uncertainty of predictions than weighted or stratified sampling methods.

Chapter 4 also contained a novel study designed to investigate if statistical models could be improved by incorporating a complicated stream distance based covariance structure and to assess how statistical models based on covariate information performed compared to models based on a complex spatial covariance structure. Covariates were considered in two ways: the covariate value for the land draining directly to a single monitoring site (RCA covariates) and the covariate value accumulated upstream from monitoring site to source point(s) which reflected the contribution of a covariate to log(TON) across all land that drained into the stream segment upon which a monitoring site is located (accumulated covariates). Additive models were fitted using only RCA covariates (RCA), only accumulated covariates (acc), a mixture of RCA and accumulated covariates (mxd) and no covariates except for a smooth surface based on location (loc). All of these models were fitted assuming spatial independence. Residuals from each of the additive models were extracted and modelled using the `SSN` package (ver Hoef et al., 2014) using a variety of spatial covariance structures: Gaussian Euclidean (Euc), Epanechnikov Tail-up (TU) and a hybrid of these two (hyb). As a comparison, a model with no covariates but a hybrid covariance structure as determined in Chapter 3 was also fitted (ssn) and the residuals extracted. A comparison of the residuals showed that once RCA covariates had been accounted for there was no significant improvement in the residuals either in terms of mean value or variance by modelling the residuals using a Euclidean covariance structure. The variance of the residuals from the model based on accumulated covariates was significantly reduced when modelled with any

spatial covariance function. The variance of the residuals from the mxd model with a mixture of covariates was significantly reduced when modelled using a Tail-up or hybrid covariance function and the variance of the residuals from the additive model based on location only was significantly reduced when spatial covariance was modelled using any of the covariance functions considered. This comparison of residuals along with inspection of variance components and covariance parameter estimates showed that there is a tradeoff between RCA covariates and Euclidean distance based covariance so if covariates are included in the model, the Euclidean covariance function will not lead to further reduction in residual variance. There was some evidence of a similar tradeoff between accumulated covariates and the Tail-up covariance function but this was far weaker than for RCA covariates and Euclidean distance covariance. The residuals from the `SSN` models were comparable in terms of mean and variance to the best fitting residual model. Finally, log(TON) was predicted at observed and unobserved locations and there was no significant difference between predicted values from the mixed covariate and hybrid covariance models, although uncertainty of predictions was approximately four times greater for the hybrid covariance model compared to the mixed covariate model.

This novel comparison of covariates and covariance models could result in cost and efficiency savings for regulatory agencies who carry out routine monitoring in densely monitored river networks. Predictions from the hybrid covariance models gave comparable predictions to the covariate models, although uncertainty was approximately 4 times greater. This means that it is possible for agencies to lower costs by collecting less covariate data and using models with a complex covariance structure, providing the increased uncertainty associated with predictions is acceptable.

Chapter 4 showed that regulatory agencies could achieve cost and efficiency savings by building an appropriate spatial covariance structure into models when aiming to make predictions at unobserved locations. Savings would occur as a result of reducing the size of a densely monitored network or by removing the need to collect and process covariate data.

6.3 Incorporating river network structure into temporal pattern estimation

Chapter 2 concentrated on the estimation of common temporal patterns among time series data while Chapters 3 and 4 investigated, in detail, an appropriate way to model the spatial covariance structure found in river network data at a single time point. Chapter 5 combined the ideas of these three chapters to investigate common temporal patterns in water quality data, accounting for the spatial structure of the river network. Regulatory agencies are coming under increasing pressure to reduce costs and one way to do this is to increase the information that can be extracted from statistical models without a corresponding increase in monitoring. This could be achieved by adapting standard statistical methods to account for spatial structure. Another possibility is to use dimension reduction techniques to determine common temporal patterns - if the data from several monitoring sites can reasonably be reduced to a small number of common patterns this suggests there is redundancy in the network from duplicated information, strengthening the case for reducing the number of monitoring sites.

Chapter 2 considered dynamic factor analysis as a dimension reduction technique but due to computational demands this is not suitable for large numbers of time series. Instead, Chapter 5 focussed on principal components analysis (PCA) which has been shown, in the literature discussed in Chapter 1 and Chapter 5, to be suitable for large numbers of time series and also when there are more variables than observations. Data consisting of observations of a single variable over time at several spatial locations can be accommodated within PCA by performing the analysis in T- or S-mode (Richman, 1986). In T-mode PCA the aim is to identify spatial patterns, common to several time points. Although mostly suitable for climate data where weather systems move in space, T-mode PCA can also be used for river network data to identify changing spatial patterns of water quality variables over time. S-mode PCA aims to assess dominant temporal patterns, usually shared by several monitoring sites. There are examples in the literature of attempts to weight PCA to account for spatial heterogeneity (Harris et al. (2011), Harris et al. (2015)), spatial autocorrelation (Wartenberg (1985), Jombart et al. (2008)) and to account for known structure in the rows and/or columns of the data to estimate patterns independent of known structure (Gabriel and Zamir (1979), Allen et al. (2014)).

Chapter 5 developed a novel adaptation of PCA methodology to account for structure in rows and columns of the data matrix. Specifically, the methodology was developed to include an asymmetric weight matrix reflecting the direction of water flow in a river network as well as the influence of upstream monitoring sites on downstream sites. Adjusting for known dependencies in the data meant that it was appropriate to apply standard PCA to the adjusted data matrix, followed by a back transformation of principal components or loadings depending on whether the spatial weights were applied to rows or columns of the data matrix. The development of the methodology especially highlighted the importance of the orientation of the asymmetric spatial weights matrix. T-mode PCA with rows of the data matrix representing monitoring sites requires the spatial weight matrix to represent water flowing *from* columns *to* rows. For S-mode PCA the data and spatial weights matrices are transposed. The spatial weights must be asymmetric since a symmetric weight matrix would mean that the variance of log(TON) at a monitoring site is assumed to be a weighted combination of the variance and covariance of monitoring sites both upstream and downstream, which is not a reasonable assumption when water flows in one direction.

An application of spatially weighted T-mode PCA to the Trent catchment area concluded that a single spatial pattern was sufficient to describe the pattern of winter log(TON) over 13 years and this pattern largely reflected land use, discussed in Chapter 4. Adjusting T-mode PCA for river network structure meant that the principal components and percentage variance explained by the principal components were only slightly affected, with the percentage variance explained by the first principal component reducing from 89% to 85% and barely noticeable differences in the spatial pattern of the principal components. Differences between the principal components for the unweighted and weighted T-mode PCA were more noticeable in the components explaining the least amount of variance in the data.

Next, S-mode PCA was applied to 21 years of monthly log(TON) data, and missing values were imputed using the methodology in Josse and Husson (2012). S-mode PCA was applied to the unweighted data (PCA_{uw}). Following this, the columns of the data matrix were spatially weighted (PCA_{f}) and the study was completed by an application of S-mode PCA to data where both rows and columns were weighted to account for temporal and spatial structure ($\text{PCA}_{f\rho}$), respectively. PCA was adjusted for temporal structure using a symmetric AR(1) correlation matrix. The first principal component

for PCA_{uw} explained 42% of the variance in the data. This reduced to 38% for PCA_f and 31% for $PCA_{f\rho}$. All subsequent principal components explained less than 10% of the variance in the data. Plots of back transformed principal components and maps of loadings showed that differences between the three analyses were quite subtle, with differences becoming apparent from the third principal component and later components. This study used a complex dataset containing 566 time series covering a geographical area with heterogeneous land use. Chapters 3 and 4 showed that variance of log(TON) in the Trent area is dominated by land use and spatial covariance related to Euclidean distance which might explain why adjusting PCA for spatial covariance based on stream distance and the shape of the river network had a small impact on the PCA results. An explanation for the decrease in variance explained by the weighted principal components is that adjusting for spatial and temporal structure affects the error structure of the data, demonstrated using variance components in Chapter 3, and adjusting PCA using inverse weights means that some of this variance is removed. This suggests that the variance explained by principal components from weighted PCA is the percentage of variance in the data that is not part of the error structure.

Finally, the novel adaptation of PCA methodology was compared to the methods described in Allen et al. (2014). Results between the two methods were shown to be comparable up to multiplication of $\pm$ a constant. The methodology developed in this thesis however specifically shows how an asymmetric weight matrix can be incorporated into PCA to adjust for river network structure, which was not considered in Allen et al. (2014).

Clearer differences between unweighted and weighted PCA might be found if the data are dominated by a very noisy spatial or temporal error structure. This would be in agreement with the results in Allen et al. (2014) who found adjusting PCA using an inverse temporal correlation matrix meant that the signal in the data could be distinguished much more clearly from noise compared to using unweighted PCA. Data collected on river networks can exhibit abrupt changes at confluence points and so adjusting PCA for this type of variability means that spatial variability other than this can be identified. By adjusting for flow connectedness in the river network it might make it possible to more clearly identify abrupt changes in the spatial distribution of log(TON). Regulatory agencies could also extract better estimates of temporal patterns from data exhibiting strong noise such as that found in high frequency data. The use of weighted PCA could

lead to cost savings for regulatory agencies since the identification of common temporal patterns provides evidence of redundancy in the data, strengthening the argument for monitoring network reduction.

This chapter developed a novel adaptation to PCA methodology to account for spatial dependence based on direction and stream distance. The work presented showed how spatial dependency in a river network could be accounted for within the PCA framework using an asymmetric weights matrix where weights reflected flow direction and the relative influence of upstream monitoring sites on downstream sites. It was shown that the orientation of the spatial weight matrix depended on whether the spatial weights were applied to rows or columns and that a symmetric weight matrix as described in Peterson et al. (2007) would not correctly account for flow direction if applied as column weights in PCA.

6.4 Future work

There are several possible extensions to the identification of common patterns in water quality data collected on river networks that has been carried out in this thesis. Possible future work includes extensions of the studies described in this thesis and also the consideration of additional statistical challenges.

The calculation of confidence intervals for curvature developed in Chapter 2 could be improved by further investigation of the distribution theory in Chaudhuri and Marron (1999) and Hannig and Marron (2006) or by taking into account the dependent nature of time series data using Park et al. (2004) and Rondonotti et al. (2007) as the basis for this. Adjustments to the calculation could also be made by including a covariance term in the Taylor expansion of the curvature formula to account for dependence between the first and second derivative terms.

The investigation of appropriate spatial covariance functions to model data collected on river networks in Chapter 3 could be applied to large hydrological areas with quite different land use. This would make it possible to assess whether it is necessary or not to include a term in the model to account for differences between LHA's either as a fixed effect to allow differences in mean levels of log(TON), or as a random effect in the error structure to account for different levels of variance.

The simulation study in Chapter 4 could be extended by considering alternative sampling schemes or by using a variable other than proportion of arable land as weights, perhaps through discussion with the Environment Agency to allow the inclusion of expert knowledge. It would also be interesting to find the best reduced monitoring network of a particular size, following the examples in Ferreyra et al. (2002) and Falk et al. (2014). The covariate study could easily be extended to include more covariates if data were available. It would also be interesting to consider the lagged effect of rainfall as in Burt et al. (1988) if a time series of rainfall were available rather than long term average rainfall. Further statistical development of the covariate study could also include developing models that incorporate smooth relationships between covariates and the response, as well as a stream distance based covariance structure. O'Donnell et al. (2014) and Rushworth et al. (2015) discuss flexible regression for river networks but use the Tail-up weights in the mean structure of the model rather than in the error covariance structure as in ver Hoef et al. (2006).

The work in Chapter 5 could be extended by applying PCA, weighted to adjust for spatial and temporal structure to high frequency temporal data. There was some evidence that PCA weighted for known structure results in the extraction of less noisy signals from the data. The spatial weights matrix could be developed to account for spatial structure related to land use and Euclidean distance - such an adjustment would allow identification of variability directly related to the structure of the river network. The spatial weights accounting for flow direction and connectivity could be applied to data that are less dominated by land use such as elevation data (as in ver Hoef et al. (2014)) or data recorded for an enclosed pipe system where the connectedness structure and stream (or pipe) distance would dominate. PCA adjusted for a directed spatial structure could be developed for traffic flow data, although this would provide many statistical challenges. The spatial weight matrix would need developed to take into account traffic flowing in opposing directions and the incorporation of roundabouts would be particularly challenging. The methodology developed in Chapter 5 could also be developed for branched biological applications, such as the circulation system, where it would be necessary to develop a weight matrix for a closed loop network rather than a network with source and outlet points.

The methodology discussed and developed in this thesis provides many possible opportunities to provide cost savings to regulatory agencies by improving the information

obtained from statistical models using existing spatial information. The identification of clear common temporal patterns of water quality data could provide evidence of redundancy in the monitoring network, strengthening the argument for a reduction in the number of monitoring sites. This thesis explored and developed statistical methods to identify common temporal patterns in water quality data collected on river networks and provides opportunities for future research.

6.5 Software

The underlying code to produce the simulation study in Chapter 4 and to implement the statistical methods developed in Chapter 5, for use in `R`, is openly available from the University of Glasgow Research Data Archive at http://dx.doi.org/10.5525/gla.researchdata.277.

Appendix A

Data processing

This appendix contains further details on data processing steps carried out in order to perform the statistical analyses described in this thesis.

Elevation

Elevation data for England and Wales were obtained from the US Geological Survey. The data product *SRTM3 void filled* with resolution of 3 arc seconds was selected since the data contained were in a format suitable for immediate use in ArcGIS. The data product is based on the WGS84 geographic coordinate system and was converted to British National Grid projected coordinates using the Transverse Mercator projection. This was so that the elevation data were projected to the same coordinate system as the water quality data. Projected coordinates were used so that stream distance could be measured in metres. The elevation data are downloaded as several raster files called tiles and were joined to make a single elevation data file using *mosaic* functionality in ArcGIS. Following this, elevation data were extracted for the LHA's investigated in this thesis using polygon shapefiles (provided by the EA) as a mask.

Polyline shapefile

The Environment Agency for England and Wales provided the polyline shapefile containing the shapes of rivers in the main monitoring network. Within the polyline file a

stream segment is a line joined to other segments by nodes. In order to use this shapefile to calculate stream distances it was necessary to fix existing errors and remove any topological features that would prevent the calculation of stream distances (Peterson and Ver Hoef, 2014). The polyline shapefile is underpinned by a database structure and the STARS toolkit requires that each line segment has only one row entry in the corresponding attributes table.

Errors were fixed using a two stage process and, due to time constraints, only for LHA's that would be investigated in this thesis. LHA's 28 and 34-38 were inspected for errors. During the first stage each stream segment was individually checked for multipart lines. These are stream segments consisting of more than one part joined together. It was often the case that multipart stream segments had a small part of the segment overlapping itself. These overlapping parts were deleted and endpoints joined so that all stream segments consisted of single lines. Other multipart lines were 'exploded' (in ArcGIS terminology) when there were no overlapping sections but the stream segment was stored as multiple parts. Following this each part became a separate stream segment with a separate row entry in the attributes table.

The second staged involved making sure that the network was topologically correct where correct here means that the network satisfies any conditions stated in the STARS tutorial (Peterson, 2011). The main problem encountered was downstream divergence of stream segments which is not permitted. Downstream divergence in the EA data is usually caused by canals cutting across country. A tool exists in ArcGIS to automatically select a route through braided channels based on minimum or maximum distance. Canals caused problems as it was not always clear which direction the water flowed and in most cases the whole stretch of canal was removed i.e. if part of a canal needed removed due to downstream divergence then every line segment that was part of that canal was removed. This was to ensure that no gaps were created and that remaining rivers were fully connected. In this second stage junctions were added at confluence points (where two stream segment join to flow into a single stream segment) if they did not already exist. Junctions are necessary to ensure that nodes are correctly classified and that only a single node is placed at each confluence.

A landscape network (Peterson and Ver Hoef, 2014) storing the stream distance and flow direction information was created once these checks were carried out and nodes

were placed at each end of the stream segments and the classification of each node was checked. At this point the main problem remaining was nodes being placed part way along a stream segment where the end of two adjoining segments overlapped. In this case one segment was shortened so that the two adjoining segments met at their endpoints. Another problem was converging streams where two line segments converged and did not flow into a common downstream segment. This was solved by 'snapping' the ends of each stream segment meeting at a confluence to the adjoining downstream segment.

The landscape network was re-created once these problems were resolved and inspected again to ensure the network was topologically correct. Once it was established that nodes were correctly classified, monitoring sites were snapped to the nearest stream segment. Since rivers have width but are represented as a line the geographical coordinates for monitoring sites did not always fall exactly on a stream segment when mapped in ArcGIS. Snapping the sites to the nearest line makes it possible to calculate the stream distance between monitoring stations. A 2.5km radius was used as tolerance for snapping and any monitoring sites that were further than 2.5km from the nearest stream segment were removed from the data set. This applied only to a single monitoring site that was originally situated on a canal.

Population

Population data were obtained from http://www.ons.gov.uk/ons/index.html. Three sources of information were used to calculate population density: (1) Population count of all ages for each local authority district (LAD) in England and Wales from mid 2012 (mid 2012 population estimates for England and wales), (2) a shapefile containing boundary information for LAD's from 2012 (Local authority district (GB) Dec 2012 Boundaries (Full Extent)) and (3) Standard Area Measurements (SAM) 2012 (SAM LAD 2012 UK) containing the area of each LAD in hectares (ha). ArcGIS was used to calculate population density for each monitoring station and prediction location.

The table of area measurements was linked to the shapefile of LAD boundaries using an identifying code (LAD12CD) and a variable representing area in square kilometers (km^2) was calculated by dividing the LAD area (AreaLHect) by 100. A field for population count was added to the LAD shapefile and linked to the table of population

counts of all ages using the LAD identifying code. The polygon LAD boundary file was converted to raster format with cell size 49.1855526*49.1855526 which is equal to the cell size in the elevation raster file. The zonal statistics tool in ArcGIS was used to calculate population/km^2 for each RCA. This value was then linked to stream segments and corresponding monitoring sites using the methods described in Peterson (2011). To obtain total population contributing to each monitoring site the RCA population value was multiplied by the area of the RCA in km^2 and the RCA totals were accumulated downstream between source and monitoring site, following the steps in Peterson (2011).

Rainfall

Rainfall data were provided by the EA and is a long term average annual depth in mm for 1961 to 1990. The data are in a polygon shapefile with each polygon representing 1km^2 and this shapefile was converted to raster format using the same cell size as the elevation data. There were 2 missing values from the resulting rainfall grid file and these were replaced with the average annual depth recorded for 1941 to 1970. The RCA value was calculated as the average depth (mm) of the rainfall value from all of the cells within each RCA. In order to obtain an estimate of total rainfall contributing to a monitoring site it was necessary to calculate rainfall volume within each RCA and use the accumulation tools described in Peterson (2011) to obtain rainfall volume over all contributing land.

Rainfall volume is calculated as depth(mm)$\times$area(m^2) $\times$ 0.001 (calculation information kindly provided by John Douglass of the Environment Agency).

Livestock and Crops

Livestock and crop data were provided by the EA and are taken from the ADAS 2010 agricultural census. The livestock data show the number of animals, broken down by species, in square polygons with area 1km^2 . Table 4.2 shows all of the species included in the agricultural census. In order to get and RCA value for livestock, the polygon file was converted to raster format using the same cell size as for the elevation data and the average taken of all of the cells within an RCA. The resulting value is the number

of livestock per km^2 . In this covariate study all of the livestock were counted together rather than having separate values for cows, pigs, sheep etc. The exception to this was chickens since large numbers of chickens can be accommodated in quite a small space and the chicken number hugely inflated the livestock numbers. Chicken numbers were therefore calculated separately from the livestock numbers. The total number of livestock/chickens over all contributing land was calculated by multiplying the RCA value for livestock/chickens by the area of the RCA in km^2 and summed from source to monitoring point using the tools described in Peterson (2011).

Crop data was also included in the agcensus dataset and like the livestock data is a polygon file of squares with area $1km^2$. The data gives the number of hectares in each square that are used for each crop covered in the census, with $1km^2$ =100ha. This was processed in the same way as for the livestock data by converting the polygons to raster format and calculating the average value for all of the cells within each RCA. The number of hectares used for all crops (see Table 4.2), excluding grass, was divided by 100 to give area of crops per km^2 . The total area of crops for contributing area was calculated by multiplying the RCA value by the area of the RCA in km^2 and summing this total for all RCA's between source and monitoring site. The same steps were taken to obtain equivalent values for area of grass. Grass was calculated separately since land used for grass is correlated with livestock numbers.

Land cover

Land cover information was provided by the EA and obtained from the the LCM2007 dataset (see Morton et al. (2011) for full details of this dataset). LCM2007 is a polygon shapefile and gives the proportion of area in land parcels that is classified as each of 23 land cover categories as well as the dominant classification in each land parcel. The land parcels are much smaller than the RCA's, with each RCA containing many land parcels. A land cover classification was obtained by converting the polygon shapefile to raster format using the same cell size as for the elevation data. The number of cells for each land cover category were counted and RCA land cover was allocated the category with the greatest number of cells. Land cover was not calculated for total contributing area. The RCA value was attributed to the monitoring sites following the steps in Peterson (2011).

Land cover category is related to number of livestock (land cover=grass), area of RCA used for crops (land cover=arable) and population density (land cover=urban) and therefore models could be fitted either with land cover category as a factor variable or with livestock, crops and population aggregated to RCA level. The proportion of land used for grass and crops would also have needed further processing before being used in the covariate model since these are compositional data. A more meaningful interpretation of the model is possible therefore when land cover category is included rather than using compositional data.

Bibliography

Abdi, H. and L. J. Williams (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics 2*(4), 433–459.

Abraham, G. and M. Inouye (2014). Fast principal component analysis of large-scale genome-wide data. *PloS one 9*(4), e93766.

Abramowitz, M. and I. A. Stegun (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Number 55. Courier Corporation.

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. *Second International Symposium on Information Theory, Akademia Kiado*, 267281.

Allen, G. I., L. Grosenick, and J. Taylor (2014). A generalized least-square matrix decomposition. *Journal of the American Statistical Association 109*(505), 145–159.

Alonso, A. M., C. García-Martos, J. Rodríguez, and M. J. Sánchez. (2011). Seasonal dynamic factor analysis and bootstrap inference: application to electricity market forecasting. *Technometrics 53(2)*, 137–151.

Amna, T., M. S. Hassan, N. A. Barakat, D. R. Pandeya, S. T. Hong, M.-S. Khil, and H. Y. Kim (2012). Antibacterial activity and interaction mechanism of electrospun zinc-doped titania nanofibers. *Applied microbiology and biotechnology 93*(2), 743–751.

Andrade, A. G., J. C. Polese, L. A. Paolucci, H. Menzel, and L. F. Teixeira-Salmela (2014). Functional data analyses for the assessment of joint power profiles during gait of stroke subjects. *Journal of applied biomechanics 30*(2), 348–352.

Andrés Houseman, E. (2005). A robust regression model for a first-order autoregressive time series with unequal spacing: application to water monitoring. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(4), 769–780.

Baines, S. B., K. E. Webster, T. K. Kratz, S. R. Carpenter, and J. J. Magnuson (2000). Synchronous behavior of temperature, calcium, and chlorophyll in lakes of northern wisconsin. *Ecology 81*(3), 815–825.

Baldwin, M. P., D. B. Stephenson, and I. T. Jolliffe (2009). Spatial weighting and iterative projection methods for eofs. *Journal of Climate 22*(2), 234–243.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2004). *Hierarchical modeling and analysis for spatial data.* Chapman and Hall/CRC; Boca Raton; FL, London.

Barreira, S. and R. H. Compagnucci (2011). Spatial fields of antarctic sea-ice concentration anomalies for summer-autumn and their relationship to southern hemisphere atmospheric circulation during the period 1979-2009. *Annals of Glaciology 52*(57), 140–150.

Bartram, J. and R. Ballance (1996). *Water quality monitoring: a practical guide to the design and implementation of freshwater quality studies and monitoring programmes.* CRC Press.

Bengraïne, K. and T. F. Marhaba (2003). Using principal component analysis to monitor spatial and temporal changes in water quality. *Journal of Hazardous Materials 100*(1), 179–195.

Blangiardo, M., M. Cameletti, G. Baio, and H. Rue (2013). Spatial and spatio-temporal models with r-inla. *Spatial and spatio-temporal epidemiology 7*, 39–55.

Blenckner, T., R. Adrian, D. M. Livingstone, E. Jennings, G. A. Weyhenmeyer, D. George, T. Jankowski, M. Järvinen, C. N. Aonghusa, T. Nõges, et al. (2007). Large-scale climatic signatures in lakes across europe: A meta-analysis. *Global Change Biology 13*(7), 1314–1326.

Bowman, A. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations.* Oxford University Press, New York.

Bowman, A., C. Ferguson, D. Lee, A. Magdalina, and E. Scott (2010). Spatiotemporal modelling of nitrate and phosphorous for river catchments. Science Report - SC080041/SR, Environment Agency.

Brockwell, P. J. and R. A. Davis (1996). *Introduction to time series analysis.* Springer Texts in Statistics, Springer Verlag, New York.

Burt, T., B. Arkell, S. Trudgill, and D. Walling (1988). Stream nitrate levels in a small catchment in south west england over a period of 15 years (1970-1985). *Hydrological Processes 2*(3), 267–284.

Burt, T., N. Howden, F. Worrall, and M. Whelan (2008). Importance of long-term monitoring for detecting environmental change: lessons from a lowland river in south east england. *Biogeosciences 5*(6), 1529–1535.

Burt, T., N. Howden, F. Worrall, M. Whelan, and M. Bieroza (2010). Nitrate in united kingdom rivers: Policy and its outcomes since 1970. *Environmental science & technology 45*(1), 175–181.

Caliman, A., L. Carneiro, J. Santangelo, R. Guariento, A. Pires, A. Suhett, L. Quesado, V. Scofield, E. Fonte, P. Lopes, et al. (2010). Temporal coherence among tropical coastal lagoons: a search for patterns and mechanisms. *Brazilian Journal of Biology 70*(3), 803–814.

Cameletti, M., F. Lindgren, D. Simpson, and H. Rue (2011). Spatio-temporal modeling of particulate matter concentration through the spde approach. *Submitted*.

Carey, S. K., D. Tetzlaff, J. Seibert, C. Soulsby, J. Buttle, H. Laudon, J. McDonnell, K. McGuire, D. Caissie, J. Shanley, et al. (2010). Inter-comparison of hydro-climatic regimes across northern catchments: synchronicity, resistance and resilience. *Hydrological Processes 24*(24), 3591–3602.

Carvalho, L., C. Miller, B. Spears, I. Gunn, H. Bennion, A. Kirika, and L. May (2011). Water quality of loch leven: responses to enrichment, restoration and climate change. In *Loch Leven: 40 years of scientific research*, pp. 35–47.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research 1*(2), 245–276.

Chaudhuri, P. and J. Marron (2002). Curvature vs. slope inference for features in nonparametric curve estimates. *Unpublished manuscript*.

Chaudhuri, P. and J. S. Marron (1999). Sizer for exploration of structures in curves. *Journal of the American Statistical Association 94*(447), 807–823.

Chen, H. (2002). Principal component analysis with missing data and outliers. *URL: http://www. cmlab. csie. ntu. edu. tw/˜ cyy/learning/papers/PCA_Tutorial. pdf*.

Cheng, Q., G. Bonham-Carter, W. Wang, S. Zhang, W. Li, and X. Qinglin (2011). A spatially weighted principal component analysis for multi-element geochemical data for mapping locations of felsic intrusions in the gejiu mineral district of yunnan, china. *Computers & Geosciences 37*(5), 662–669.

Chilès, J.-P. and P. Delfiner (1999). *Modeling Spatial Uncertainty.* John Wiley & Sons, Inc; New York.

Chilundo, M., P. Kelderman, and J. O'Keeffe (2008). Design of a water quality monitoring network for the Limpopo river basin in Mozambique. *Physics and Chemistry of the Earth, Parts A/B/C 33*(8-13), 655 – 665.

Clement, L., O. Thas, P. A. Vanrolleghem, J.-P. Ottoy, et al. (2006). Spatio-temporal statistical models for river monitoring networks. *Water Science & Technology 53*(1), 9–15.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association 74*(368), 829–836.

Cleveland, W. S. and S. J. Devlin (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association 83*(403), 596–610.

Cortina-Borja, M., M. Geraci, L. Griffiths, C. Rich, C. Dezateux, et al. (2011). Modelling accelerometer data from 7-year old british children using functional analysis of variance. *Journal of Epidemiology and Community Health 65*(Suppl 2), A26–A27.

Cox, D. D. and J. S. Lee (2008). Pointwise testing with functional data using the westfall–young randomization method. *Biometrika 95*(3), 621–634.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology 17*(5), 563–586.

Cressie, N., J. Frey, B. Harch, and M. Smith (2006). Spatial prediction on a river network. *American Statistical Association and the International Biometric Society Journal of Agricultural, Biological, and Environmental Statistics 11*, 127–150.

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

Cressie, N. and M. Pavlicová (2002). Calibrated spatial moving average simulations. *Statistical Modelling 2*(4), 267–279.

Cressie, N. and C. K. Wikle (2011). *Statistics for spatio-temporal data*. John Wiley & Sons.

Cressie, N. A. C. (1993). *Statistics for spatial data*. John Wiley and Sons, Inc; USA.

Curriero, F. C. (2007). On the use of non-euclidean distance measures in geostatistics. *Mathematical Geology 38*(8), 907–926.

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

de Boor, C. (2001). *A practical guide to splines, revised Edition, Vol. 27 of Applied Mathematical Sciences*. Springer-Verlag, Berlin.

Delgado, M. A. (1993). Testing the equality of nonparametric regression curves. *Statistics & probability letters 17*(3), 199–204.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Demšar, U., P. Harris, C. Brunsdon, S. Fotheringham, and S. McLoone (2013). Principal component analysis on spatial data: An overview. *Annals of the Association of American Geographers 103*(1), 106–128.

Dette, H. and N. Neumeyer (2001). Nonparametric analysis of covariance. *The Annals of Statistics 29*(5), 1361–1400.

Diaz-Ramos, S., D. Stevens, and A. Olsen (1996). *EMAP statistical methods manual*. (EPA/620/R-96/002 ed.). Office of Research and Development, NHEERLWED, Corvallis, Oregon: U.S. Environmental Protection Agency.

Diggle, P. and P. Ribeiro (2007). *Model-based Geostatistics*. Springer.

Diggle, P. J. et al. (1983). *Statistical analysis of spatial point patterns*. Academic press.

Dixon, W., G. Smyth, and B. Chiswell (1999). Optimized selection of river sampling sites. *Water Research 33*(4), 971–978.

Djurdjevic, M., J. Sokolowski, and Z. Odanovic (2012). Determination of dendrite coherency point characteristics using first derivative curve versus temperature. *Journal of thermal analysis and calorimetry 109*(2), 875–882.

Dobbie, M., B. Henderson, and D. Stevens (2008). Sparse sampling: spatial design for monitoring stream networks. *Statistics Surveys 2*, 113–153.

Dray, S., S. Saïd, and F. Débias (2008). Spatial ordination of vegetation data using a generalization of wartenberg's multivariate spatial correlation. *Journal of Vegetation Science 19*(1), 45–56.

Eastoe, E. F., C. J. Halsall, J. E. Heffernan, and H. Hung (2006). A statistical comparison of survival and replacement analyses for the use of censored data in a contaminant air database: a case study from the canadian arctic. *Atmospheric Environment 40*(34), 6528–6540.

EEA (2015). The european environment - state and outlook 2015: synthesis report. *European Environment Agency*.

Ehrendorfer, M. (1987). A regionalization of austria's precipitation climate using principal component analysis. *Journal of Climatology 7*(1), 71–89.

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

Erzini, K. (2005). Trends in ne atlantic landings (southern portugal): identifying the relative importance of fisheries and environmental variables. *Fisheries Oceanography 14*(3), 195–209.

European Parliament (1991). Council directive of 12 december 1991 concerning the protection of waters against pollution caused by nitrates from agricultural sources (91/676/eec). *Official Journal of the European Communities 327*, 1–13.

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

Fahrmeir, L. and S. Lang (2001). Bayesian inference for generalized additive mixed models based on markov random field priors. *Applied statistics*, 201–220.

Falk, M., J. McGree, and A. Pettitt (2014). Sampling designs on stream networks using the pseudo-Bayesian approach. *Environmental and ecological statistics*. In press.

Fan, J. and I. Gijbels (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 2008–2036.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, Volume 66. CRC Press.

Fan, J. and S.-K. Lin (1998). Test of significance when data are curves. *Journal of the American Statistical Association 93*(443), 1007–1021.

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics 39*(3), 254–261.

Ferguson, C., L. Carvalho, E. Scott, A. Bowman, and A. Kirika (2008). Assessing ecological responses to environmental change using statistical models. *Journal of Applied Ecology 45*(1), 193–203.

Ferreyra, R., H. Apezteguía, R. Sereno, and J. Jones (2002). Reduction of soil water spatial sampling density using scaled semivaiograms and simulated annealing. *Geoderma 110*, 265–289.

Fölster, J., E. Göransson, K. Johansson, and A. Wilander (2005). Synchronous variation in water chemistry for 80 lakes in southern sweden. *Environmental monitoring and assessment 102*(1-3), 389–403.

Francisco-Fernandez, M. and J. D. Opsomer (2005). Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *Canadian Journal of Statistics 33*(2), 279–295.

Frichot, E., S. Schoville, G. Bouchard, and O. François (2012). Correcting principal component maps for effects of spatial autocorrelation in population genetic data. *Frontiers in genetics 3*.

Friedland, K. D. and J. A. Hare (2007). Long-term trends and regime shifts in sea surface temperature on the continental shelf of the northeast united states. *Continental Shelf Research 27*(18), 2313–2328.

Friedman, J. and B. Silverman (1989). Flexible parsimonious smoothing and additive modelling. *Technometrics 31*(1), 3–21.

Frontier, S. (1976). Étude de la décroissance des valeurs propres dans une analyse en composantes principales: Comparaison avec le moddle du bâton brisé. *Journal of Experimental Marine Biology and Ecology 25*(1), 67–75.

Fuentes, M., A. Chaudhuri, and D. Holland (2007). Bayesian entropy for spatial sampling design of environmental data. *Environmental and Ecological Statistics 14*(3), 323–340.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika 58*(3), 453–467.

Gabriel, K. R. and S. Zamir (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics 21*(4), 489–498.

Ganguli, B. and M. Wand (2007). Feature significance in generalized additive models. *Statistics and Computing 17*(2), 179–192.

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

Garreta, V., P. Monestiez, and J. M. Ver Hoef (2010). Spatial modelling and prediction on river networks: up model, down model or hybrid? *Environmetrics 21*(5), 439–456.

George, D., J. Talling, and E. Rigg (2000). Factors influencing the temporal coherence of five lakes in the english lake district. *Freshwater Biology. 43*, 449–461.

Gerber, P., C. Opio, and H. Steinfeld (2007). Poultry production and the environmenta review.

Geweke, J. (1976). *The dynamic factor analysis of economic time series models.* University of Wisconsin.

Ghanbari, R. N. and H. R. Bravo (2011). Coherence among climate signals, precipitation, and groundwater. *Groundwater 49*(4), 476–490.

Giannitrapani, M., A. Bowman, M. Scott, and R. Smith (2006). Sulphur dioxide in europe: Statistical relationships between emissions and measured concentrations. *Atmospheric Environment 40*(14), 2524–2532.

Girimurugan, S. and E. Chicken (2013). Wavelet analysis of variance (wanova). *Technical Reports of the FSU. Department of Statistics M 1012*, 1–7.

Gorrostieta, C., J. Ortega, A. J. Quiroz, and G. H. Smith (2014). Characterization of storm wave asymmetries with functional data analysis. *Environmental and ecological statistics 21*(2), 263–283.

Gottschalk, L. (1993). Correlation and covariance of runoff. *Stochastic hydrology and hydraulics 7*(2), 85–101.

Gower, J. C., S. G. Lubbe, and N. J. Le Roux (2011). *Understanding biplots*. John Wiley & Sons.

Green, P. J. and B. W. Silverman (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.

Grinsted, A., J. C. Moore, and S. Jevrejeva (2004). Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics 11*(5/6), 561–566.

Guo, R., M. Ahn, and H. Z. Hongtu Zhu (2015). Spatially weighted principal component analysis for imaging classification. *Journal of Computational and Graphical Statistics 24*(1), 274–296.

Haggarty, R. A. (2012). *Evaluation of sampling and monitoring designs for water quality*. Ph. D. thesis, University of Glasgow.

Hall, P. and J. D. Hart (1990). Bootstrap test for difference between means in nonparametric regression. *Journal of the American Statistical Association 85*(412), 1039–1049.

Hannachi, A., I. Jolliffe, and D. Stephenson (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology 27*(9), 1119–1152.

Hannig, J. and J. Marron (2006). Advanced distribution theory for sizer. *Journal of the American Statistical Association 101*(474), 484–499.

Hanson, R., M. Newhouse, and M. Dettinger (2004). A methodology to asess relations between climatic variability and variations in hydrologic time series in the southwestern united states. *Journal of Hydrology 287*(1), 252–269.

Härdle, W. and J. S. Marron (1990). Semiparametric comparison of regression curves. *The Annals of Statistics 18*(1), 63–89.

Hari, R. E., D. M. Livingstone, R. Siber, P. Burkhardt-Holm, and H. Guettinger (2006). Consequences of climatic change for water temperature and brown trout populations in alpine rivers and streams. *Global Change Biology 12*(1), 10–26.

Harrell Jr, F. E. and with contributions from Charles Dupont and many others. (2014). *Hmisc: Harrell Miscellaneous*. R package version 3.14-4.

Harris, P., C. Brunsdon, and M. Charlton (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science 25*(10), 1717–1736.

Harris, P., A. Clarke, S. Juggins, C. Brunsdon, and M. Charlton (2015). Enhancements to a geographically weighted principal component analysis in the context of an application to an environmental data set. *Geographical Analysis 47*(2), 146–172.

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge university press.

Hassan, M., J. Terrien, B. Karlsson, and C. Marque (2010). Application of wavelet coherence to the detection of uterine electrical activity synchronization in labor. *IRBM 31*(3), 182 – 187.

Hastie, T. and R. Tibshirani (1990). *Generalized additive models*. Chapman and Hall; London.

Hausman, R. (1982). Constrained multivariate analysis. *Optimization in statistics 137*.

Helsel, D. (2010). Much ado about next to nothing: incorporating nondetects in science. *Annals of occupational hygiene 54*(3), 257–262.

Helsel, D. R. (1990). Less than obvious-statistical treatment of data below the detection limit. *Environmental Science & Technology 24*(12), 1766–1774.

Helsel, D. R. (2012). Reporting limits. *Statistics for Censored Environmental Data Using Minitab® and R, Second Edition*, 22–36.

Helsel, D. R. and T. A. Cohn (1988). Estimation of descriptive statistics for multiply censored water quality data. *Water Resources Research 24*(12), 1997–2004.

Henderson, B. (2006). Exploring between site differences in water quality trends: a functional data analysis approach. *Environmetrics 17*(1), 65–80.

Herlihy, A., D. Larsen, S. Paulsen, N. Urquhart, and B. Rosenbaum (2000). Designing a spatially balanced, randomized site selection process for regional stream surveys: The emap mid-atlantic pilot study. *Environmental monitoring and assessment 63*(1), 95–113.

Hidalgo-Muñoz, J., D. Argüeso, S. Gámiz-Fortis, M. Esteban-Parra, and Y. Castro-Díez (2011). Trends of extreme precipitation and associated synoptic patterns over the southern iberian peninsula. *Journal of Hydrology 409*(1), 497–511.

Higham, N. J. (2002). Computing the nearest correlation matrixa problem from finance. *IMA journal of Numerical Analysis 22*(3), 329–343.

Holmes, E. E., E. J. Ward, and K. Wills (2012). Marss: Multivariate autoregressive state-space models for analyzing time-series data. *The R Journal 4*(1), 11–19.

Hooda, P., A. Edwards, H. Anderson, and A. Miller (2000). A review of water quality concerns in livestock farming areas. *Science of The Total Environment 250*, 143 – 167.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology 24*(6), 417.

Howden, N. J., T. P. Burt, F. Worrall, and M. J. Whelan (2011). Monitoring fluvial water chemistry for trend detection: hydrological variability masks trends in datasets covering fewer than 12 years. *Journal of Environmental Monitoring 13*(3), 514–521.

Husson, F. and J. Josse (2015). *missMDA: Handling Missing Values with Multivariate Data Analysis*. R package version 1.9.

Huth, R. (1996). Properties of the circulation classification scheme based on the rotated principal component analysis. *Meteorology and Atmospheric Physics 59*(3-4), 217–233.

Ignaccolo, R., S. Ghigo, and E. Giovenali (2008). Analysis of air quality monitoring networks by functional clustering. *Environmetrics 19*(7), 672–686.

Ilin, A. and T. Raiko (2010). Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research 11*, 1957–2000.

James, G. M. and C. A. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association 98*(462), 397–408.

Jiménez, P., E. García-Bustamante, J. González-Rouco, F. Valero, J. Montávez, and J. Navarro (2008). Surface wind regionalization in complex terrain. *Journal of Applied Meteorology and Climatology 47*(1), 308–325.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

Jolliffe, I. T., N. T. Trendafilov, and M. Uddin (2003). A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics 12*(3), 531–547.

Jolliffe, I. T. and M. Uddin (2000). The simplified component technique: an alternative to rotated principal components. *Journal of Computational and Graphical Statistics 9*(4), 689–710.

Jombart, T., S. Devillard, A. Dufour, and D. Pontier (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity 101*(1), 92–103.

Josse, J. and F. Husson (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique 153*(2), 79–99.

Journel, A. G. and C. J. Huijbregts (1978). *Mining Geostatistics.* Academic Press; London.

Kahya, E., S. Kalaycı, and T. C. Piechota (2008). Streamflow regionalization: case study of turkey. *Journal of Hydrologic Engineering 13*(4), 205–214.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika 23*(3), 187–200.

Kao, J.-J., P.-H. Li, C.-L. Lin, and W.-H. Hu (2008). Siting analyses for water quality sampling in a catchment. *Environmental Monitoring and Assessment 139*(1-3), 205–215.

Kaufman, C. G., S. R. Sain, et al. (2010). Bayesian functional {ANOVA} modeling using gaussian process prior distributions. *Bayesian Analysis 5*(1), 123–149.

Khalil, B. and T. B. M. J. Ouarda (2009). Statistical approaches used to assess and redesign surface water-quality-monitoring networks. *Journal of Environmental Monitoring 11*, 1915–1929.

Ku, W., R. H. Storer, and C. Georgakis (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and intelligent laboratory systems 30*(1), 179–196.

Kulasekera, K. (1995). Comparison of regression curves using quasi-residuals. *Journal of the American Statistical Association 90*(431), 1085–1093.

Kulasekera, K. and J. Wang (1997). Smoothing parameter selection for power optimality in testing of regression curves. *Journal of the American Statistical Association 92*(438), 500–511.

Laaha, G., J. O. Skøien, and G. Blöschl (2012). Comparing geostatistical models for river networks. In *Geostatistics Oslo 2012*, pp. 543–553. Springer.

Lachaux, J.-P., A. Lutz, D. Rudrauf, D. Cosmelli, M. L. V. Quyen, J. Martinerie, and F. Varela (2002). Estimating the time-course of coherence between single-trial brain signals: an introduction to wavelet coherence. *Neurophysiologie Clinique/Clinical Neurophysiology 32*(3), 157 – 174.

Lang, S. and A. Brezger (2004). Bayesian p-splines. *Journal of computational and graphical statistics 13*(1), 183–212.

Lansac-Tôha, F. A., L. M. Bini, L. F. M. Velho, C. C. Bonecker, E. M. Takahashi, and L. C. Vieira (2008). Temporal coherence of zooplankton abundance in a tropical reservoir. *Hydrobiologia 614*(1), 387–399.

Lee, L. (2012). *NADA: Nondetects And Data Analysis for environmental data*. R package version 1.5-4.

Lindgren, F., H. Rue, and J. Lindstrom (2011). An explicit link between gaussian fields and gaussian markov random fields: the spde approach (with discussion). *J R Statist Soc B 73(4)*, 423–498.

Little, L. S., D. Edwards, and E. E. Porter (1997). Kriging in estuaries: As the crow flies, or as the fish swims?. *Journal of Experimental Marine Biology and Ecology 213*, 1–11.

Little, R. and D. Rubin (1987). *Statistical Analysis with Missing Data*. Wiley.

Lopes, H. F., D. Gamerman, and E. Salazar (2011). Generalized spatial dynamic factor models. *Computational Statistics & Data Analysis 55*(3), 1319–1330.

Lopes, H. F., E. Salazar, D. Gamerman, et al. (2008). Spatial dynamic factor analysis. *Bayesian Analysis 3*(4), 759–792.

Lord, E. and S. Antony (2000). Magpie: a modelling framework for evaluating nitrate losses at national and catchment scales. *Soil Use and Management 16(issue supplement s1)*, 167–174.

Lounici, K. et al. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli 20*(3), 1029–1058.

Lu, X. and J. Marron (2013). Principal nested spheres for time warped functional data analysis. *arXiv preprint arXiv:1304.6789*.

Magnuson, J., B. Benson, and T. Kratz (1990). Temporal coherence in the limnology of a suite of lakes in wisconsin, u.s.a. *Freshwater Biology. 23*, 145–159.

Magnuson, J. J., T. K. Kratz, B. J. Benson, and K. E. Webster (2006). Coherent dynamics among lakes. *Long-term dynamics of lakes in the landscape: long-term ecological research on north temperate lakes, Oxford University Press, Oxford, UK*, 89–106.

Maraun, D. and J. Kurths (2004). Cross wavelet analysis: significance testing and pitfalls. *Nonlinear Processes in Geophysics 11*(4), 505–514.

Marron, J. and J. de Uña-Álvarez (2004). Sizer for length biased, censored density and hazard estimation. *Journal of Statistical Planning and Inference 121*(1), 149–161.

Marron, J. and A. Tsybakov (1995). Visual error criteria for qualitative smoothing. *Journal of the American Statistical Association 90(430)*, 499–507.

Marron, J. and J. T. Zhang (2005). Sizer for smoothing splines. *Computational Statistics 20*(3), 481–502.

Matheron, G. (1963). Principles of geostatistics. *Economic geology 58*(8), 1246–266.

McIntyre, A. C., M. L. Bilyk, A. Nordon, G. Colquhoun, and D. Littlejohn (2011). Detection of counterfeit scotch whisky samples using mid-infrared spectrometry with an attenuated total reflectance probe incorporating polycrystalline silver halide fibres. *Analytica chimica acta 690*(2), 228–233.

McMullan, A., A. Bowman, and E. Scott (2007). Water quality in the river clyde: a case study of additive and interaction models. *Environmetrics 18*(5), 527–539.

Meyer, P. L. (1970). *Introductory probability and statistical application.* Addison-Wesley, Philippines.

Miller, C., A.-M. Magdalina, R. Willows, A. Bowman, E. Scott, D. Lee, C. Burgess, L. Pope, F. Pannullo, and R. Haggarty (2014). Spatiotemporal statistical modelling of long-term change in river nutrient concentrations in england & wales. *Science of the Total Environment 466-7*, 914–923.

Minas, C., S. J. Waddell, and G. Montana. (2011). Distance-based differential analysis of gene curves. *Bioinformatics 27(22)*, 3135–3141.

Molenaar, P. C. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika 50*(2), 181–202.

Molenaar, P. C., J. G. De Gooijer, and B. Schmitz (1992). Dynamic factor analysis of nonstationary multivariate time series. *Psychometrika 57*(3), 333–349.

Monestiez, P., J.-S. Bailly, P. Lagacherie, and M. Voltz (2005). Geostatistical modelling of spatial processes on directed trees: application to fluvisol content. *Geoderma 128*, 179–191.

Money, E., G. P. Carter, and S. M. L. (2009). Using river distances in the space/time estimation of dissolved oxygen along two impaired river networks in new jersey. *Water Research 43*, 1948–1958.

Monk, W. A., P. J. Wood, D. M. Hannah, and D. A. Wilson (2007). Selection of river flow indices for the assessment of hydroecological change. *River Research and Applications 23*(1), 113–122.

Morton, D., C. Rowland, C. Wood, L. Meek, C. Marston, G. Smith, R. Wadsworth, and I. Simpson (2011). Final report for lcm2007-the new uk land cover map. countryside survey technical report no 11/07.

Müller, H.-G. (2012). *Nonparametric regression analysis of longitudinal data*, Volume 46. Springer Science & Business Media.

Muñoz-Carpena, R., A. Ritter, and Y. Li (2005). Dynamic factor analysis of groundwater quality trends in an agricultural area adjacent to everglades national park. *Journal of Contaminant Hydrology 80*(1), 49–70.

Murphy, R., F. Curriero, and W. Ball (2010). Comparison of spatial interpolation methods for water quality evaluation in the Chesapeake Bay. *Journal of Environmental Engineeering 136*, 160–171.

Neal, C., H. P. Jarvie, M. Neal, L. Hill, and H. Wickham (2006). Nitrate concentrations in river waters of the upper thames and its tributaries. *Science of the Total Environment 365*(1), 15–32.

Neal, C., R. Skeffington, M. Neal, R. Wyatt, H. Wickham, L. Hill, and N. Hewitt (2004). Rainfall and runoff water quality of the pang and lambourn, tributaries of the river thames, south-eastern england. *Hydrology and Earth System Sciences Discussions 8*(4), 601–613.

Neal, R. A. and I. D. Phillips (2009). Summer daily precipitation variability over the east anglian region of great britain. *International Journal of Climatology 29*(11), 1661–1679.

Neill, M. (1989). Nitrate concentrations in river waters in the south-east of ireland and their relationship with agricultural practice. *Water Research 23*(11), 1339–1355.

Nelson, P. R., P. A. Taylor, and J. F. MacGregor (1996). Missing data methods in pca and pls: Score calculations with incomplete observations. *Chemometrics and intelligent laboratory systems 35*(1), 45–65.

Neumeyer, N., H. Dette, et al. (2003). Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics 31*(3), 880–920.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 558–625.

O'Donnell, D., A. Rushworth, A. W. Bowman, E. Marian Scott, and M. Hallard (2014). Flexible regression models over river networks. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 63*(1), 47–63.

Oliveira, M., R. M. Crujeiras, and A. Rodríguez-Casal (2014). Circsizer: an exploratory tool for circular data. *Environmental and ecological statistics 21*(1), 143–159.

Orr, H. G., G. L. Simpson, S. Clers, G. Watts, M. Hughes, J. Hannaford, M. J. Dunbar, C. L. Laizé, R. L. Wilby, R. W. Battarbee, et al. (2015). Detecting changing river temperatures in england and wales. *Hydrological Processes 29*(5), 752–766.

Pace, M. L. and J. J. Cole (2002). Synchronous variation of dissolved organic carbon and color in lakes. *Limnology and Oceanography 47*(2), 333–342.

Pardo-Fernández, J. C., I. Van Keilegom, W. González-Manteiga, et al. (2007). Testing for the equality of k regression curves. *Statistica Sinica 17*(3), 1115.

Park, C., V. A, H. J, and K. K. (2009). Sizer analysis for the comparison of time series. *Journal of Statistical Planning and Inference 139*, 3974–3988.

Park, C. and K.-H. Kang (2008). Sizer analysis for the comparison of regression curves. *Computational Statistics & Data Analysis 52*(8), 3954–3970.

Park, C., J. S. Marron, and V. Rondonotti (2004). Dependent sizer: Goodness-of-fit tests for time series models. *Journal of Applied Statistics 31*(8), 999–1017.

Park, T.-S., I. Eckley, H. C. Ombao, et al. (2014). Estimating time-evolving partial coherence between signals via multivariate locally stationary wavelet processes. *Signal Processing, IEEE Transactions on 62*(20), 5240–5250.

Partridge, M. and R. A. Calvo (1998). Fast dimensionality reduction and simple pca. *Intelligent data analysis 2*(1), 203–214.

Patoine, A. and P. R. Leavitt (2006). Century-long synchrony of fossil algae in a chain of canadian prairie lakes. *Ecology 87*(7), 1710–1721.

Paul, M. and J. Meyer (2001). Streams in the urban landscape. *Annual Review of Ecology and Systematics 32*, 333–365.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2*(11), 559–572.

Peifer, M., J. Timmer, and H. Voss (2003). Non-parametric identification of non-linear oscillating systems. *Journal of sound and vibration 267*(5), 1157–1167.

Peres-Neto, P. R., D. A. Jackson, and K. M. Somers (2005). How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis 49*(4), 974–997.

Pesce, S. F. and D. A. Wunderlin (2000). Use of water quality indices to verify the impact of córdoba city (argentina) on suquıa river. *Water Research 34*(11), 2915–2926.

Petersen, W., L. Bertino, U. Callies, and E. Zorita (2001). Process identification by principal component analysis of river water-quality data. *Ecological Modelling 138*(1), 193–213.

Peterson, E. E. (2011). *STARS: Spatial tools for the analysis of river systems - a tutorial*. CSIRO.

Peterson, E. E., D. M. Theobald, and J. M. ver Hoef (2007). Geostatistical modelling on stream networks: developing valid covariance matrices based on hydrologic distance and stream flow. *Freshwater Biology 52*(2), 267–279.

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

Peterson, E. E. and J. M. Ver Hoef (2014). Stars: An arcgis toolset used to calculate the spatial information needed to fit spatial statistical models to stream network data. *J Stat Softw 56*(2), 1–17.

Pinto da Costa, J. F., H. Alonso, and L. Roque (2011). A weighted principal component analysis and its application to gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 8*(1), 246–252.

Plaia, A. and A. L. Bondì (2006). Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment 40*(38), 7316–7330.

Polansky, L., G. Wittemyer, P. C. Cross, C. J. Tambling, and W. M. Getz (2010). From moonlight to movement and synchronized randomness: Fourier and wavelet analyses of animal location time series data. *Ecology 91*(5), 1506–1518.

Ramsay, J. and B. Silverman (1997). *Functional Data Analysis.* Springer, New York.

Ramsay, J. and B. Silverman (2006). *Functional Data Analysis.* Springer, New York.

Ramsay, J. O., G. Hooker, and S. Graves (2009). *Functional Data Analysis with R and Matlab.* Springer, New York; London.

Ramsay, J. O., H. Wickham, S. Graves, and G. Hooker (2014). *fda: Functional Data Analysis.* R package version 2.4.4.

Rathbun, S. L. (1998). Spatial modelling in irregularly shaped regions: kriging estuaries. *Environmetrics 9*, 109–129.

Richman, M. B. (1986). Rotation of principal components. *Journal of Climatology 6*(3), 293–335.

Robinson, T. and G. Metternicht (2005). Testing the performance of spatial interpolation techniques for mapping soil properties. *Computers and electronics in agriculture 50*, 97–108.

Robson, A. and C. Neal (1997). A summary of regional water quality for eastern uk rivers. *Science of the Total Environment 194*, 15–37.

Rondonotti, V., J. S. Marron, and C. Park. (2007). Sizer for time series: A new approach to the analysis of trends. *Electronic Journal of Statistics 1*, 268–289.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Rubin, D. B. and R. J. Little (2002). Statistical analysis with missing data. *Hoboken, NJ: J Wiley & Sons*.

Rue, H., S. Martino, and N. Chopin. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *J R Statist Soc B 71(2)*, 319–392.

Rushworth, A., E. Peterson, J. Ver Hoef, and A. Bowman (2015). Validation and comparison of geostatistical and spline models for spatial stream networks. *Environmetrics*.

Saeys, W., B. De Ketelaere, and P. Darius (2008). Potential applications of functional data analysis in chemometrics. *Journal of chemometrics 22*(5), 335–344.

Sain, S. R., D. Nychka, and L. Mearns (2011). Functional anova and regional climate experiments: a statistical analysis of dynamic downscaling. *Environmetrics 22*(6), 700–711.

Sánchez-López, G., A. Hernández, S. Pla-Rabes, M. Toro, I. Granados, J. Sigró, R. Trigo, M. Rubio-Inglés, L. Camarero, B. Valero-Garcés, et al. (2015). The effects of the nao on the ice phenology of spanish alpine lakes. *Climatic Change 130*(2), 101–113.

Sanderson, J., P. Fryzlewicz, and M. Jones (2010). Estimating linear dependence between nonstationary time series using the locally stationary wavelet model. *Biometrika 97*(2), 435–446.

Sauquet, E., L. Gottschalk, and E. Leblois (2000). Mapping average annual runoff: a hierarchical approach applying a stochastic interpolation scheme. *Hydrological sciences journal 45*(6), 799–815.

Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate 14*(5), 853–871.

SDWA (1974). Safe Drinking Water Act of 1974, P.L. 93-523.

SDWA (1996). Safe Drinking Water Act Ammendments of 1996, P.L. 104-182.

Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 683–690.

Shen, F., D. Yang, Y. Ying, B. Li, Y. Zheng, and T. Jiang (2012). Discrimination between shaoxing wines and other chinese rice wines by near-infrared spectroscopy and chemometrics. *Food and bioprocess technology 5*(2), 786–795.

Shen, H. and J. Z. Huang (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis 99*(6), 1015–1034.

Shrestha, S., F. Kazama, and T. Nakamura (2008). Use of principal component analysis, factor analysis and discriminant analysis to evaluate spatial and temporal variations in water quality of the mekong river. *Journal of Hydroinformatics 10*(1), 43–56.

Shumway, R. H., R. S. Azari, and M. Kayhanian (2002). Statistical approaches to estimating mean water quality concentrations with detection limits. *Environmental science & technology 36*(15), 3345–3353.

Silverman, B. and J. Ramsay (2005). *Functional Data Analysis*. Springer.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–52.

Simpson, G. (2014). Simultaneous confidence intervals for derivatives of splines in gams. http://www.fromthebottomoftheheap.net/2014/06/16/simultaneous-confidence-intervals-for-derivatives/.

Sinelli, N., S. Limbo, L. Torri, V. Di Egidio, and E. Casiraghi (2010). Evaluation of freshness decay of minced beef stored in high-oxygen modified atmosphere packaged at different temperatures using nir and mir spectroscopy. *Meat science 86*(3), 748–752.

Singh, A. K., A. Singh, and M. Engelhardt (1997). The lognormal distribution in environmental applications. *Technology Support Center Issue Paper, 182CMB97. EPA/600/R-97/006*.

Skočaj, D., A. Leonardis, and H. Bischof (2007). Weighted and robust learning of subspace representations. *Pattern recognition 40*(5), 1556–1569.

Skøien, J. O., R. Merz, and G. Blöschl (2006). Top-kriging-geostatistics on stream networks. *Hydrology and Earth System Sciences 10*(2), 277–287.

Smith, T. M., R. W. Reynolds, R. E. Livezey, and D. C. Stokes (1996). Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *Journal of Climate 9*(6), 1403–1420.

Som, N., P. Monestiez, J. ver Hoef, D. Zimmerman, and E. Peterson (2014). Spatial sampling on streams: principles for inference on aquatic networks. *Environmetrics*. Early view.

Sonderegger, D. (2012). *SiZer: Significant Zero Crossings (version 0.1-4)*. http://CRAN.R-project.org/package=SiZer.

Stahlschmidt, S., W. K. Härdle, and H. Thome (2015). An application of principal component analysis on multivariate time-stationary spatio-temporal data. *Spatial Economic Analysis* (ahead-of-print), 1–21.

Strahler, A. N. (1957). Quantitative analysis of watershed geomorphology. *Transactions, American Geophysical Union 38*, 913–920.

Strickland, C., D. Simpson, I. Turner, R. Denham, and K. Mengersen (2011). Fast bayesian analysis of spatial dynamic factor models for multitemporal remotely sensed imagery. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 60*(1), 109–124.

Strobl, R. O. and P. D. Robillard (2008). Network design for water quality monitoring of surface freshwaters: A review. *Journal of Environmental Management 87*(4), 639–648.

Tamuz, O., T. Mazeh, and S. Zucker (2005). Correcting systematic effects in a large set of photometric light curves. *Monthly Notices of the Royal Astronomical Society 356*(4), 1466–1470.

Taylor, M., M. Losch, M. Wenzel, and J. Schröter (2013). On the sensitivity of field reconstruction and prediction using empirical orthogonal functions derived from gappy data. *Journal of Climate 26*(22), 9194–9205.

Thioulouse, J., D. Chessel, and S. Champely (1995). Multivariate analysis of spatial patterns: a unified approach to local and global structures. *Environmental and Ecological Statistics 2*(1), 1–14.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(3), 611–622.

Valle, S., W. Li, and S. J. Qin (1999). Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Industrial & Engineering Chemistry Research 38*(11), 4389–4401.

Van den Dool, H., S. Saha, and Å. Johansson (2000). Empirical orthogonal teleconnections. *Journal of Climate 13*(8), 1421–1435.

ver Hoef, J., E. Peterson, D. Clifford, and R. Shah (2014). SSN: An R package for spatial statistical modeling on stream networks. *Journal of Statistical Software 56*(3).

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

ver Hoef, J. M. and E. E. Peterson (2010). A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association 105*(489), 6–18.

Vines, S. (2000). Simple principal components. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 49*(4), 441–451.

Von Storch, H. and F. W. Zwiers (2001). *Statistical analysis in climate research*. Cambridge university press.

Vsevolozhskaya, O., M. Greenwood, D. Holodov, et al. (2014). Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis. *The Annals of Applied Statistics 8*(2), 905–925.

Walker, E. and S. P. Wright (2002). Comparing curves using additive models. *Journal of Quality Technology 34*(1), 118–129.

Wartenberg, D. (1985). Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis 17*(4), 263–283.

Webb, J. A. and M. Padgham (2013). How does network structure and complexity in river systems affect population abundance and persistence? *Limnologica-Ecology and Management of Inland Waters 43*(5), 399–403.

Whitehead, P., L. Jin, H. Baulch, D. Butterfield, S. Oni, P. Dillon, M. Futter, A. Wade, R. North, E. O'Connor, et al. (2011). Modelling phosphorus dynamics in multi-branch river systems: A study of the black river, lake simcoe, ontario, canada. *Science of the Total Environment 412*, 315–323.

Wiberg, T. (1976). Computation of principal components when data are missing. In *Proc. of Second Symp. Computational Statistics*, pp. 229–236.

Wilbers, G.-J., M. Becker, Z. Sebesvari, F. G. Renaud, et al. (2014). Spatial and temporal variability of surface water pollution in the mekong delta, vietnam. *Science of the Total Environment 485*, 653–665.

Withers, C. S. and S. Nadarajah (2015). Estimating trend from seasonal data: is daily, monthly or annual data best? *Environmetrics*.

Wold, H. and E. Lyttkens (1969). Nonlinear iterative partial least squares (nipals) estimation procedures. *Bulletin of the International Statistical Institute 43*, 29–51.

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

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(1), 3–36.

Wu, L., M. Bocquet, and M. Chevallier (2010). Optimal reduction of the ozone monitoring network over France. *Atmospheric Environment 44*, 3071–3083.

Young, S. G. and A. W. Bowman (1995). Non-parametric analysis of covariance. *Biometrics 51(3)*, 920–931.

Zhang, C., H. Peng, and J.-T. Zhang (2010). Two samples tests for functional data. *Communications in Statistics - Theory and Methods 39*(4), 559–578.

Zhang, J. P., T. Zhu, Q. H. Zhang, C. C. Li, H. L. Shu, Y. Ying, Z. P. Dai, X. Wang, X. Y. Liu, A. M. Liang, H. X. Shen, and B. Q. Yi (2012). The impact of circulation patterns on regional transport pathways and air quality over beijing and its surroundings. *Atmospheric Chemistry and Physics 12*(11), 5031–5053.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of computational and graphical statistics 15*(2), 265–286.

Zuur, A., R. Fryer, I. Joliffe, R. Dekker, and J. Beukema (2003). Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics 14*, 665–685.

Zuur, A., E. Leno, and G. Smith (2007). *Analysing Ecological data.* Springer, New York.

Zuur, A. F., I. D. Tuck, and N. Bailey. (2003). Dynamic factor analysis to estimate common trends in fisheries time series. *Can. J. Fish. Aquat. Sci. 60*, 542–552.