Evaluation of Sampling and Monitoring Designs for Water Quality

by

Ruth Alison Haggarty

A thesis submitted to the University of Glasgow for the degree of Doctor of Philosophy

in

Statistics

December 2012

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.

The work presented in Chapter 4 has been published in Environmetrics with the title 'Functional Clustering of Water Quality Data in Scotland' (October 2012). Part of this work has also been presented at the 26th International Workshop on Statistical Modelling (IWSM) in Valencia, 2011, with the same title. A manuscript based on the work presented in Chapter 5 is currently in preparation.

Signed: _____

Date: _____

Abstract

Assessing water quality is of crucial importance to both society and the environment. Deterioration in water quality through issues such as eutrophication presents substantial risk to human health, plant and animal life, and can have detrimental effects on the local economy. Long-term data records across multiple sites can be used to investigate water quality and risk factors statistically, however, identification of underlying changes can only be successful if there is a sufficient quantity of data available. As vast amounts of resources are required for the implementation and maintenance of a monitoring network, logistically and financially it is not possible to employ continuous monitoring of all water environments. This raises the question as to the optimal design for long-term monitoring networks which are capable of capturing underlying changes. Two of the main design considerations are clearly where to sample, and how frequently to sample.

The principal aim of this thesis is to use statistical analysis to investigate frequently used environmental monitoring networks, developing new methodology where appropriate, so that the design and implementation of future networks can be made as effective and cost efficient as possible. Using data which have been provided by the Scottish Environment Protection Agency, several data from Scottish lakes and rivers and a range of determinands are considered in order to explore water quality monitoring in Scotland. Chapter 1 provides an introduction to environmental monitoring and both existing statistical techniques, and potential challenges which are commonly encountered in the analysis of environmental data are discussed. Following this, Chapter 2 presents a simulation study which has been designed and implemented in order to evaluate the nature and statistical power for commonly used environmental sampling and monitoring designs for surface waters. The aim is to answer questions regarding how many samples to base the chemical classification of standing waters, and how appropriate the currently available data in Scotland are for detecting trends and seasonality. The simulation study was constructed to investigate the ability to detect the different underlying features of the data under several different sampling conditions.

After the assessment of how often sampling is required to detect change, the remainder of the thesis will attempt to address some of the questions associated with where the optimal sampling locations are. The European Union Water Framework Directive (WFD) (European Parliament, 2000) was introduced in 2003 to set compliance standards for all water bodies across Europe, with an aim to prevent de terioration, and ensure all sites reach 'good' status by 2015. One of the features of the WFD is that water bodies can be grouped together and the classification of all members of the group is then based on the classification of a single representative site. The potential misclassification of sites means one of the key areas of interest is how well the existing groups used by SEPA for classification capture differences between the sites in terms of several chemical determinands. This will be explored in Chapter 3 where a functional data analysis approach will be taken in order to investigate some of the features of the existing groupings. An investigation of the effect of temporal autocorrelation on our ability to distinguish groups of sites from one another will also be presented here.

It is also of interest to explore whether fewer, or indeed more groups would be optimal in order to accurately represent the trends and variability in the water quality parameters. Different statistical approaches for grouping standing waters will be presented in Chapter 4, where the question of how many groups is statistically optimal is also addressed. As in Chapter 3, these approaches for grouping sites will be based on functional data in order to include the temporal dynamics of the variable of interest within any analysis of group structure obtained. Both hierarchical and model based functional clustering are considered here. The idea of functional clustering is also extended to the multivariate setting, thus enabling information from several determinands of interest to be used within formation of groups. This is something which is of particular importance in view of the fact that the WFD classification encompasses a range of different determinands.

In addition to the investigation of standing waters, an entirely different type of water quality monitoring network is considered in Chapter 5. While standing waters are assumed to be spatially independent of one another there are several situations where this assumption is not appropriate and where spatial correlation between locations needs to be accounted for. Further developments of the functional clustering methods explored in Chapter 4 are presented here in order to obtain groups of stations that are not only similar in terms of mean levels and temporal patterns of the determinand of interest, but which are also spatially homogenous. The river network data explored in Chapter 5 introduces a set of new challenges when considering functional clustering that go beyond the inclusion of Euclidean distance based spatial correlation. Existing methodology for estimating spatial correlation are combined with functional clustering approaches and developed to be suitable for application on sites which lie along a river network.

The final chapter of this thesis provides a summary of the work presented and discussion of limitations and suggestions for future directions.

Acknowledgements

Firstly, I would like to thank my supervisors Prof. Marian Scott and Dr. Claire Miller for their invaluable support and guidance throughout my research. I am extremely grateful for their encouragement and patience, without which the production of this thesis would not have been possible. I would also like to acknowledge the Scottish Environment Protection Agency for providing the data for this project and in particular, thanks are due to Fiona Wyllie and Malcolm Smith for their helpful advice. In addition, I gratefully acknowledge the funding from the Engineering and Physical Sciences research council that allowed me to undertake this work.

Thank you to everyone in the Department of Statistics and to all of the new friends I have made during my time here. To Ally, Claire, Heather, Nicola and Robin - Friday drinks, nights out, lunches, and although often frowned upon, the occasional 'stats chats', kept me going if things were getting stressful, and made my experience at Glasgow University an enjoyable one.

A big thanks are due to all of my family and friends. To Thea, thank you for fulfilling your duties as a little sister by driving me crazy, and being brilliant in equal measure! To my mum and dad, thank you for your constant love and support. The example of hard work that you have set me has encouraged and helped me throughout my life and my education. I could not have accomplished what I have if it were not for you.

Finally, to David, your belief in me, patience and understanding over the past few months cannot be underestimated - thank you for being there constantly, for keeping me (almost) calm, and for always making me smile.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Water is an invaluable resource; providing drinking water and important inputs for many industries as well as facilities for recreation and leisure. Both maintaining and improving water quality, which is established using many determinands and characteristics, is therefore of crucial importance.

1.1 Water Quality Monitoring

The European Union Water Framework Directive (WFD) (European Parliament, 2000) was introduced in 2003 to set compliance standards for all water bodies across Europe, with an aim to prevent deterioration, and ensure all sites reach 'good' status by 2015. It is a wide ranging piece of legislation and has several implications for how monitoring networks are defined and implemented. A classification that is underpinned by a broad range of variables is required for all rivers, lochs, transitional, coastal and groundwater bodies. For surface waters, the classification is determined by the poorer of their chemical or ecological status. Chemical status describes whether or not the concentration of any pollutant exceeds standards that have been set for it at European Community (EC) level, while ecological status is principally a measure of the cumulative effects of human activities on river, lake, estuary or coastal water ecosystems. Each of the five ecological status classes (high, good, moderate, poor and bad) defined by the WFD represents a different level of disturbance from a reference state. In Scotland, the Scottish Environment Protection Agency (SEPA) is the regulatory agency responsible for monitoring water environments and for reporting classifications to the

European Union. SEPA was established in 1996 and is a non-departmental public body. Further to being accountable to the Scottish Parliament, SEPA also regulates and provides advice to business, industry and the public on environmental matters. The equivalent agency in England and Wales is the Environment Agency (EA).

Monitoring levels of water pollution has been a key focus of research and legislation in recent years and in addition to the WFD, there have been several other pieces of European Community legislation which have been brought into force to assess and set targets for water quality criteria. For example, in 1991 the Nitrates Directive (European Parliament, 1991) was introduced with an aim to both identify polluted water environments and reduce the levels of nitrate pollution from agricultural sources and in 2006 a revised European Community Bathing Water Directive (European Parliament, 2006) was introduced which set compliance standards for bathing waters in terms of safe limits for microbial indicator quantities. Both the Nitrates Directive and the Bathing Water Directive set specific limits for particular pollutants which must not be exceeded and require regular monitoring to be carried out.

The introduction of the WFD, which is an overarching directive that pulls together other such legislation, requires that regulatory agencies, such as SEPA, have extensive monitoring networks in place in order to have a satisfactory quantity of data on which to base classification. However, at the same time as the demand for comprehensive monitoring data and water quality reports is increasing, there are constraints on financial resources and so it is becoming increasingly important for those responsible for designing and implementing networks to know where, and how frequently, to collect samples. Moreover, monitoring is not only important due to the legislative requirement to assess standards for mandatory environmental policy, it is also vital as it enables detection of the presence and extent of underlying changes in water quality. Any changes detected can subsequently provide evidence that improvement measures already in place are working or, conversely, they can indicate that action plans are required to deal with areas of concern. The most recent Intergovernmental Panel on Climate Change report on water (Bates et al., 2008) highlighted the importance of monitoring data and identified improvement in the collection of data, as well as the use of available data, as an area for future development. The report acknowledged that 'water resources management clearly impacts on many other policy areas, e.g. energy, health, food, security, nature

conservation' while also stating that 'better observational data and data access are necessary to improve understanding of ongoing changes'.

Maher et al. (1994) provides a comprehensive discussion of the requirements of environmental monitoring programmes. The authors state that monitoring is often wasteful, and data rich, but information poor. The spatial selection of sampling sites and the quantity of data required are identified as key issues which must be addressed when designing effective sampling programmes whose aim is to assess environmental status. More recently, Field et al. (2007) states that more resources than ever before are being channeled to the task of documenting environmental change, however, current monitoring efforts still fall far short of what is required. The authors suggest that if policy driven monitoring, such as that required by the WFD, are improperly designed and implemented the consequences may be worse than not monitoring at all. The usefulness of existing environmental data in terms of detecting long-term trends, and the quantity of data required to ensure reliable conclusions, is discussed in Chapter 2.

In addition to these issues, which are common to many types of monitoring network, there are unique features of the WFD that also impact on the design of sampling procedures. One such feature of the WFD is that standing waters can be grouped together, and the classifications of all members of the group can then be based on the classification of a single representative lake, enabling water quality to be predicted without monitoring. In Scotland, SEPA currently practise a grouping approach for classification of lakes. Consequently, before any monitoring is carried out, lake groupings have to be established and representative sites identified. Grouping lakes which are similar in terms of the observed determinands of interest is of great importance, as wrongly specifying either the groups, or the representative lake within each group, could potentially result in misclassification of all members and hence could miss potential environmental risks. The question of how well existing SEPA groups are performing in terms of capturing the variability in chemical determinands at different lakes, as well as other possible ways in which lakes could be grouped form one of the main areas of interest for this thesis.

1.2 Water Quality Determinands

There are a range of determinands that impact water quality and that can be measured in order to evaluate the condition of the water environment. For some determinands, such as nitrate, there are strict targets set in terms of acceptable levels, whilst other determinands, such as alkalinity, are used more generally as an indicator of water quality. In addition to hazardous substances that are controlled by regulation, chemicals which are viewed as nutrients, such as phosphates, can cause eutrophication of water bodies if they occur in high concentrations (Smith et al., 1999). Eutrophication is the process by which a body of water becomes enriched in dissolved nutrients that stimulate the growth of algae and other aquatic plant life. As this abundance of algae usually results in the depletion of dissolved oxygen, eutrophication can be detrimental to animal and plant life.

There are two key sources of water pollution; point source and diffuse pollution. Point source pollution is related to emissions from a single discharge source which can easily be identified. In contrast, diffuse pollution does not have one identifiable origin but instead consists of pollution resulting from several different sources and land-use activities (Environment Agency, 2007). Each of these sources is indirect, and although they may only contribute a small amount of waste individually, they can be collectively important. While point source contamination can be controlled through regulation, it is often more difficult to deal with diffuse pollution and a recent report on risk assessment for the WFD highlighted diffuse pollution as being a bigger risk to rivers, lakes and groundwaters than point sources (Environment Agency, 2007). One source of diffuse pollution is agricultural run-off. For example, rainfall washes manure used as fertilizer or livestock waste from surrounding fields either directly into the water itself or into connecting streams. Other sources can include partially or untreated sewage and application of some lawn fertilizers. A summary of some of the most commonly monitored determinands used to assess water quality is now provided.

Phosphorus/Phosphates: The element phosphorus and compounds that are composed of phosphorus ions in a different chemical arrangement called phosphates are necessary for plant and animal growth (EPA, 1976). While phosphorus occurs naturally and phosphate forms are produced by natural processes, a major source of this compound in water environments is due to diffuse pollution. For example, after rainfall, varying quantities of

phosphates found in most fertilisers wash from farmland into nearby waterways (Smith et al., 1999). Phosphates stimulate the growth of plankton and aquatic plant life which provide food for fish, but, as noted earlier, eutrophication can occur if there is an excess of phosphate.

Nitrate: Both nitrate and nitrite are forms of the element nitrogen. The available nitrate data investigated in this thesis consists of two types of measurement; nitrate (N) and total oxidised nitrate (TON). TON is technically the sum of nitrate and nitrite levels but, since the latter tends to be negligible, SEPA regards TON and nitrate as equivalent. As with phosphorus, nitrates stimulate the growth of plankton and aquatic plants that provide food for fish. Furthermore nitrate is also a major component of agricultural fertiliser and hence the largest contributors of nitrate in water environments are sources of diffuse pollution as opposed to point sources (EEA, 2010).

Chlorophyll$_a$: Chlorophyll$_a$ is bound within the living cells of algae and other phytoplankton found in surface water and is a key component in photosynthesis, the process in which energy from sunlight is used by plants to produce oxygen. Chlorophyll$_a$ is essential to the existence of phytoplankton and hence it can be used as an indirect indicator measure for the health of a water environment. Throughout this thesis chlorophyll$_a$ will be referred to as chlorophyll.

Alkalinity: While alkalinity is not a pollutant, it is a measure of substances within the water that have acid neutralising ability and is essentially a measure of the ability of a water source to keep its pH from changing. It is an important component for fish and aquatic life since alkalinity acts as a buffer to changes in pH and provides protection from sudden shifts (EPA, 1976). The main sources of natural alkalinity are rocks, which contain carbonate, bicarbonate, and hydroxide compounds.

1.3 Existing Techniques for Modelling Environmental Data

Environmental and ecological data are typically sequential over time and space. Data which arise from samples collected over a period of several years at regular

intervals are referred to as time series data. Often the key area of interest is to quantify the nature and extent of any trend in the variables and, as mentioned previously, this can be used to identify potential areas of concern, or to indicate that measures already put in place to deal with problems are having the desired effect. In this context a trend can be defined as a generally upward or downward drift in the long-term, or in the case of space, long-range, average that is commonly, but not necessarily, linear. Another feature of time series data is the presence of seasonality or cyclical patterns which are repetitive short-term patterns of known length.

Usually monotonic trends are of most interest and there are several approaches which can be taken to test for this. One common method of checking for non-parametric trends is the Mann Kendall test for deseasonalised data (Mann, 1945). Following from this a Seasonal Kendall test was developed by Hirsch et al. (1982) for trend analysis of water quality data. Smith et al. (1993) extends the ideas of detecting and estimating the magnitude of temporal trends in measures of water quality to the multivariate setting while Yue and Wang (2004) adapts the Mann Kendall test to account for serially correlated samples in hydrological time series.

In the section that follows existing and commonly used methods and models for the evaluation of environmental data will be described. This will include methods for time series analysis, regression models and smoothing techniques.

1.3.1 Time Series Models

Modelling or testing for trends and seasonal patterns are frequently the main focus of any statistical analysis in this field, but the presence of correlated data often complicates this. Many statistical models assume errors that are independent of one another, however, due to the nature of the data and the close proximity of observations to one another in either space or time, lagged relationships are typical. Correlation between measurements is therefore very likely. Incorrectly assuming data are independent when they are correlated can potentially result in estimates of the standard errors which are smaller than they should be. Although it is widely accepted that environmental data collected more than two weeks apart are not significantly correlated in time (van Belle and Hughes, 1984), in some situations data are collected at frequencies where the time period between observations is shorter than this. A brief description of commonly used time series methods is

provided in this section. For full details see Brockwell and Davis (1991) which has been used as the main reference text for the description provided here.

In order to explore patterns for a single response variable of interest, Y at any time point t we consider a time series of the form,

$$Y_t = m_t + s_t + \epsilon_t \text{ where } t = 1, 2, .., n \tag{1.1}$$

Here m_t represents any trend, s_t is a short term seasonal or cyclical pattern of known period and ϵ_t is a random White Noise term. The simplest model for this data would be where m_t and s_t can be described in terms of some parametric trend and cyclical pattern which have a known functional form over time, for example a linear or polynomial trend and constant seasonality. While the aim is often to use the observed data to estimate the trend and seasonal component the presence of autocorrelated errors means that using standard techniques such as ordinary least squares (OLS) to fit models is inappropriate. One way to account for autocorrelation is to incorporate previous observed values of the variable of interest, y, into the regression equation by expressing the current value of y_t as a finite linear combination of these earlier values. This is known as an AutoRegressive (AR) model and a general form of this model AR(p) can be expressed as;

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

where $\delta_1, ...\delta_p$ are coefficients and e_t is a White Noise sequence. Alternatively, the current value of the variable of interest can be expressed in terms of both past and current noise terms. This is known as a Moving Average (MA) model and can be written in general terms as an MA(q) model as;

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

The combination of these two classes of model produces an AutoRegressive Moving Average or ARMA(p, q) model as;

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

These techniques for modelling autocorrelated data require that the time series

data are stationary. Stationarity of a time series implies that properties such as the mean and variance are constant over time and hence that there are no trends or seasonal patterns present but in practice the assumption of stationarity is unrealistic.

An extension of the ARMA class of models are AutoRegressive Integrated Moving Average (ARIMA) models and Seasonal ARIMA models, known as SARIMA models which can be used to include non-stationary mean and seasonal dynamics of the data. The key idea is to build an ARMA(p, q) model on a stationary time series that is obtained from the original time series via differencing. With an ARIMA model first order differencing is applied to remove any trend from the data. If Y_t denotes the value of the time series Y at time t, then the first difference of Y_t at time t is equal to $\Delta Y_t = Y_t - Y_{t-1}$. More generally, period differencing can be applied by computing the time series $\Delta^d Y_t = Y_t - Y_{t-d}$, for a given period, d. An ARMA process is then used to model the series $\Delta^d Y_t$. The ARIMA model can be denoted by ARIMA(p, d, q) where p and q are the lags corresponding to the AR and MA components of the models whilst d corresponds to the order of differencing applied. SARIMA models are a further generalization of the ARIMA class which are used when there is a seasonal component in the data. A SARIMA $(p, d, q) \times (P, D, Q)$ model is an ARIMA(P, D, Q) where the residuals are ARIMA(p, d, q). For example, a SARIMA model can be fitted to a series by initially using first order differencing to remove trend to obtain ΔY_t then subsequently applying period differencing to remove any seasonal pattern. For monthly data, 12th order differencing could be applied to the series ΔY_t and the resulting series could then be modelled with an ARMA(p, q) process.

After the application of differencing to ensure the time series is stationary, an appropriate order ARMA(p, q) process needs to be selected. In order to identify p and q the sample autocorrelation function (ACF) can be computed. Let $y_t, t = 1, ..., N$ be observations of a time series and let the sample mean of the series be denoted by $\bar{y}$. Then the sample autocorrelation function at lag h can be written as

$$\hat{\rho}_h = \frac{\sum_{t=1}^{N-1}(y_t - \bar{y})(y_{t+h} - \bar{y})}{\sum_{t=1}^{N}(y_t - \bar{y})^2}$$

The ACF is the collection of sample correlation coefficients that correspond to the cross-correlation of the data with itself at a series of different lags in time. Plotting

the estimated correlation coefficients against the lags provides a correlogram which can be used to suggest a suitable model. If the sample ACF shows unstructured non-zero coefficients to lag q this is an indication that an MA(q) component is required while smooth decay suggests that an AR component is required. Oscillation of the coefficients within the ACF is evidence that an AR process of order 2 or higher would be needed. To identify the order of the AR component the Partial ACF (PACF) can be used. The PACF of lag h is the autocorrelation between all observations y_t and y_{t+h} that is not accounted for by lags 1,...,h-1 inclusive. Smooth decay in the PACF suggests an MA process is required, while unstructured non-zero coefficients to lag p provide evidence that an AR(p) component is suitable. After selection of appropriate orders the models can be fitted to the series using least squares.

ARIMA and SARIMA models provide one approach to modelling data where there is correlation between the errors but an alternative method is to model the trend and seasonal components of the time series explicitly and then to adjust the standard errors accordingly. In this approach the correlation can be viewed as a nuisance parameter rather than a component of direct interest. After estimating the trend and seasonal components of the data, the remaining autocorrelation in the residuals can subsequently be modelled using an ARMA process. This enables the components of the data which are often the main focus of interest to be estimated using standard approaches and allows for the estimation of features of the data such as non-linear trends. Often a simple AR(1) process is sufficient for modelling the covariance structure of water quality parameters. Houseman (2005) use a first order autoregressive process to model depth data at a Boston Harbour, while Clement et al. (2006) use an AR(1) process within a spatio-temporal model which is fitted to dissolved oxygen concentrations on a river network in Belgium. Bowman et al. (2009) also found an AR(1) process to adequately capture the correlation structure of the errors when modelling sulphur dioxide trends across Europe in time and space.

1.3.2 Nonparametric Models

There are several complicating features of water quality data that make traditional parametric methods such as linear regression techniques, analysis of covariance, and standard time series approaches difficult to implement. Moreover, as noted

earlier, the relationships between parameters of interest and time are often complex and do not follow a linear pattern where there is a constant monotonic increase or decrease in time. Consequently, changes in environmental data through time are often analysed using nonparametric methods. Nonparametric regression enables the assumption of linearity to be removed and more flexible, smooth functions to be fitted instead. In addition to the investigation of non-parametric trends, these flexible regression methods can also be used to explore non-constant seasonal patterns. The key idea in nonparametric modelling is to average the values of the response variables locally as opposed to globally.

The application of smoothing methods can be used to estimate the dependence of the mean of the response variable y on a covariate, or covariates. Denoting the response variable as y, a general nonparametric model can be written as

$$\begin{aligned} y &= g(x_i) + \epsilon \\ \text{where } \epsilon &\sim N(0, \sigma^2) \end{aligned} \tag{1.2}$$

Here x_i is the covariate and ϵ is the error term. The function $g(x_i)$ which describes the relationship between y and x_i is unspecified and can be estimated by a smooth function, $\hat{g}(x)$. To estimate a smooth temporal trend a model where the covariate is time in the form of decimal year can be used. This model can be written as

$$\begin{aligned} y &= g(\text{time}) + \epsilon, \\ \epsilon &\sim N(0, \sigma^2) \end{aligned} \tag{1.3}$$

Following this, a bivariate model is an extension of the nonparametric regression model to two dimensions. This model is of the form

$$\begin{aligned} y &= g(x_1, x_2) + \varepsilon, \\ \text{where } \epsilon &\sim N(0, \sigma^2) \end{aligned} \tag{1.4}$$

Here the bivariate term $g(x_1, x_2)$ involves two different covariates. One potential application of this model is to assess changes in the seasonal pattern over time by including both decimal year and month. A bivariate model allows there to be a varying seasonal pattern across the time period, which can be expressed as

$$y = g(\text{year}, \text{month}) + \epsilon, \epsilon \sim N(0, \sigma^2) \tag{1.5}$$

Alternatively bivariate terms can be used to include geographic co-ordinates as covariates in models which aim to assess spatial patterns in the response variable of interest, for example, using the terms $g(\text{longitude}, \text{latitude})$.

Additive Models

Additive models extend univariate or bivariate nonparametric models to include a sum of smooth functions of a set of covariates. They create an estimate of a response variable by combining a collection of functions of predictors which are assumed to act additively. Models of this form are a particular case of generalised additive models with normal errors and are discussed in detail in Hastie and Tibshirani (1990). Given a response variable y_i and a set of covariates $(x_1, \ldots, x_k)$ a general expression for an additive model is given by

$$y_i = \mu + \sum_{j=1}^{k} g_j(x_{ji}) + \epsilon_i \tag{1.6}$$
$$\text{where } j = 1, ..., k \text{ and } \epsilon_i \sim N(0, \sigma^2)$$

Here $g_1(x_1), ..., g_k(x_k)$ are arbitrary smooth functions of the covariates which, to ensure identifiability, are subject to the constraint $\sum_{i=1}^{n} g_j(x_{ji}) = 0$. The parameter μ is an overall mean term.

In order to assess changes in environmental data the nonparametric models described above often use terms that are smooth functions of year and month of year. For example, an additive model of the form

$$y = \mu + g_1(\text{year}) + g_2(\text{month}) + \epsilon, \epsilon \sim N(0, \sigma^2) \tag{1.7}$$

can be used to consider the trend and seasonal component of the determinand of interest. Here μ is an overall mean, g_1 corresponds to the trend across the time period and g_2 corresponds to the seasonal pattern. Unlike the bivariate model in Equation 1.5, this additive model assumes that the seasonal pattern within each year is constant across the years.

Models 1.2, 1.5 and 1.7 assume that the observations are independent. Procedures to include correlation in these models (via the errors) are discussed in Section 1.4.1.

1.3.3 Smoothing Methods

A method of smoothing is required for all the nonparametric models that have been discussed. After application of a smoothing method to data, a smooth function is obtained which is less variable than the response variable. There are a variety of different smoothing methods available, some are commonly used for graphical exploration and to provide an indication of the underlying structure of the relationship between the variables whist others are for explicit modelling of the function $g(x)$. In addition to the method of smoothing, the extent to which the observed data are smoothed also has be defined.

Firstly, let the estimated nonparametric relationship between the response y and a single explanatory variable x be denoted by $\hat{g}(x)$. Then this estimated smooth function can be expressed as

$$\hat{g}(x) = Sy$$

where S represents a smoothing matrix. Following this there needs to be a way of defining the smoothing matrix S. Smoothers use local averaging whereby the observations within a set distance, commonly referred to as a local neighbourhood, of a single observation of interest are averaged. The two key questions that arise regarding the definition of a smoother are what size should the local neighbourhood which surrounds the observation of interest be and how should the observations which fall into each of these local neighbourhoods be averaged? The first of these questions is dealt with by specifying a smoothing parameter that controls the size of the neighbourhoods, although this in turn generates a question of how to choose the optimal smoothing parameters. Different approaches for choosing smoothing parameters are described later. For the second of the above questions there are several ways to define how averaging within neighbourhoods is carried out. Amongst the most frequently used methods are local running mean smoothers, kernel regression approaches, smoothing splines and regression splines. Each of these methods are described in Hastie and Tibshirani (1990), Bowman and Azzalini (1997) and Wood (2006). Although the methods of estimating smooth functions differ in philosophy and style, the end results are often very similar in terms of the estimates produced.

Kernel Smoothing

Kernel approaches use kernel functions centered around each observation of interest on the covariate axis to define a set of weights that can then be assigned to the surrounding observations. In general, a kernel function is a smooth positive function which peaks at the target observation, x, and decreases monotonically the further away the observations are from x. The smoothing parameter or bandwidth, h, defines the width of the kernel function which surrounds each of the observations of interest and hence the extent of the smoothing applied.

Loess

One popular method of smoothing which is often employed to obtain a graphical overview of underlying patterns in a dataset is a locally weighted running line smoother known as loess. This method of defining a local regression model was proposed by Cleveland and Devlin (1988). Within the loess method, an area surrounding a target observation x_0 is obtained by identifying the target points k nearest neighbours. This area can be written as $N(x_0)$. Next, the Euclidean distance between x_0 and the furthest away point within each neighbourhood is calculated as

$$\Delta(x_0) = \max_{N(x_0)}|x_0 - x_i|.$$

Within neighbourhood weights are subsequently assigned to each of the observations (x_0, y_0), using the tri-cube weight function which can be defined as

$$w_i = W\left(\frac{|x_0 - x_i|}{\Delta(x_0)}\right)$$

where

$$W(u) = \begin{cases} (1-u^3)^3 & : \text{ for } 0 \leq u < 1 \\ 0 & : \text{ otherwise} \end{cases}$$

Using these weights in combination with weighted least squares can then be used to produce a locally weighted straight line smooth. The smoothing parameter for locally weighted running line smoothers will determine the quantity of data which contributes to the estimate at each point by specifying the percentage of points that fall within a neighbourhood, the nearest neighbours to the target observation.

The loess method is also described in Cleveland et al. (1990) and in Hastie and Tibshirani (1990).

Local linear regression

Another commonly used smoothing method is local linear regression where a normal distribution is specified to define the weights (Cleveland and Devlin, 1988, see Bowman and Azzalini, 1997 for details). The idea is to estimate the regression function, $g(x)$, in the model $y_i = g(x_i) + \varepsilon_i$ based on data $\{(x_i, y_i); i = 1, ..., n\}$ by minimising the weighted sum-of-squares criterion;

$$\underset{\alpha,\beta}{min} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_i - x)\}^2 w(x_i - x; \lambda). \tag{1.8}$$

The estimate at x is then defined as the minimising value of α. If the smoothing parameter specified is λ, and the kernel function specified is a normal probability density function with mean 0 and standard deviation λ, then the weights can be defined as;

$$w_\lambda(x_i - x; \lambda) \quad = \quad \exp\left(-0.5\left(\frac{x_i - x}{\lambda}\right)^2\right)$$

This means that observations within an area spanning 2λ on either side of each point of interest will contribute to the estimate of that point, equivalent to a distance of approximately 4λ in total. Equation 1.8 can alternatively be expressed in matrix notation. Writing the vector with i^{th} element $(x_i - x)$ as X, and defining a diagonal weight matrix W which has entries corresponding to the kernel weights $w(x_i - x; \lambda)$ then the local linear regression least squares criterion is the minimising value of

$$\overset{min}{\alpha,\beta} \{y - \alpha I_n - X\beta\}^T W \{y - \alpha I_n - X\beta\} \tag{1.9}$$

Here I_n represents the identity of size n.

The asymptotic properties of local linear regression are discussed in Fan (1992). One of the advantages of this estimator over the more simple running mean approach is the superior behaviour of the local linear regression smoother near the boundaries of the region where data are collected. Further discussion of local linear

smoothers are provided in Fan and Gijbels (1992) where the idea of having a non-constant bandwidth λ is proposed. The extension of the local linear smoothing approach to the multivariate case is discussed in Ruppert and Wand (1994).

Cyclical Patterns

Another consideration when fitting models is that components referring to seasonal information, such as month of year, are defined on a cyclical scale, and hence require a different treatment. To deal with these components a local mean estimator constructed as

$$\overset{min}{\alpha} \{y - \alpha I_n\}^T W \{y - \alpha I_n\} \tag{1.10}$$

can be used in combination with a Von Mises weight function to define W. Denoting month as x_2 and the period as r, so for example with monthly data $r = 12$, the Von Mises weight function is defined as

$$w(x_{2i} - x_2; \lambda) = \exp\left\{\frac{1}{\lambda}\cos\left(2\pi\frac{x_{2i} - x_2}{r}\right)\right\}$$

The purpose of this weight function is to ensure that the estimate of the component corresponding to month is adapted to take into account the cyclical scale and means that observations at one boundary of the period influence the estimate at the other end. A local mean approach is taken rather than a local linear approach as the seasonal pattern will not take the form of a straight line. Examples where this local linear regression models are employed as the smoothing method and cyclical patterns are included in this way are provided in Ferguson et al. (2008) and McMullan et al. (2007).

Kernel Model Fitting

Univariate and bivariate non-parametric models for independent data can be fitted via miminisation of the local regression least squares criterion given in Equation 1.8. Writing the parameters α and β as θ, the covariate data in matrix form as X and the response variable as Y then the model can be constructed as

$$Y = X\theta + \epsilon.$$

Following from this the weighted least squares estimator of the parameters, $\hat{\theta}$ can be expressed in matrix form as

$$\hat{\theta} = (X^T W X)^{-1} X^T W y$$

where W is the weight matrix.

For additive models, the back-fitting algorithm, as detailed in Hastie and Tibshirani (1990), can be used. Initially, smooth estimates are obtained for each of the model components $g_j(x_{ij})$ by minimising a least squares criterion and then, in order to obtain final estimates for each component, smoothing is applied iteratively with respect to each component by using the residuals based on the remaining model components as the response. After convergence of the backfitting estimates, the approximate smoothing matrices for each component can be written as $\hat{g}_j = \mathcal{P}_j y$ where $\mathcal{P}_j$ is referred to as a projection matrix. The final estimate y can subsequently be written as

$$\hat{y} = \mathcal{P} y \text{ where } \mathcal{P} = \sum_{j=0}^{k} \mathcal{P}_j \tag{1.11}$$

The matrix $\mathcal{P}_0$ represents an $n \times n$ matrix with entries $1/n$. The expression given in Equation 1.11 is similar to the form of the estimate for univariate and bivariate models when expressed in terms of a single smoothing matrix, S (Giannitrapani et al., 2005).

Spline Smoothing

Spline functions are an alternative to kernel methods which can also be used as a method of representing smooth functions $g(t)$. A relatively naive approach would be to use polynomial regression with a low-order polynomial to represent the smooth function g in Equation 1.2 and to estimate the coefficients of the polynomial terms using a least squares criterion. However, polynomial regression would often have to use a high order polynomial - and therefore an excessive number of parameters - in order to capture all of the main features of the data. Spline methods can account for elaborate relationships without having to estimate an unnecessarily large number of parameters. A detailed overview of smoothing splines is provided in both Green and Silverman (1993) and Gu (2002). Using Green and

Silverman (1993) as the key reference, a description of cubic spline smoothing and regression spline smoothing with a B-spline basis will now be discussed.

Cubic Spline Functions

Spline functions consist of polynomial segments which are joined together smoothly at pre-defined subintervals. The points at which the joins occur are called break-points, or knots, of the spline and the order of the polynomial, g, within each section is defined by the degree of the polynomial segment plus one. It is clear that every smoothing spline function is defined by both the location and number of the knots as well as the order of the polynomial segments. Order 4 polynomial segments are amongst the most commonly used. As this means the fitting function is piece-wise cubic, the smooth function using order 4 splines is hence commonly referred to as a cubic spline function.

In order to approximate a function over a closed interval $[a, b]$ using cubic spline functions, the whole interval is first divided into subintervals. Given a series of real numbers which lie within the interval $s_1, ..., s_n$ such that $a \leq s_1 < s_2 < ... \leq s_n < b$ then a smooth function g can subsequently be fitted across the whole interval $[a, b]$ with a cubic polynomial segment in each interval $[(a, s_1), (s_1, s_2), ..., (s_n, b)]$.

$$g(t) = \begin{cases} g_0(t) & : \quad a \leq t \leq s_1 \\ g_1(t) & : \quad s_1 \leq t \leq s_2 \\ \vdots & : \quad \vdots \\ g_{n+1}(t) & : \quad s_b \leq t \leq b \end{cases}$$

Each of the internal points, s_i, are the knots. For g to be a cubic spline, values of the polynomial segments are not only required to be equal at the joins but there is a further constraint that the first and second derivatives at the end of one curve are equal to the first and second derivatives at the start of the next, to ensure the joins at each knot are smooth. In addition to these conditions, other constraints can be imposed. For example, natural cubic splines require that the value of the second and third derivatives of g at the start and end points a and b are both equal to zero. A particular definition for a natural cubic spline is provided in Green and Silverman (1993) in terms of the value of the function, g, and of the second derivative, g'' at each of the knots s_i. Natural smoothing splines can be

used to produce interpolating splines and in fact, Green and Silverman (1993) state that if there are more than two points there is a unique natural cubic spline which is an interpolant of these points with a knot at each data point.

While there are applications where an interpolating function may be of interest, more often than not the aim when fitting a smooth function is not to interpolate the observed data, but instead to estimate a smooth function which is close to the data but avoids local fluctuations which could be due to random noise. Clearly the aim of spline smoothing is to fit a smooth, flexible function which minimizes the residual sum of squares. However, if the model in Equation 1.2 is fitted using unconstrained least squares then the function which would minimize this is the curve which simply interpolates the data. Consequently, a roughness penalty approach is needed which will produce a smoother, more flexible function that will capture the main features of the data, but will avoid random fluctuations which will occur with interpolation. There is clearly a bias-variance trade-off and while it is important that the function fitted captures important curvature in the data, it is conversely important to ensure the curve is not excessively locally variable. As in standard smoothing approaches, roughness penalty approaches also require the minimisation of a fitting criterion, however this criterion will incorporate some pre-specified measure of 'smoothness'.

Following this, there has to be some definition of how best to quantify the roughness of a function. Although there are several measurements that could be used, one popular measure is the integrated square of the second derivative, also known as the curvature at t, which is defined as

$$PEN_2 = \int [g''(t)]^2 dt$$

The reason this value is a natural choice to measure roughness of a function is that this value will be equal to zero if $g(t)$ is a straight line (which obviously has no curvature). Although this is a good measure of smoothness it is not always appropriate and so a more general penalty term can be defined. A broader roughness penalty can be defined by any m-th order derivative, D^m, as

$$PEN_m = \int [D^m g(t)]^2 dt \tag{1.12}$$

Any linear combination of derivatives, known as a linear differential operator, of the form $L_{g(t)} = \sum_{m=1}^{M} \beta_m(t) D^k g(t)$ can be used as a penalty. Using the definition

of a measure of roughness in Equation 1.12, the least squares criterion that is used to determine the spline coefficients is modified to include this penalty measure and can be written as,

$$\sum_{i=1}^{n}(y_i - g(t_i))^2 + (\lambda \int [D^m g(t)]^2 \, dt) \tag{1.13}$$

The parameter λ in the above expression is a smoothing parameter which is a positive scalar that determines the emphasis of the role of the roughness penalty term and therefore controls the trade-off between goodness of fit and departures from smoothness. As λ increases, the greater the influence of the roughness penalty imposed relative to goodness of fit, and hence the smoother the function will become. Conversely as λ approaches zero, then $\hat{g}$ becomes increasingly locally variable and will eventually become the interpolating function when $\lambda = 0$. Subjective and automatic methods of selecting optimal smoothing parameters are discussed later in Section 1.3.4.

One of the drawbacks of using cubic smoothing splines is that there are as many parameters are there are observations. Since the number of parameters that are required to define a spline smoother is the number of interior knots, plus the order of the polynomial minus one, for a cubic spline with knots at each observation this means the number of parameters is $(n-2)+3-1 = n$. While this implies the function fitted will therefore have n degrees of freedom, the influence of the smoothing parameter, λ, results in a function which is smoother than this large number of parameters implies. This excessive number of parameters can become very computationally inefficient, particularly if there are multiple covariates. In an attempt to overcome this potential problem with smoothing splines, penalised regression splines - which use B-spline bases - are often used as they enable functions to be built from a linear combination of a set of spline functions which is substantially smaller than if the function is fitted to all the data.

B-Splines

Another common way to build a smooth function is through sets of known functions, called basis functions, that are mathematically independent. Smooth functions can be approximated using weighted sums of the individual functions. Amongst the aspects which control the flexibility of the function that is estimated using

spline bases are both the type of basis function used and number of functions used. While there are a wide variety of basis systems available, the choice of basis system is often dependent on the data to which the smooth functions are to be fitted, for example, polynomial basis systems, known as B-splines, are commonly used to represent non-periodic basis systems, while fourier basis systems are used to represent periodic functions. These two systems are frequently complemented by the addition of constant and monomial bases.

Polynomial B-spline basis functions are amongst the most commonly used basis systems. The B-spline system was first developed by de Boor (1978) and has several properties which mean they provide a particularly flexible and computationally efficient approach for non-periodic data. One of the key attributes of polynomial B-splines is the compact support property which means that a B-spline basis of order m is non-zero between a maximum of $m + 2$ adjacent knots (or equivalently over m adjacent intervals). This property results in a relatively sparse design matrix which makes B-splines computationally efficient. Both the number of basis functions and the number and placement of knots also have to be decided upon. For B-splines, the number of spline functions within the basis and the number of degrees of freedom in the fit is equal to the order of the polynomials which define the basis functions plus the number of interior knots. As expected, this means that if there are no interior knots, then the fit is a simple polynomial fit with the degrees of freedom equal to the order of the polynomial.

After a decision has been made as to what basis system is most appropriate, there is then some question as to where to place the interior knots. While the smoothing splines already discussed include a knot at each observation, for regression spline smoothing using B-splines, it is common for there to be fewer knots than observations and for the knots to be equally spaced. This is a suitable approach to adopt if the observed data are regularly spaced across the interval of interest. Alternatively, Ramsay and Silverman (1997) discuss the placement of knots and suggest having knots at different quantiles of the distribution by placing at every j^{th} data point, where j is a suitable pre-specified integer. This may be appropriate if the data are sparse. Subjective selection of knot placements may also be employed, for example, if there are particular areas where there is thought to be a large amount of curvature, additional knots can be included in these regions.

In order to define a B-spline basis with P basis functions of degree m (order $m+1$) it is first necessary to define $P+m+1$ knots. Let $t = (t_1, t_2, ..., t_{P+m+1})$ be the knot vector where $t_1 \leq t_2 \leq ... \leq t_{P+m+1}$ and the interval of interest over which the smooth function has to be estimated is $[t_{m+2}, t_P]$. Then an $(m+1)^{th}$ order smooth spline, $g(t)$, can be written as

$$g(t) = \sum_{p=1}^{P} \phi_p^m(t) c_P \tag{1.14}$$

where c_p are coefficients which have to be estimated. The individual B-spline basis functions, ϕ_p^m can be defined recursively using the Cox-de Boor Recursion Formula (de Boor, 1978)

$$\phi_p^m(t) = \frac{t - t_p}{t_{p+m+1} - t_p} \phi_p^{m-1} + \frac{t_{p+m+2} - t}{t_{p+m+2} - t_{p+1}} \phi_{p+1}^{m-1}$$

where $p = 1, ..., P$ and

$$\phi_p^{-1}(t) = \begin{cases} 1 & t_p \leq t < t_{p+1} \\ 0 & \text{otherwise} \end{cases}$$

Alternatively, 1.14 can be written more generally in matrix notation. Given a set of P known basis functions, ϕ_p, where $p = 1, ..., P$, that exist over the same range as data pairs (t_i, y_i) where $i = 1, ..., N$ then $y = g(t)$ can be expressed as the basis function expansion

$$y = g(t) = \sum_{P}^{p=1} \phi_p(t) c_p = \sum_{p=1}^{P} c_p^T \phi_p(t) = c^T \Phi(t)$$

where c is a vector of length P that contains the coefficients c_p while $\Phi(t)$ is an $N \times P$ matrix containing the values $\phi_p(t)$. In order to compute the coefficients c_p regression spline smoothing is used whereby a least squares criterion is minimised similarly to standard regression model fitting. Assuming the model in Equation 1.2 with independent, normally distributed errors with mean zero and constant variance, then it is the aim to minimise the residual sum of squares

$$\begin{aligned} RSS(c|y) &= \sum_{i=1}^{N} \left[y_i - \sum_{p=1}^{P} c_p \phi_p(t_i) \right]^2 \\ &= \sum_{i=1}^{N} \left[y_i - \phi(t_i)^T c \right]^2 \\ &= (y - \Phi c)^T (y - \Phi c). \end{aligned} \tag{1.15}$$

Taking the derivative of Equation 1.15 with respect to **c** and solving enables the least squares estimate $\hat{c}$ of c

$$\hat{c} = (\Phi^T \Phi)^{-1} \Phi^T y \tag{1.16}$$

Penalised Regression Splines

As discussed, rather than using unrestricted least squares, a roughness penalty approach is required in order to ensure that an estimate is obtained which captures the curvature of the data without simply interpolating the observations. Penalised regression splines combine polynomial B-splines with the second order roughness penalty to fit the curve to the data. The aim is to estimate the coefficients which will minimise the penalised least squares criterion given by Equation 1.13. Wood (2006) states that advantages of penalised regression splines include that they are both straightforward to use and are sufficiently flexible since any order of penalty can be used in conjunction with any order of B-spline basis. However, the author also expresses concerns that in practice, the level of complexity for implementing and interpreting penalised regression spline smooths becomes somewhat more difficult if unequally spaced knots are used. To compute the best estimate of the penalised regression spline coefficients, $\hat{c}$, it is first necessary to express the roughness penalty in Equation 1.12 in matrix form. First taking the general penalty measurement

$$PEN_m = \int [D^m g(t)]^2 \, dt$$

then substituting $g(t) = c^T \Phi(t)$ gives

$$\begin{aligned} PEN_m &= \int \left[D^m c^T \Phi(t)\right]^2 \, dt \\ &= \int c^T D^m \Phi(t) D^m \Phi^T(t) c \, dt \\ &= c^T \left[\int D^m \Phi(t) D^m \Phi^T(t) dt\right] c \\ &= c^T R c \end{aligned} \tag{1.17}$$

Where $R = \int D^m\Phi(t)D^m\phi^T(t)dt$ is a square matrix of order K. Using this expression for the penalty matrix, 1.13 can subsequently be written as

$$PEN_mRSS(y|c) \quad = \quad (y - \Phi c)^T(y - \Phi c) + \lambda c^T Rc \tag{1.18}$$

It can be shown that the smoothing matrix, S_λ, can be written in the form as in kernel smoothing approaches. If the derivative of Equation 1.18 with respect to the coefficient vector c is taken the following expression is obtained

$$-2\Phi^T y + \Phi^T\Phi c + \lambda Rc = 0$$

re-arranging this expression enables an estimate of the coefficient vector to be obtained as,

$$\hat{c} = (\Phi^T\Phi + \lambda R)^{-1}\Phi^T y$$

Multiplying both sides of this equation by Φ it can be seen that

$$\Phi\hat{c} = \Phi(\Phi^T\Phi + \lambda R)^{-1}\Phi^T y = S_\lambda y$$

where S_λ is the $n \times n$ symmetric smoothing matrix.

An alternative approach for applying spline smoothing using a B-spline basis with a penalty is provided in Eilers and Marx (1996). Rather than using the integral of a squared higher derivative of the fitted curve as the penalty, they instead impose a penalty which is based on the difference between coefficients of adjacent B-splines. The authors state that for second order differences, both their approach, and the method described in this section which uses Equation 1.12 as the penalty, are very similar.

1.3.4 Model Comparisons

It is often of interest to compare pairs of competing models in order to assess which model provides a better fit to the data and what components should be retained or dropped. One way to do this is through the use of an approximate F-test. Suppose there are two nested models; the full model, mod_1, which has degrees of freedom for error, d_1, and the reduced model, mod_2, which has degrees of freedom for error, d_2. To compare these models, an F-statistic is computed and

subsequently compared to an F-distribution with $(df_2 - df_1)$ degrees of freedom for error. Hastie and Tibshirani (1990) advocate the use of an approximate F-test to do this and define the test statistic used in this procedure to be

$$F = \frac{(RSS_2 - RSS_1)/(df_2 - df_1)}{RSS_1/df_1}$$

where RSS_1 and RSS_2 are the residual sums of squares of the two models, mod_1 and mod_2 respectively.

Hastie and Tibshirani (1990) define the degrees of freedom for independent errors of a model in terms of the trace of the smoothing matrix S. For additive models (such as Model 1.6) the sum of the k component projection matrices $\mathcal{P} = \sum_{j=0}^{k} \mathcal{P}_j$ is equivalent to the smoothing matrix S for univariate and bivariate nonparametric regression models. These can be written as;

$$df \quad = \quad n - tr(2S - SS^T) \tag{1.19}$$

To compute the residual sum of squares S can also be used. For each independent model the residual sum of squares can be expressed as;

$$RSS = y^T(I_n - S)^T(I_n - S)y \tag{1.20}$$

Throughout this thesis the approximate F-test approach will be used to compare pairs of nested models. An alternative model comparison procedure is outlined in Bowman and Azzalini (1997) who discuss the use of an F statistic which is defined in terms of quadratic forms.

Choice of Smoothing Parameters

The choice of the amount of smoothing to apply in nonparametric regression models is an key issue. When comparing pairs of models it is particularly important that both models have smoothing parameters that are equivalent in order to ensure a fair comparison. Choosing a smoothing parameter that is too small will 'under-smooth' the data and will result in an estimate that follows the observed data closely and has high variation in local areas. Conversely, if the smoothing parameter chosen is excessively large, a high proportion of the observed data will

contribute to the estimate at each point and therefore the data could be 'over-smoothed', meaning some of the curvature in the data may be missed. A range of different methods are available for deciding what the optimal degree of smoothing should be. These comprise two main approaches; subjective selection or automatic procedures.

There are a range of automatic procedures for the selection of smoothing parameters. Three commonly used model selection criteria are Akaike's Information Criterion (AIC) (Akaike, 1973), a corrected version of this statistic is known as AICc (Sugiura, 1978), and the Bayseian Information Criterion (BIC) (Schwarz, 1978). Writing the sample size as n, the number of the parameters in the model as n_{par} and the maximized value of the likelihood function for the estimated model as L, then AIC can be defined as

$$\text{AIC} = 2n_{par} - 2\log(L), \tag{1.21}$$

AICc can be defined as,

$$\text{AICc} = \text{AIC} + \frac{2n(n_{par}+1)}{n - n_{par} - 1} \tag{1.22}$$

and BIC can be defined as

$$\text{BIC} = -2\log(L) + n_{par}\log n, \tag{1.23}$$

In order to determine the optimal degree of smoothing, each of these methods is computed as the residual deviance of the model penalised by adding a measure based on the number of parameters in the model. The penalty based on the number of model parameters for BIC is stronger than that imposed on the AIC equivalent. For AICc, a second penalty is imposed which is added in order to take small sample sizes into account, however it should be noted that AICc is often used regardless of the sample size since, as sample size increases, AICc will converge to AIC. Alternatively, a frequently used non parametric choice for selecting the quantity of smoothing that is optimal is Generalised Cross Validation (GCV) (Craven and Wahba, 1979).

Ordinary leave-one-out cross validation (OCV) works by leaving out each observation one at a time, and estimating the average smoothed value of the point which has been omitted using the remaining $n - 1$ points. A more generalised

version is K-fold cross validation where the data is split into K subsets and each subset (rather than a single point) acts as test data in turn. Choices of K can be made to reduce the computational burden. An OCV sum of squares is obtained by calculating the average sum of squared differences between the missing data/-datum and its predicted values. OCV sums of squares can then be calculated for a suitable range of different smoothing parameter values. It is clear however that leave-one-out OCV is computationally expensive as the model has to be fitted n times for each smoothing parameter value considered. In addition, the use of leave-one-out OCV is not recommended as although it is technically unbiased, it can be highly variable. Generalised cross validation, which was developed by Craven and Wahba (1979), overcomes this problem as you only need to fit the model once with the full data. From a fitted non parametric model with smoothing parameter λ, the smoothing matrix can be used to compute the effective degrees of freedom for the parameters, $df_\lambda = \text{trace}(S_\lambda)$, and the degrees of freedom for the error, $n - df_\lambda$, where, as before, n is the number of observations. Denoting the corresponding residual sum of squares as RSS_λ then the GCV value can be defined as

$$GCV_\lambda = \left(\frac{n}{n - df_\lambda}\right)\left(\frac{RSS}{n - df_\lambda}\right). \tag{1.24}$$

Plots of λ versus GCV_λ can be used to determine the optimal value of λ. GCV is discussed in detail in Hastie and Tibshirani (1990). More recently, Wood (2011) has discussed the use of restricted maximum likelihood for selecting appropriate smoothing parameters for generalised additive models as an alternative to GCV.

AIC, AICc, BIC and GCV are all methods which rely solely on the data to select an optimal value of smoothing. The advantage of these methods is that they provide a data-driven solution to which smoothing parameter gives the 'best' trade-off between roughness and capturing the main features of the observed data, but these approaches do have their drawbacks. One problem is due to the fact that often automatic procedures for the selection of smoothing parameters assume that the data, and hence any errors from fitted models, are independent and so the presence of correlation in the data provides a further complication. Opsomer and Yang (2001) explores the effects of correlation on smoothing parameter selection for non-parametric regression and states that the presence of correlation between the errors can cause automatic smoothing selection methods such as GCV to break down. Likewise, Diggle and Hutchinson (1989) found that GCV

frequently under-smooths the data, sometimes to the extent that a perfect interpolated fit is produced if there is first order autocorrelation in the data. Further to the problems associated with the presence of correlation in the data, Hall and Johnstone (1992) note that the cross-validatory choice of smoothing parameter can be highly variable and unstable. Ramsay et al. (2009) do not recommend relying only on automatic selection procedures and state that often, near the minimising value, GCV values will change very little thus indicating that the data are not particularly informative as to the true optimal value of the smoothing parameter. They suggest that taking a subjective approach and using judgment as to what values of the smoothing parameter provide a smooth function which can be sensibly interpreted is a reasonable alternative. Faraway (1997) also suggests a good method of selecting a smoothing parameter may be taking a subjective approach as automatic selection methods can be inconsistent, particularly in the presence of unusual observations.

The amount of flexibility used in nonparametric models can be specified subjectively. One method is to use a visual assessment where graphical representations of the smooth functions estimated with a range of different potential smoothing parameters are first obtained and then used to visually assess what amount of smoothing is appropriate. Alternatively, a sensitivity analysis could involve assessing how F-test results change with changes in smoothing parameters. In addition, an approximate number of degrees of freedom can be chosen to define how complex the model is and the smoothing parameters are then set in order to obtain this pre-specified number.

Therefore, meaningful models can be estimated as smooth functions for trend and seasonality, however, the presence of other features such as correlation is still an issue which causes complications when modelling environmental data.

1.4 Statistical Issues in Environmental Data

In addition to accounting for the possibility of autocorrelation, there are several other common challenges in the statistical analysis of environmental and ecological datasets. Issues arise due to missing and irregular observations, non-constant variance and samples which are affected by the limits of detection of scientific equipment, and these must be addressed in some way to ensure that the conclusions

reached for any analysis are valid. A brief description of some of these potential issues, and ways in which they can be dealt with, is provided in this section.

1.4.1 Correlation

Correlation is frequently encountered in environmental data and so, in models where correlation is present it can either be modelled explicitly using ARIMA and SARIMA models, or it can be viewed as a nuisance parameter and appropriate modification can be made to standard techniques. If correlation is present in the data the effective sample size of the dataset will decrease, and hence the size of the errors corresponding to model parameters estimated from that dataset will increase. Giannitrapani et al. (2005) discuss an approach to incorporating correlation into additive models through adjusting the model fitting procedure, but note that the main impacts of correlation are in the calculation of standard errors and in model comparisons. McMullan et al. (2007) explores the use of the standard non-parametric regression models, which represent valid estimates even in the presence of correlated data, and subsequently adjust the standard errors for correlation. Giannitrapani et al. (2005) states that the residual sums of squares given in Equation 1.20 and 1.19 can be adjusted for correlated errors as follows;

$$\begin{aligned} RSS &= y^T(I_n - S)^T V^{-1}(I_n - S)y \qquad (1.25) \\ df_{err} &= tr\{(I_n - S)^T V^{-1}(I_n - S)V\} \\ &= n - tr(S^T + V^{-1}SV - S^T V^{-1}SV) \qquad (1.26) \end{aligned}$$

Where V denotes a correlation matrix. In practice, V is unknown and so it is necessary to estimate the correlation structure using the residuals from the model which assumes the observations are independent and identify a suitable structure for the error component, for example, an AR(1) process. In this thesis, for non-parametric regression models that have been fitted using the kernel smoothing techniques discussed above, temporal correlation, if present, has been incorporated by adjusting the standard errors and using the modified degrees of freedom and residual sum of squares shown in Equations 1.25 and 1.26.

For spline smoothing approaches, the residual sum of squares can also be adjusted to incorporate correlation. Letting Σ denote the variance-covariance matrix, then the least squares criterion in Equation 1.15 can be extended to deal

with correlation in the residual sum of squares as below,

$$RSS(c|y) \quad = \quad (y - \Phi c)^T \Sigma^{-1} (y - \Phi c)$$

Setting Σ to be the identity matrix I produces the standard residual sums of squares equation. Conversely, letting Σ^{-1} be any other positive definite symmetric weight matrix W results in weighted least squares.

1.4.2 Non-constant Variance

It is common that preliminary analysis for many environmental variables will indicate that the distribution of the data is positively skewed, or that the variability of the determinand across the time period is not stable. Consequently many variables are often natural log transformed either to stabilize the variability or to satisfy parametric test assumptions which require that the data are normally distributed. It is worth noting that there are several reasons why the data could be skewed other than the possibility that they have arisen from an underlying log normal distribution. For example, the presence of one or more outliers and bimodality could influence the calculation of the mean and the symmetry of the distribution. In addition, what appears to be a change in the variability of the data over time could, in practice, be due to a change in the limits of detection, or could be a due to a change in the seasonal pattern over time. The suitability of the log-normal distribution is discussed both in Singh et al. (2007), with reference to its use in general environmental contexts, and in Chalwa and Hunter (2005) which is specifically related to its use as a basis for classification of bathing waters. Chalwa and Hunter (2005) considered using parametric percentile values to assess the classification of Irish bathing water sites and concluded that using this method to gauge compliance was statistically unreliable due to failure of the log-normality assumption at many beaches. This suggests that the need for a log-normal transformation should be carefully considered in combination with other aspects of the data and should not be applied as a matter of course.

1.4.3 Missing Data

Several standard statistical techniques that have been designed to analyse ecological data require observations to be collected, at regularly spaced time intervals

with no gaps. For example, when computing the ACF it is assumed the data are equally spaced and that there are no missing observations. However, in practice it is unlikely that data will be complete, and missing data are commonplace, meaning that strategies for dealing with missing data often need to be employed before any analysis can be carried out.

There are several reasons why data may be missing and the 'Missing Data Mechanism' (Rubin, 1976) describes the mechanism by which missing data may have arisen. The mechanism in operation is dependent on whether or not there is a link between the missingness and the underlying values in the dataset. There are three main Missing Data Mechanisms;

- Missing Completely at Random (MCAR), where the probability of a value being missing is unrelated to either the observed or unobserved elements of the data.
- Missing at Random (MAR), where the probability of a value being missing may be related to the observed elements of the data but not to the unobserved elements of the data.
- Not Missing at Random (NMAR), where the actual mechanism which caused the missing data is systematic and informative and hence has to be examined and modelled appropriately.

It is important, is possible, to determine the mechanism which is relevant to any missing data in a statistical analysis as the approach to analysing the data may differ depending on the this. However, while there are a number of potential reasons as to why data are missing, in situations where the underlying ecological system is complex the nature of every missing observation often cannot be determined. In the sampling of environmental data it is becoming increasingly common for monitoring calendars to be established in advance of any monitoring season but adverse weather events can mean sites are inaccessible which prevents samples being collected, particularly in the winter months. Another potential source of missing data is failure of scientific equipment used to analyse samples and samples becoming lost or damaged in transit. It is often the case that monitoring networks change in size throughout time and additional stations entering a network can cause problems due to differences in the quantity of data available at different locations.

There are numerous missing data techniques (Little and Rubin, 1987) the majority of which are designed to impute values in place of the missing observations, creating the regularly spaced datasets which are necessary for the application of traditional methods of statistical analysis. In general, approaches for dealing with missing data can be split into two broad categories; single imputation, where one value is generated in place of each missing value and multiple imputation, where several values are generated for each missing value. Multiple imputation aims to reflect the uncertainty associated with the missing values however Plaia and Bondi (2006) states that single imputation methods have the advantage of only having to generate one value for each missing observation and mean that standard complete data analysis techniques can be applied directly after the missing data values are in place.

Engels and Diehr (2003) discuss several approaches for dealing with missing data in longitudinal studies. One method considered is the 'last observation carried forward, next observation carried backward'. This essentially means for each individual the missing value is replaced with an interpolated or average value of the preceding and successive known value. While this method is somewhat ad hoc, it is easy to implement and was found by Engels and Diehr (2003) to be highly effective, particularly if there is a strong individual specific component to the data. An assessment of single imputation methods in the context of environmental data is provided in Plaia and Bondi (2006). Here, the authors use information from both the site, and the time point where the observation is missing to simulate an appropriate replacement value and provide an application of the method to air pollutant concentrations.

1.4.4 Limits of Detection

In the analysis of environmental data, problems can occur with some samples due to the equipment used to take measurements. Scientific equipment often has saturation levels either above, or below which the exact quantity cannot be confirmed. The levels above or below which it is not possible to determine the exact value are known as limits of detection (LOD) and there is some question as to how to treat these values, which are effectively right or left censored observations. SEPA include half the stated LOD value in any analysis, however this is just one approach for dealing with this issue. Clearly, it would invalidate any conclusions reached to

simply ignore values which are affected by limits of detection and analysis must take them into account. Eastoe et al. (2006) investigated different ways of handling censored observations in an environmental context concerning air pollutants and indicated that it was necessary to incorporate censored observations in any analysis rather than ignoring them altogether. On the other hand, Helsel (1990) advises against the method of simply substituting non-detect observations with a nominal fixed constant.

Changes in the limits of detection due to the introduction of new, more sensitive scientific equipment can also introduce problems. For example, a reduction in the limit of detection may not only falsely indicate there is a trend in the data, but could also give the appearance of a change in the seasonal pattern over time. If relatively low values of a particular determinand occur in the warmer spring and summer months then this is the time when values are most likely to be recorded below detection limits. An increase in the minimum recorded values in the summer months across the time period could be determined as a change in the seasonal pattern across the years, when in reality, this is just a further by-product of changing limits of detection. This re-enforces the importance of taking these censored values into account.

Three possible statistical methods that can be used for the analysis of datasets which include left censored observations are described briefly below. The general idea behind each of the methods is to estimate summary statistics for the distribution of the data which takes into account the censored observations which are present. Using this estimated distribution, values are simulated, subject to the constraint that they fall below the stated limit of detection values. The values which are generated are subsequently imputed in place of the censored observations.

Kaplan-Meier Estimator

This is a non-parametric method which is often used in survival analysis for estimating the summary statistics for data where there are right censored observations. It can also be applied to data where there are left censored observations by 'flipping' the data and subtracting them from a fixed constant. Using this estimator in this way, in the context of non-detects, was first suggested in Helsel (1990)

The Kaplan-Meier estimator estimates the survival function, which maps the probability that observations will survive onto time. In the context of survival analysis the Kaplan-Meier estimator estimates the probability that observations will survive beyond certain time points and this can be translated into the context of left censored observations as being the probability that observations will fall below the limits of detection. Summary statistics of this distribution, which by its definition takes into account the loss of information from limit of detection values, can then be found. Helsel (2005) recommends using this non-parametric estimator in situations where there are less than 50% censored observations and more than 50 observations to estimate the summary data. As well as not requiring any distributional assumptions, an additional benefit of the Kaplan-Meier estimator is that it is suitable where there are multiple detection limits as there often are in water quality data.

Maximum-Likelihood Estimator

This is a parametric method which requires the specification of a distribution which is a close fit to the observed data. The parameter estimates obtained describe a distribution with the maximum likelihood of producing a dataset with the observed detected values and the proportion of censored data which falls below each of the stated detection limits.

One of the potential problems of this approach arises if the distribution of the data is poorly specified. In this case the maximum likelihood estimator approach can produce estimates which are incorrect. When the data are thought to have arisen from a log-normal distribution logarithms of the raw data are taken and a Normal distribution is specified so that maximum likelihood procedures can be used. As stated in Shumway et al. (2002) this can cause problems since the parameter estimates produced using this method are on the log-transformed scale and process of back-transforming can potentially produce estimators that are quite severely biased due to the non-linear relationship between the different scales.

Regression on Order Statistics (ROS)

This is a semi-parametric method for computing summary statistics of a distribution where there are left censored non-detect data. Shumway et al. (2002)

assessed this method and stated it was a reliable approach to take when dealing with data where there are values which are marked as being at the limit of detection. Within this method, left censored observations are modelled using a linear regression model of the observed un-censored values against their normal quantiles. A brief description of the ROS method is provided below, notation has been taken from Shumway et al. (2002).

Suppose there are n_0 observations $y_i, i = 1, \ldots, n_0$(log transformed or otherwise) which are below the limit of detection U and n_1 observations $y_i, i = n_0 + 1, \ldots, n_0 + n_1$ which are above U. Assuming these observations are independent and normally distributed with mean μ and variance σ^2 then the mean and variance will satisfy the equation

$$y_i = \mu + \sigma\Phi^{-1}(P_i) \tag{1.27}$$

where $P_i = Prob\{Y_i < y_i\}$ and Φ^{-1} denotes the inverse of the cumulative normal distribution. Applying a linear regression to the normal scores for the complete case observations would then enable the mean and variance of the observations to be obtained. Shumway et al. (2002) state that the accepted procedure is to replace the probabilities by the adjusted ranks in Equation 1.27 so that the regression equation becomes

$$y_i = \mu + \sigma\Phi^{-1}\left(\frac{i - 3/8}{n + 1/2}\right) + \epsilon_i \tag{1.28}$$

where $i = n_0 + 1, \ldots, n_0 + n_1$ and the errors, ϵ_i, are assumed to be independent and and have equal variance. The estimates of μ and σ can be obtained using least squares. Using Equation 1.28 predicted values can be obtained for observations y_i where $i = 1, \ldots, n_0$. If a transformation has been applied, the set of all observations (uncensored and predicted) can be back-transformed and the mean and variance can be worked out on the original scale.

Similarly to the Maximum-Likelihood estimator approach to determining appropriate summary statistics, there is some concern over the correct specification of the distribution of the response variable. The ROS method requires the same key assumptions as linear regression; that the response is a linear function of the explanatory variable or variables, and that the variance is constant. However, it is extremely common in environmental contexts that the variables of interest are skewed and, as a result of this, a log transform of the data is often taken prior to

application of the ROS method. Helsel (2005) recommends this method for use where there are up to 80% of observations listed as non-detects.

Throughout this thesis the ROS method will be used to impute suitable values for observations which have been affected by limit of detection issues. The ROS method has been applied using the `NADA` (L., 2012) package in `R`.

1.4.5 Statistical Power

It is of great importance to ensure that any statistical analysis is based on a quantity of data that is sufficient to achieve an acceptable level of statistical power. The power of a statistical test is defined as $1 - \beta$ where β is the probability of making a type II error. A type II error is the probability of not detecting a difference or change which does in fact exist. In addition, the statistical size of the test, known as type I error, can be defined as the probability of identifying a difference or change which does not exist. While focus is often placed on the statistical size and the type I error is commonly set as 5%, both of these error rates have to be considered. A high statistical power is meaningless when the size of the test is also high as this indicates, for example, that a trend could be detected, both when there is a true underlying pattern, and when there is no relationship. These quantities have to be taken into account when designing any monitoring programme to ensure reliable interpretation of the results of any subsequent analysis. Further discussion of the power of environmental monitoring programmes is presented in Chapter 2.

1.5 Aims and Objectives

The principal aim of this thesis is to use and develop statistical analysis to investigate commonly used environmental monitoring networks so that the design and implementation of future networks can be made as effective and cost efficient as possible. Using data which have been provided by SEPA, rivers and lake data and a range of determinands will be considered in order to explore water quality monitoring in Scotland. The importance of understanding changes in water environments, combined with the mandatory classification of water bodies required by policy, motivates the key objectives of this thesis. These objectives include;

1. To investigate, via the development and implementation of a simulation study, the statistical power of several common sampling schemes;

2. To explore the current group structure used by SEPA for classification of standing waters and to assess how well existing SEPA groups capture differences between the lakes in terms of several chemical variables which are used for WFD classification;

3. To explore the effects of autocorrelation on our ability to distinguish between groups of lakes;

4. To investigate and develop alternative statistical approaches for grouping observed chemistry data based on the temporal dynamics of the variables of interest;

5. To explore and develop statistical techniques to group different types of water bodies.

Chapter 2

Assessing Statistical Power to Detect Change

It is of importance to ensure that any environmental monitoring programme which is implemented can lead to meaningful results. As well as societal pressure to document changes in environmental indicator variables, there are increased legislative requirements and, under the WFD, regulatory agencies need to indicate the level of confidence and precision of the results provided by their monitoring programmes. The necessity for meaningful and considered sampling programmes, that have a sufficient level of statistical power to detect underlying changes is widely acknowledged in the literature (Nicholson and Fryer, 1992 Field et al., 2007, Legg and Nagy, 2006). Legg and Nagy (2006) claim that the results of inadequate monitoring can be both misleading and dangerous, not only because of their inability to detect ecologically significant change, but also because they can create the illusion that improvements are working when they are in reality having little effect. This in turn can result in an inefficient use of financial resources, which are becoming increasingly limited.

Field et al. (2007) states that obtaining adequate statistical power is the cornerstone of any rigorous monitoring programme and clearly the designers of environmental monitoring programmes want to achieve as high a level of statistical power as possible. Although a high statistical power is desired, there is an obvious trade off with the level of statistical power of a sampling programme and the Type I error rate. The Type I error rate is usually fixed at 5% and so it is important to ensure that with this statistical size fixed, a satisfactory level of power is also

achieved. In addition to the relationship between statistical power and statistical size, there are several factors which affect power to detect trend such as sampling variability, sampling frequency, the effect size and the presence and strength of autocorrelation in the data. Although some of these factors are an inherent part of the system under study there are practical constraints such as budget and time considerations which dictate how many samples are collected.

There are a number of existing studies in the literature which explore the power of environmental sampling programmes. For example, Gerrodette (1987) investigates methods for calculating the power of detecting linear and exponential growth and decline in animal populations, while Di Stefano (2001) considers ways in which power analysis can be incorporated into monitoring programmes used in the context of forest management. Further to this, Keizer-Vlek et al. (2012) attempts to quantify spatial and temporal variation in macroinvertebrates in lakes in the Netherlands via the implementation of a simulation study. The authors calculate the number of sites required to detect a relatively large change in the frequency of collection of individual species. A slightly different approach is taken by Howden et al. (2011) who investigate the length of time series required to detect changes in water quality parameters, including dissolved organic carbon and nitrate, by subsampling from an existing long-term dataset and calculating the length of time series required to detect the known underlying trend in the data.

Another example where a simulation study is used to calculate power to detect change is provided in Field et al. (2005). Here the authors employ a simulation study in order to investigate statistical power of a monitoring programme to measure occupancy of species in a landscape. The magnitude of a linear decline over time was fixed and subsequently the effect of both different numbers of sampling sites, and different numbers of visits to each site on the statistical power to detect this change was explored. Only linear trends were considered in this example and it was assumed there was no spatial or temporal correlation in the data. It is noted by Nicholson and Fryer (1992) that in contaminant monitoring the changes of interest are not necessarily linear and that there is often no straightforward way to assess power when the change over time is non-linear.

It is the aim of this Chapter to first consider the type of features which are common in existing water quality time series that are used to compute WFD

classification, and then to investigate the power to detect different forms of underlying change in similar time series under a range of different sampling and data conditions.

2.1 Case Study

In order to obtain some background of the type of features commonly seen in available water quality data, a set of observations from three Scottish lakes was considered. The aim of this exploratory analysis was to look at typical features of existing time series for water quality variables that are of the most interest in the context of monitoring standing waters within the Water Framework Directive. Total Phosphorus (TP) and orthophosphate (OP) measurements, both in milligrams per litre (mg/l) were available for Loch Voil, Linlithgow Loch and Lake of Menteith. The locations of these three sites are shown on a map of Scotland in Figure 2.1. Each of these sites was selected as a case study site for this analysis due to the length of time series available. At Loch Voil and Lake of Menteith the data covers a 25 year period from 1984 to 2008, although at Linlithgow Loch the data covers a shorter period of 16 years from 1993 to 2008. The estimates obtained will be used as the basis of the simulation study discussed later in the Chapter, to investigate power to detect long-term change. The analysis for each lake is primarily an illustrative one, and so only selected results will be shown to demonstrate key characteristics of environmental data.

Table 2.1 contains the number of samples available for each of the determinands at the three sites and the percentage of samples which are marked as being at the limit of detection. As can be seen, OP measurements at Loch Voil were severely affected by limit of detection problems, with over half of all recorded sample values falling below the detection limits. For this reason the OP measurements at Loch Voil were not considered any further. All other sites and variables were either unaffected, or had only a small percentage of observations marked as being at the limits of detection and hence any of the observations identified as being below the detection limits were imputed using the regression on order statistics method described in Section 1.4.4.

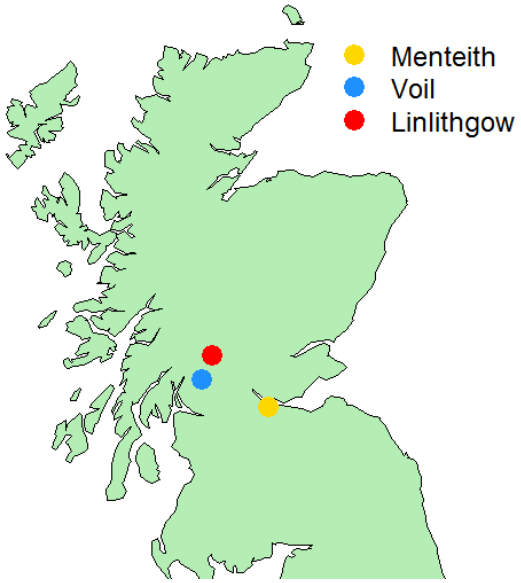


FIGURE 2.1: Map of Scotland showing location of Lake of Menteith, Loch Voil and Linlithgow Loch

	OP	OP LOD	TP	TP LOD
Linlithgow	202	7.4%(15)	220	0%
Menteith	239	10.4%(24)	241	0%
Voil	177	56.5%(100)	152	2.6% (4)

TABLE 2.1: Summary of available orthophosphate (OP) and total phosphorus (TP) data at Scottish lakes

Distribution of TP and OP

As with many environmental variables, initial plots of the distributions of TP and OP indicated that a transformation was required and that a natural log transformation was necessary to stabilize the variance of the values. Figure 2.2 shows a boxplot, histogram and normal Q-Q plot for the original TP sample values. It can be seen from this that the distribution of the data are highly skewed and there is a high level of curvature in the Q-Q plot obtained. This indicates that the data are not normally distributed. Figure 2.3 displays the same data, but this time using the natural log transformed TP levels. The variance has clearly been stabilized by the use of the log transformation as the spread of the data is far more symmetric around the mean value. It can also be seen that after log transforming the data

there is a much greater level of agreement between the data and the theoretical distribution under the assumption of normality. This can be seen in both the histogram and the normal Q-Q plot of the logged values. For this reason, all further simulations and analysis in this Chapter will be based on log transformed values.

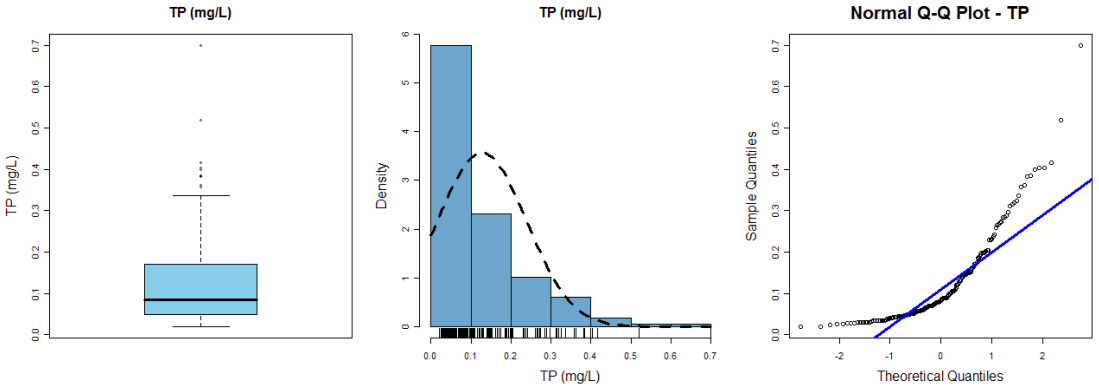


FIGURE 2.2: Distribution of TP(mg/L) at Linlithgow Loch

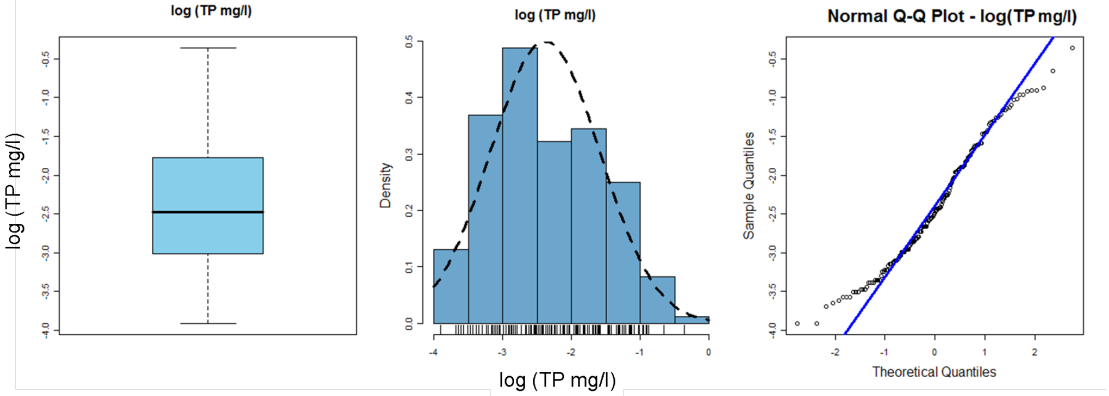
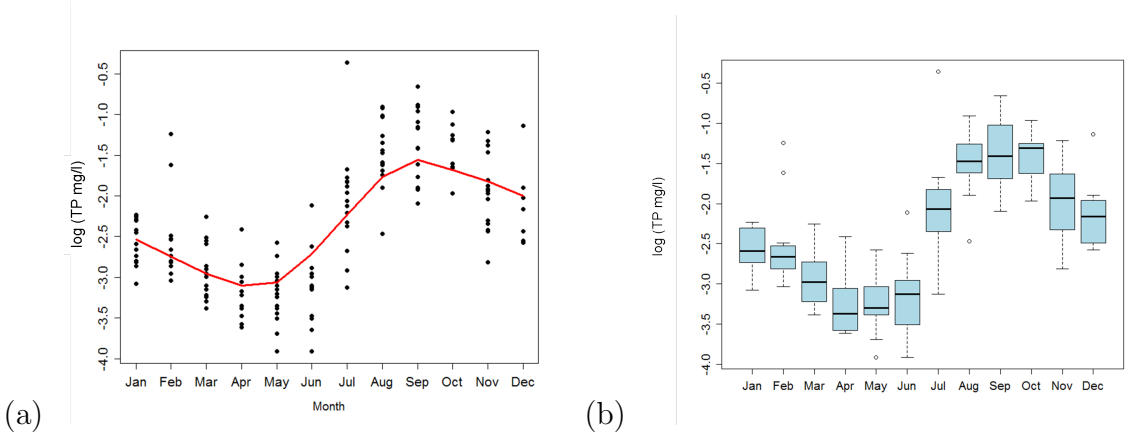


FIGURE 2.3: Distribution of log(TP, mg/l) at Linlithgow Loch

Seasonal patterns

To identify seasonality in the lakes data, two exploratory plots were produced. The first of these was a plot of the log transformed values against the month of year. Secondly, a set of boxplots for the data against month of year was also considered to examine the distribution of the samples within each month. The plots for TP at Linlithgow Loch are shown in Figure 2.4.

For the seasonal scatterplot shown in Figure 2.4(a), a loess smoothed fit was included in red to indicate the general pattern across the years and the pattern within each year. From this scatterplot there does appear to be a strong seasonal pattern present across each year. There is a clear change in the levels of TP throughout the year with the lowest values being recorded in the late spring and early summer months and the highest values in the autumn and winter months. The monthly seasonal pattern is also evident from boxplots of the TP levels at Linlithgow Loch for each month of the year shown in Figure 2.4(b). In addition, it can be seen here that the variance across the months appears to be reasonably constant. The presence and strength of this seasonal component within the TP data at Linlithgow indicated that it is necessary to include a seasonal component within our simulations in order to reflect the real data situation.

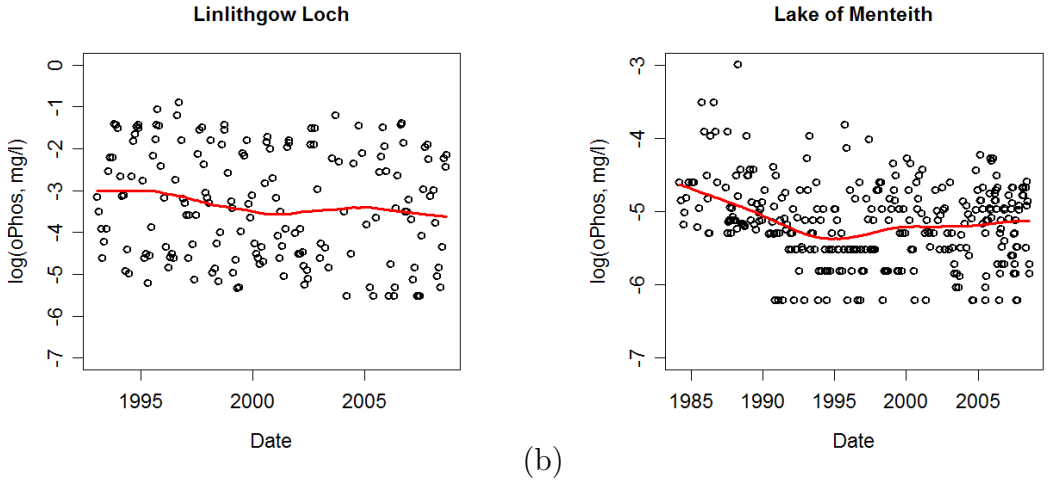


(a) (b)

FIGURE 2.4: Plot of log(TP, mg/l) vs. month (a) and monthly distribution of TP (b) at Linlithgow Loch

Trends

Figure 2.5 displays plots of natural log transformed OP (mg/l) against time at Linlithgow Loch (a) and Lake of Menteith (b). A loess line has been added in red to indicate the general form of the trend over time. At Linlithgow Loch there appears to be a linear trend over time, while the form of trend at Lake of Menteith is less clear. Although non-linear trends are often suitable, in view of Figure 2.5 (a) it was thought that a parametric model should be fitted in order to assess if



(a) (b)

FIGURE 2.5: Plot of log(OP, mg/l) at Linlithgow Loch (a) and Lake of Menteith (b) with loess line to indicate trend

there is any linear trend present in the data. This model was of the form

$$y_t = \mu + \beta x_t + \gamma \cos\left(\frac{2\pi \text{month}_t - \phi}{12}\right) + \varepsilon_t \quad \text{where } \varepsilon_t \sim N(0, \sigma^2) \tag{2.1}$$

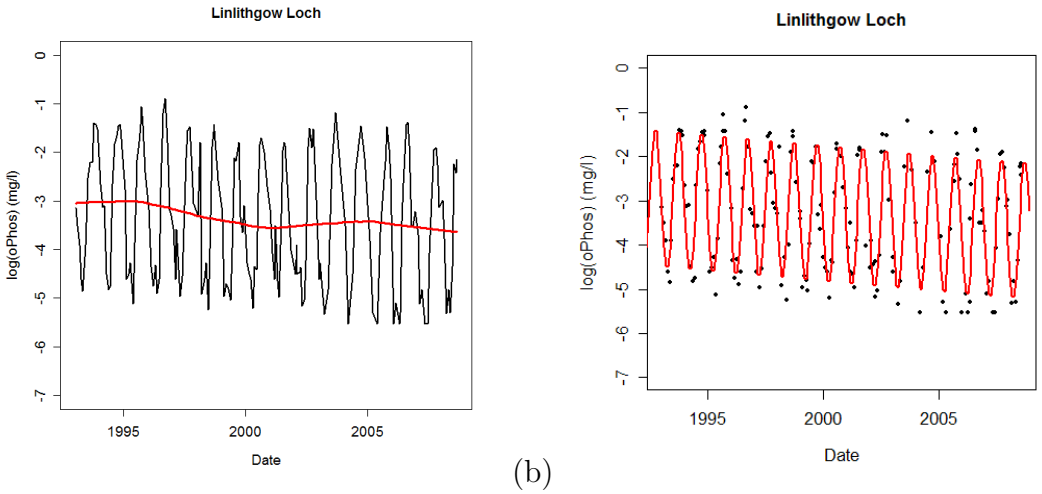
The parameter β gives an estimate of the size of linear trend in the data while the trigonometric term is used to model the seasonal pattern. The phase of the seasonal pattern is represented by ϕ while the parameter γ represents the amplitude of the seasonal pattern. Month is the month of the year (1-12) and the time corresponding to observation t is denoted by x_t and included in decimal year form which has been calculated using the formula 'year + $\left(\frac{\text{month}-1}{12}\right)$'. While the model given in Equation 2.1 is nonlinear, it can be expanded to be written as;

$$y_t = \mu + \beta x_t + \gamma_a \sin\left(\frac{2\pi \text{month}_t}{12}\right) + \gamma_b \cos\left(\frac{2\pi \text{month}_t}{12}\right) + \varepsilon_t \quad \text{where } \varepsilon_t \sim N(0, \sigma^2) \tag{2.2}$$

Here γ_a and γ_b give an estimate of the parameters related to the seasonal pattern over time. An estimate of the amount of variability in the data was obtained by calculating the estimated variance of the fitted model residuals. Although this model assumes that the errors are independent, the error term can be modified to account for correlation as described in Section 1.4.1.

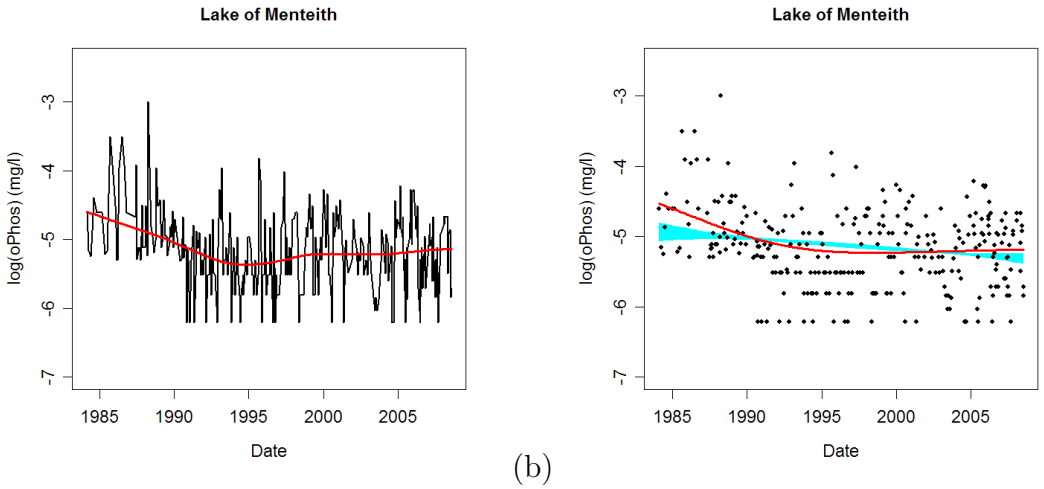
Figure 2.6 (a) shows a time series plot of log OP concentrations at Linlithgow Loch with a loess smooth function shown in red to indicate the general form of the trend. It can be seen that there appears to be a small decrease in log OP concentrations over the time period of interest. A linear model of the form shown in Equation 2.2 was fitted to the data and is shown in Figure 2.6(b). The parameter β was estimated to be -0.04 and was found to be statistically significant at a 5% significance level.

In addition to the linear model described in Equation 2.2, non-parametric models such as those described in Section 1.3.2 can be used. The model described in Equation 1.3 can be used to see if there is any general form of smooth trend over time while the additive model shown in Equation 1.7 can be used to investigate a non-parametric trend and a constant seasonal signal. The bivariate model shown in Equation 1.5 can be used to investigate if there is a smooth, non-parametric trend and a seasonal component which can change over the time period considered. A time series plot for log OP concentrations at Lake of Menteith is shown in Figure 2.7(a). It is the clear that seasonal pattern was not as strong as at Linlithgow and the form of trend does not appear to be linear. In addition, although any observations which have been marked as being at the limits of detection have been imputed using the regression on order statistics methods, there appears to be several observations at the same value. It is thought this could be due to observations being rounded. A local linear regression model of the form shown in Equation 1.3 was fitted to the OP data at Lake of Menteith using the sm.regression package in R (Bowman and Azzalini, 2010). A hypotheses test was carried out to test the null hypothesis that there is no relationship between log OP and time at this site, against the alternative that there is some smooth relationship. A similar test was then performed to assess if the relationship was linear or non-parametric. The p-values for these tests indicated that there was a relationship between LOG OP and time at Lake of Menteith and that the relationship was non-linear. Figure 2.7(b) again shows the log OP values plotted against time, with the estimated local linear regression shown in red and a blue reference band which corresponds to the null hypothesis that the trend in OP over time is linear. The shape of the local linear regression model suggests that there is an initial decrease in OP concentrations at Lake of Menteith, which levels off and remains reasonably constant over the following years.



(a) (b)

FIGURE 2.6: Plots of log(OP) at Linlithgow Loch with loess smooth (a) and with fitted linear model (Equation 2.2) (b)

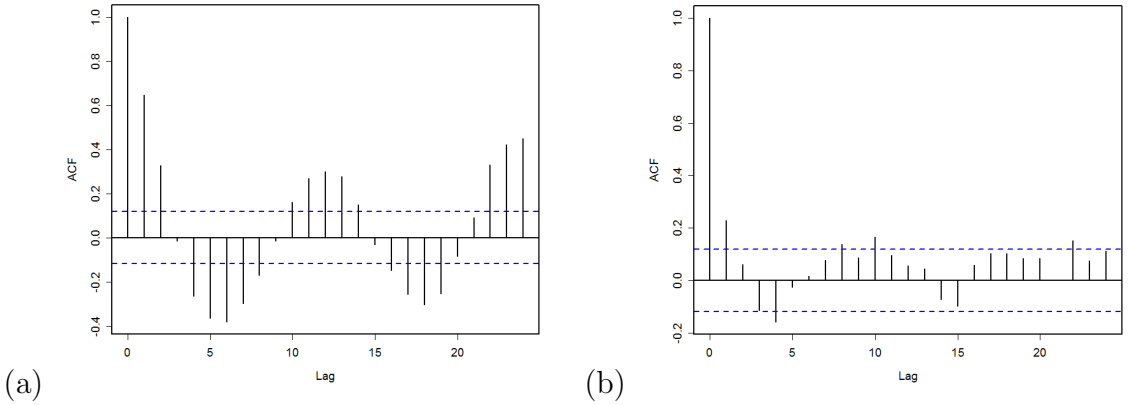


(a) (b)

FIGURE 2.7: Plots of log(OP) at Lake of Menteith with loess smooth (a) and with fitted non-linear model (Equation 1.3) (b)

Correlation

In addition to investigating trends and seasonal patterns another feature that is of interest is the presence and strength of correlation in the data. The samples from the lakes considered here were, in general, collected at monthly or two month intervals, meaning that almost all samples were more than two weeks apart. Exploratory analysis of the TP data at Linlithgow Loch was carried out to investigate if there was any evidence of autocorrelation. To assess the dependence in the data, the autocorrelation function (ACF) was computed for $\log(TP)$. Prior to calculating the ACF, the data were first deseasonalised by fitting a linear model to the data with the terms $\cos \frac{2\pi \text{month}_t}{12}$ and $\sin \frac{2\pi \text{month}_t}{12}$ where month is the month of the year (1-12). It was thought that fitting these two terms would account for the seasonal component in the data and so subtracting the fitted model values from the original observations would remove the seasonal pattern.



(a) (b)

FIGURE 2.8: ACF plots for log TP at Linlithgow Loch with lag in months

Figure 2.8 shows the ACF for both the original log(TP) series and the deseasonalised log(TP) series at Linlithgow. From Figure 2.8(a) the seasonal pattern present in the original series can be clearly identified. While the signal is much weaker in the ACF of the deseasonalised data series, shown in Figure 2.8(b), there remains some form of underlying structure in the data which has not be captured by the fitted model. Although there is little evidence of any strong temporal correlation in the deseasonalised data, there continues to be a systematic pattern which means the correlation at certain lags lie outwith the confidence bands. This could be an indication that there is a change in the seasonal pattern over time; a feature

which has not been accounted for by fitting the linear model which assumes a constant within-year pattern across the period (Equation 1.7).

Similar ACF plots were produced for the parameters at the other sites and, although not shown, several of the ACFs produced indicated that there was a low level of autocorrelation present in the errors after the seasonal component had been removed by fitting a linear model. To investigate the strength of the autocorrelation in the errors a linear regression model was fitted to the residuals of the model containing the seasonal terms $\cos\left(\frac{2\pi \text{month}_t}{12}\right)$ and $\sin\left(\frac{2\pi \text{month}_t}{12}\right)$, denoted r_i. An AR(1) error structure was assumed and a simple linear model of the form $r_i = \alpha + \beta r_{i-1}$, where $i = 1, ..., n$ was fitted. A significant slope in this model provides evidence of a linear relationship between consecutive residuals and hence that there is autocorrelation present. Typical values of first order correlation coefficients between monthly observations ranged from approximately 0.2 to 0.4.

Summary of Exploratory Analysis

The initial analysis of the available OP and TP data at the Scottish sites revealed that there are a wide range of features which need to be taken into account in the design of a simulation study to explore the power to detect change over time. Both linear and non-linear patterns over time were detected at the case study sites and so it is likely that a wide range of such trends would be encountered in an environmental context. While it was clear that at some sites there were clear seasonal signals present, it was also noted that there may be some evidence that seasonal patterns may change over time. It is possible that changes in seasonal patterns are often mistaken as changes in the variability of an underlying system and so statistical power to detect changes in the seasonal signal over time under different sampling conditions will also be considered. The parameters estimated from the models fitted to the Scottish lakes will be used within the simulation in order to provide some indication as to sensible values for the magnitude of trend, variation and correlation in real datasets.

2.2 Simulation Study

A large simulation study was designed and implemented in order to investigate the effectiveness of different sampling strategies and how different sampling designs affect the statistical power of detecting long-term changes in water quality. The overall aim of the simulation study was to answer questions regarding how many samples chemical classification of standing waters should be based upon. The simulations attempt to mimic real data by incorporating common features of environmental data, such as changing variability, seasonality and autocorrelation. It is of interest to assess the effect of different data scenarios, changing both the structure of the underlying data and the sampling scheme, on our ability to detect long-term change. Our primary interest is to explore how the application of different sampling frames (sampling plans which vary both in length of time series and sampling frequency) will affect our ability to detect an underlying change in the data. Three broad simulation scenarios which will be considered within this study are;

1. Power to detect fixed linear trend
2. Power to detect non-linear trend
3. Power to detect a non-constant seasonal pattern

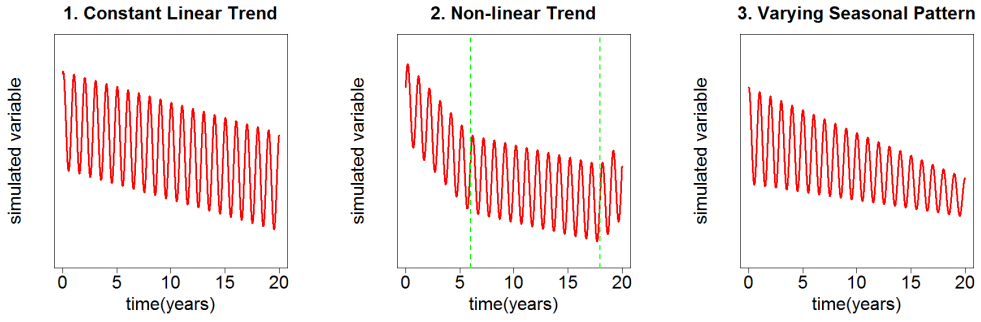


FIGURE 2.9: Shapes of underlying models used to simulate data for each of the three simulation scenarios considered

Figure 2.9 shows the general shapes of each of the three underlying models from which we will simulate our data. The models used to simulate the data in each of these scenarios will be discussed later in more detail. Within each of these scenarios the effects of varying the following features was explored;

- magnitude of trend/pattern
- variability
- strength of correlation

For the non-constant seasonal component simulation, the effects of changing the amplitude of the underlying seasonal pattern were investigated.

As mentioned previously, the empirical power is used to assess the effectiveness of the different sampling strategies explored. The power of the simulation is the probability of detecting a trend or pattern in the subsets of sampled data, given that one exists in the underlying dataset from which the samples were taken. The simulation study will provide a prospective analysis of the power of different sampling schemes, and how the changing features of underlying models affects our ability to detect different forms of change in the data.

2.2.1 Simulation Procedure

For each of the three different simulation scenarios considered, the underlying models used to simulate the data, and the subsequent models fitted to the sampled data, will be changed, while the general procedure for simulation is the same throughout. The simulation procedure used was as follows;

- For each scenario, a large daily dataset was simulated from the corresponding model. Initial model parameters were estimated using existing data provided by SEPA for Scottish lakes as discussed in the previous section. No error term was added to the simulated data at this stage.
- Different sampling frames were then applied. Both the sampling frequency within each year, as well as the length of period over which sampling was carried out were varied. The sampling dates within each time unit (week, month, year etc.) were selected at random using a suitable uniform distribution.
- After the sampling frames are created and the relevant samples selected from the daily dataset, suitable error terms were generated using an AR(1) process in combination with the value for the error variance, and the value for the correlation coefficient. These errors were then added to the simulated values.

- For each combination of model conditions considered (e.g. slope, variability, correlation) and sampling conditions (e.g. sampling frequency, length of time series) 500 simulations were generated.
- To each sampled subset, appropriate models were fitted to assess whether or not the change over time could be detected. The types of model fitted were determined by what the feature of interest was (e.g. linear trend, non-linear trend or change in the seasonal component over time)
- The empirical power of detecting a change over time for each combination of conditions considered was subsequently calculated using the 500 fitted models where the significance level was set to 5%.
- The statistical size (the probability of detecting a change over time when there is no underlying change present) was also calculated for each set of 500 fitted models.

Where possible, existing data were used to set the model parameters used in the simulations. The error term is added on to the simulated data after the sample dates have been generated so that an appropriate correlation structure is present in the simulated data.

Figure 2.10 shows an example of a simulated daily error with an AR(1) structure where the correlation coefficient $\rho = 0.4$. Although there is a correlation in the error at a daily level, due to the exponential decay of an AR(1) term it is unlikely that correlation in the data will be detected when the sampling frequency is weekly or less frequently. To illustrate this, Figure 2.11 shows a set of ACF plots corresponding to weekly, fortnightly and monthly sampled data from the set of data where there is a daily AR(1) error term. None of the plots shown in Figure 2.11 suggest there is any evidence of statistically significant correlation in the underlying data at these sampling frequencies. The finding that if there is daily correlation present, it cannot be detected under sampling frequencies which are weekly or less frequently, agrees with the statement of van Belle and Hughes (1984).

Water quality data of the type of interest here are rarely available at a daily level, and moderate levels of correlation are often observed from data which are collected at a weekly, fortnightly or even monthly frequency. Morton and Henderson (2008) found that monthly samples of stream electrical conductivity had a

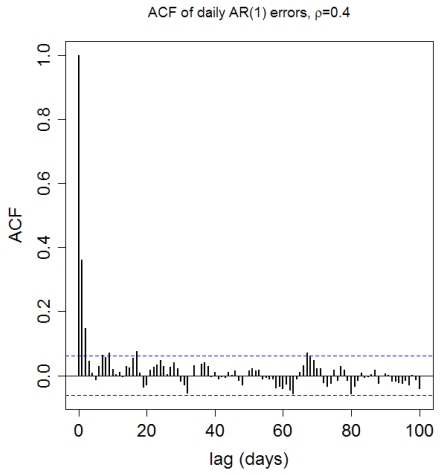


FIGURE 2.10: ACF of daily data with an AR(1) error component

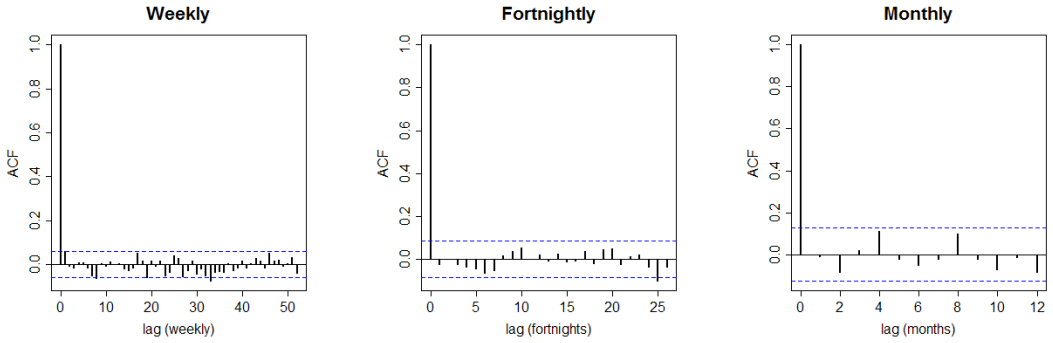


FIGURE 2.11: ACF of sampled data with AR(1) error component

first order autocorrelation structure where the correlation coefficient was greater than 0.5. In addition, Ferguson et al. (2007) also found there to be statistically significant autocorrelation between monthly data in a study of lake water quality determinands at Loch Leven in Scotland. In order to reflect the real life situation, it is assumed our 'collected' samples have correlation which follows an AR(1) structure, rather than the underlying daily dataset.

2.2.2 Sampling Conditions

Currently, SEPA aim to sample standing waters either monthly, or 6 times a year, and so both of these sampling frequencies are considered within this study. However, in addition to these frequencies, it is also of interest to consider weekly,

fortnightly and annual sampling designs, to see the effect this would have on our power to detect the underlying changes. While it is highly unlikely that SEPA will increase sampling frequency at standing waters in the near future, it is useful to include these more 'extreme' sampling frequencies as it enables comparisons to be made between the power achieved for these sampling designs - almost as a 'best case scenario' - and those which are currently used. This could not only allow us to potentially quantify any additional benefits of taking more samples in terms of power to detect change, rather than financial cost, but furthermore could potentially justify the use of less frequent sampling strategies that are presently employed. Different lengths of time series of samples were considered, starting at a minimum length of 5 years and ranging to a maximum length of 30 years. It was thought 5 years of data would be the minimum length of time that was required to ensure there were a sufficient number of samples on which to fit suitable models, particularly given some of the lower sampling frequencies which were being investigated. The upper limit was set as 30 years as this was thought to be, in general, around the maximum length of time for which there are water monitoring data records available.

Within our simulation, sampling frames are designed so that the sample date within each time unit, for example, each week, month or year, is randomly selected. This ensures that our data is as close to real life as possible since SEPA do not monitor at regular, equally spaced time intervals but instead, the dates the samples are collected on can be dependent on a variety of conditions including availability of staff, adverse weather and physical access to the sampling location. In the simulations to assess the effects of temporal correlation between samples on power to detect change only monthly, weekly and fortnightly sampling frequencies will be investigated. It is of interest to assess how temporal correlation in the data will affect our ability to detect long-term change since independent observations over time, particularly those recorded on a daily basis rarely occur in environmental settings due to the lagged effects of confounding variables such as temperature and rainfall.

2.3 Scenario 1 - Fixed Linear Trend

The first simulation scenario considers the simplest situation, where the underling trend in the data is linear and constant throughout the time period being considered. For the fixed linear case, a large daily dataset was simulated using models of the form described in Equation 2.3;

$$\begin{aligned} y_t &= \mu + \beta x_t + \gamma \cos\left(\frac{2\pi \text{doy}_\text{t} - \phi}{365}\right) + \varepsilon_t \\ y_t &= \mu + \beta x_t + \gamma_a \sin\left(\frac{2\pi \text{doy}_\text{t}}{365}\right) + \gamma_b \cos\left(\frac{2\pi \text{doy}_\text{t}}{365}\right) + \varepsilon_t \\ &\text{where } \varepsilon_t = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n) \sim N(0, \sigma^2 V) \\ &\text{for } t = 2, ..., n \ , \ \varepsilon_t = \rho \varepsilon_{t-1} + Z_t \text{ and } \quad Z_t \sim N(0, 1) \end{aligned} \tag{2.3}$$

Here $t = 1, 2, ..., n$ are the daily observations, doy_t is the of day of year for observation t and x_t is the time corresponding to observation t in decimal year which has been calculated using the formula year $+ \left(\frac{\text{doy}-1}{365}\right)$. The values for μ, γ_a and γ_b were based on estimates from the models fitted to existing TP and OP data and remained fixed throughout each of the different simulations. While initial estimates for the trend, β, and the variability, σ, were also obtained from existing data their values were varied throughout simulations. First order auto-regressive models were used in order to allow for correlated observations. This enabled the number of occasions where a statistically significant trend was detected in the presence of correlation to be investigated. In order to generate independent observations, the same model and error structure was used with the covariance matrix, denoted by V, set to be I_n (the $n \times n$ identity matrix) and ρ set at 0.

Although initial estimates for the model parameters had been obtained from existing data, a range of other suitable parameters were chosen around these values. Table 2.2 contains a summary of each of the different conditions that were used in this section of the simulation study. Model conditions were changed both in terms of the underlying model from which the data were generated, and the sampling frames that were then applied to this dataset.

Data Conditions: Fixed Linear Trend	
Number of simulations	500
Model Conditions	
Form of trend	linear
Magnitude of trend (β)	-0.2, -0.1, -0.5, -0.025, 0
Variance (σ^2)	0.1, 0.5, 1, 1.5
Correlation (ρ)	0, 0.2, 0.4, 0.6
Sampling Conditions	
Sampling frequency	annual, 6 times per year, monthly, fortnightly, weekly
Length of time series	5, 10, 15, 20, 25, 30 years
Sample Dates	unequally spaced, generated from a relevant Un(0,a) (a will be determined by specified sampling frequency)

TABLE 2.2: Conditions for Fixed Linear Trend Simulation

Figure 2.12 shows a set of plots which correspond to the four magnitudes of trend considered in the fixed linear simulation. The red line in each figure represents the underlying model from which data are sampled.

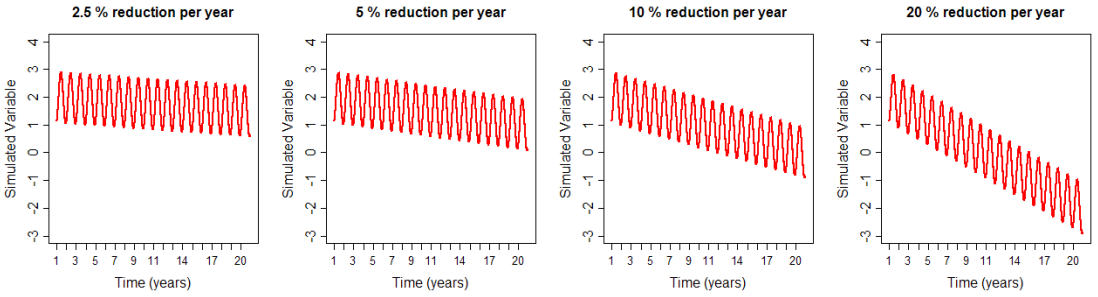


FIGURE 2.12: Examples of trends used in fixed linear simulation study

Assessing power to detect constant trend

After generating each set of sampling frames, they were subsequently applied to the daily dataset and an appropriate error term was added to each simulated data value. A linear model of the same structure as model described in Equation 2.3 was fitted to each subset of data that was sampled from the larger dataset. Following this, each of the models fitted to the sampled data were tested to see whether there were any statistically significant linear trends. This was carried out

by testing the null hypothesis,

$$\begin{aligned} H_0 : \beta &= 0 \qquad \text{against the alternative,} \\ H_1 : \beta &\neq 0 \end{aligned}$$

This test was carried out using a 5% significance level.

2.3.1 Scenario 1 - Results

How does the magnitude of linear trend affect power?

Figure 2.13 shows a set of results which describe how the power to detect statistically significant linear trends changes as the magnitude of the underlying effect changes. Each of the panels from (b)-(e) represents a different trend value and each line represents a different sampling frequency. Panel (a) shows the statistical size of the simulation: the probability of detecting a trend when there is in fact no trend present. A red dashed line has been included on panel (a) to indicate the 5% level. Panel (f) shows a key with the colours of lines used to represent each sampling frequency. The same colour key will be used throughout the remainder of this Chapter to indicate the different sampling methods. The underlying simulated dataset for all the results presented in Figure 2.13 (a)-(e) have a fixed variance and the samples are uncorrelated.

From panel (a) it can be seen that the statistical is often larger than the expected 5% level. This is in particular the case for annual sampling and highlights that often when there are very few observations it is possible to detect a trend when in fact there is no underlying pattern present. For all non-zero values of trend it is clear that power increases as the length of time series increases, and as the sampling frequency is increased. While this result is as expected, there are some particular results corresponding to the lower sampling frequencies which are important to note. Annual sampling performs poorly in terms of the level of power achieved with all magnitudes of trend considered here. Even with the highest effect size, which corresponds to a 20% decline in the simulated determinand each year, between 15-20 years worth of samples are required before a power of 0.8 is reached. In addition to concerns highlighted for annual sampling, another feature worth noting is that for monthly sampling, which is the frequency currently used

by SEPA for monitoring standing waters under the WFD, around 20 years of samples are needed in order to detect a decline of 5% with a level of power which is greater than 0.8. A reasonably small trend such as this may be of importance in the context of contaminant monitoring.

It is clear from the results shown here that more than 5 years of data are required to reach a level of power greater than 0.8 in most of the situations considered. Even with weekly sampling, when the underlying trend is very small, 15 years worth of data are required to detect a trend with a reasonable level of power.

How does variability affect power to detect a fixed linear trend?

In order to examine the effect of variability on power to detect change, the coefficient of variation, denoted throughout this Chapter as CV, will be used. This can be defined as the ratio of the standard deviation to the mean and represents the extent of variability in the data in relation to mean of the population.

Figure 2.14 shows the effect of different levels of variability on power to detect a constant rate of change in the mean level of the simulated determinand. In the results presented the underlying data are uncorrelated, and the trend is the same for each value of the variability considered, corresponding to a 10% reduction in the simulated determinand each year. Each panel represents a different variability value, while each line represents a different sampling frequency as before. The relative levels of variability in the data can be expressed as a percentage of the sample mean using the coefficient of variation (CV). The values of the variability and the corresponding CV values are presented in Table 2.3.

Variability Value	**CV**
0.1	30%
0.5	70%
1	100%
1.5	120%

TABLE 2.3: Variation values and corresponding coefficients of variation (CV) used within fixed linear trend simulation

As before, the results are as would have been expected, and clearly show that the greater the level of variability, the lower the power to detect a fixed linear trend. For weekly sampling, neither the length of time series, nor the level

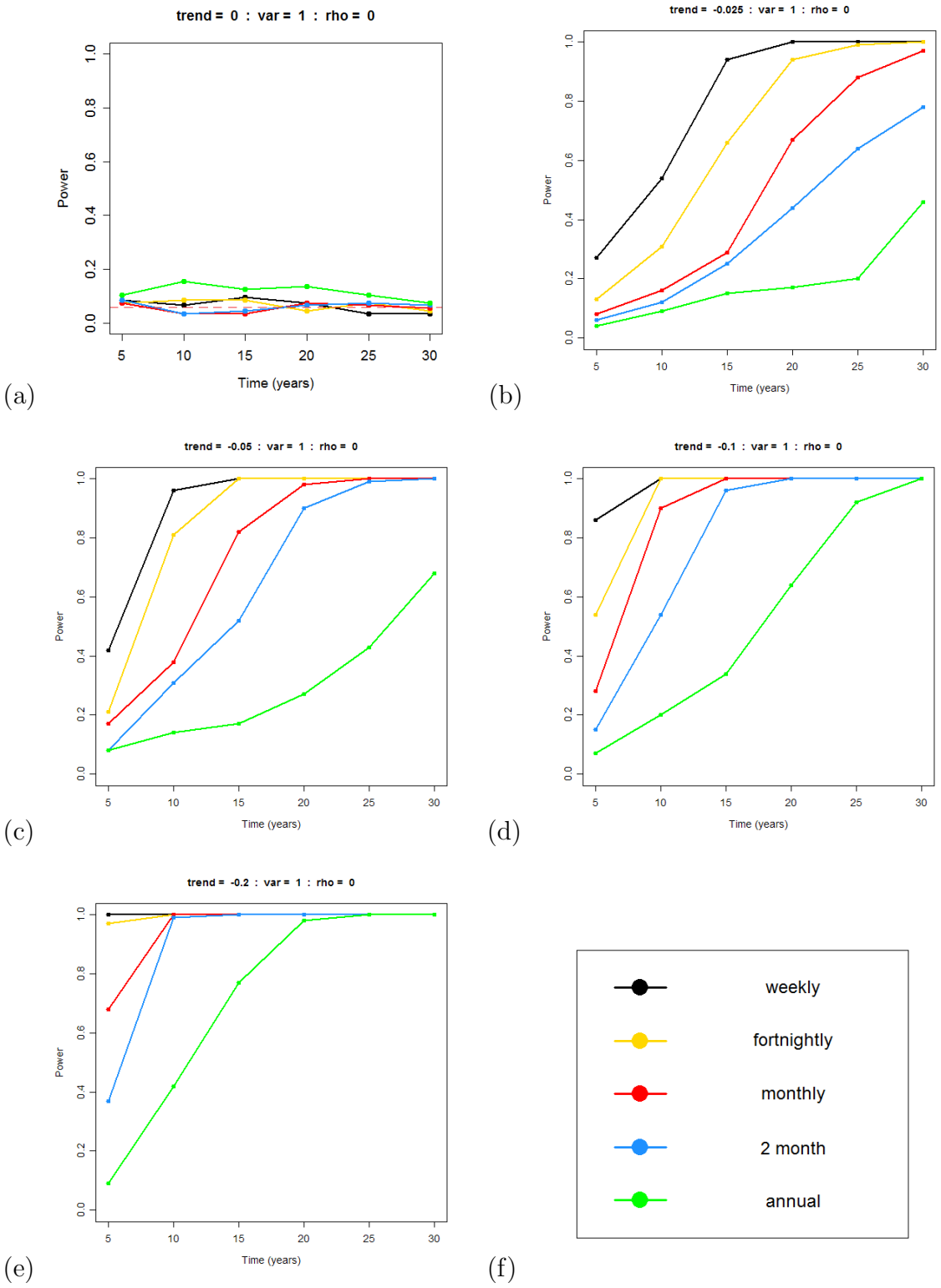


FIGURE 2.13: Simulation results showing how power and statistical size is affected by different magnitudes of linear trend

of variability has much effect on the power to detect a constant decline in the simulated determinand, and for all variation values the level of power achieved is greater than 0.8. The effects of variability on power in this situation are greater at the lower sampling frequencies. It can be seen from Figure 2.14 that when the CV is 100% ($\sigma^2 = 1$), around 25 years of annual samples are required before a power of 0.8 is exceeded. Further to this, a time series in excess of 10 years of monthly samples are required to reach an acceptable level of power, and for sampling 6 times per year more than 15 years worth of samples are required when the CV is 100%.

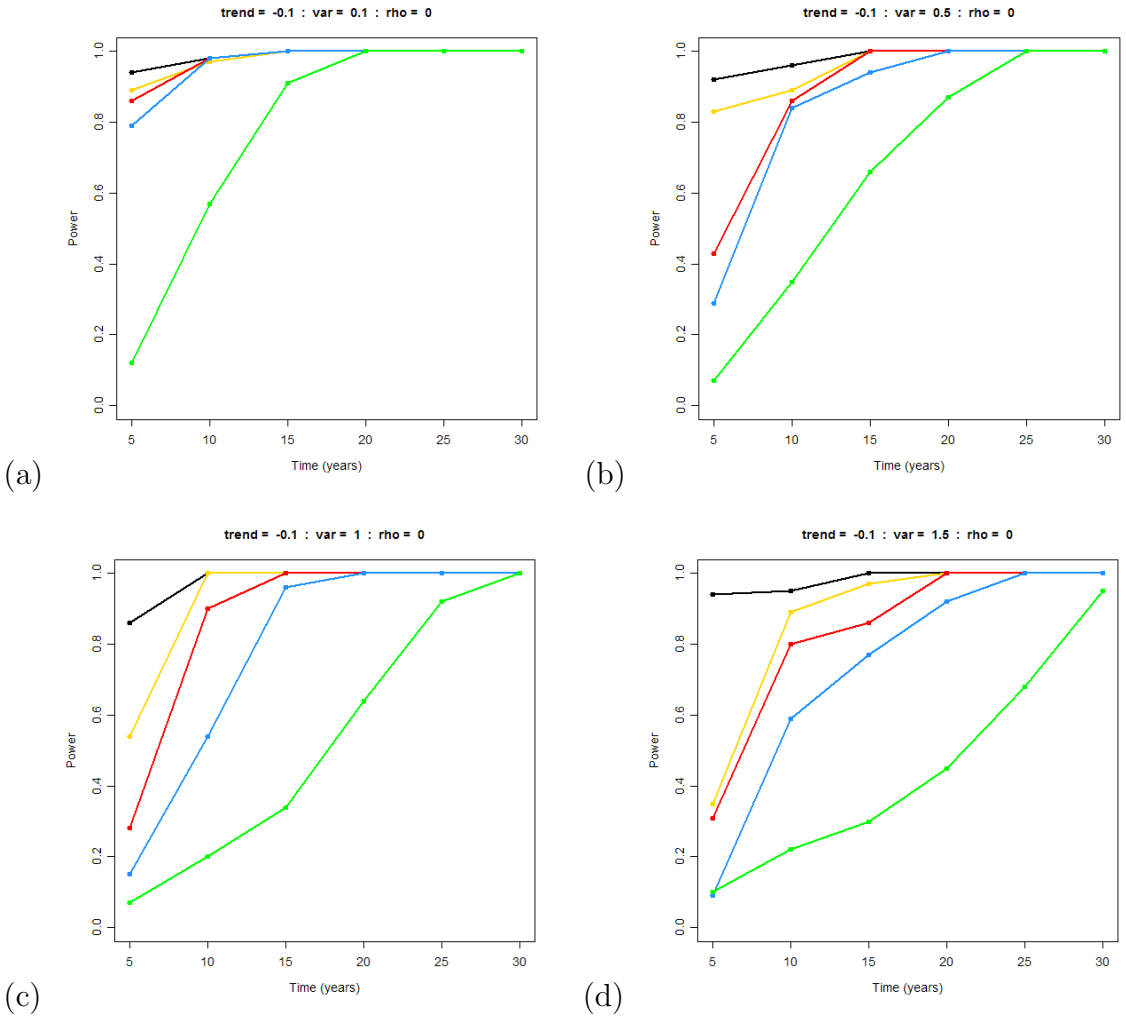


FIGURE 2.14: Simulation results showing how power to detect a fixed linear trend is affected by different levels of variability

trend (β), var (σ^2), rho (ρ)

How does correlation affect power to detect a fixed linear trend?

Figure 2.15 shows the effects of four correlation values on the power to detect a fixed linear trend which corresponds to a constant decline in the simulated determinand of 10 percent per year. Each panel represents a different level of correlation and each line represents a different sampling frequency. Variance values are fixed at 1 throughout all results presented in Figure 2.15. This value corresponds to a CV of 100%.

In general, it can be seen that as the strength of correlation increases, the power to detect a linear trend decreases. However autocorrelation appears to have a limited effect on power to detect a fixed linear trend when the sampling frequency is weekly. For monthly and fortnightly sampling, 10 or 15 years worth of data provide a level of power to detect a linear trend which is greater than 0.8 when there is a moderate level of correlation (ρ=0.4). If the correlation is particularly strong (ρ=0.6), more than 20 years of samples are required. This makes sense since the correlation present reduces the effective sample size, more data points are required to reach the equivalent level of power for correlated data than when the data are independent.

2.4 Scenario 2 - Non-Linear Trend

A similar simulation procedure to that carried out for the fixed linear trend was used in order to assess the effects a non-linear trend would have on our ability to detect a statistically significant change under different sampling conditions. Again, the effects of strength of correlation in the data and magnitude of trend will be explored. To simulate a dataset with a non-linear relationship between the simulated determinand and time, a piecewise linear model which consists of three distinct linear sections was used. Between sections, the size of the slope can be different, however within each section there is a fixed linear slope. While, as before, the slope coefficients, variance and lengths of the time series will be changed in different simulations scenarios, the ratio of the lengths of the three sections will be kept constant throughout in an attempt to preserve the general overall shape of the model. The underlying model from which the data are simulated will continue to include a constant seasonal component (fixed phase and amplitude

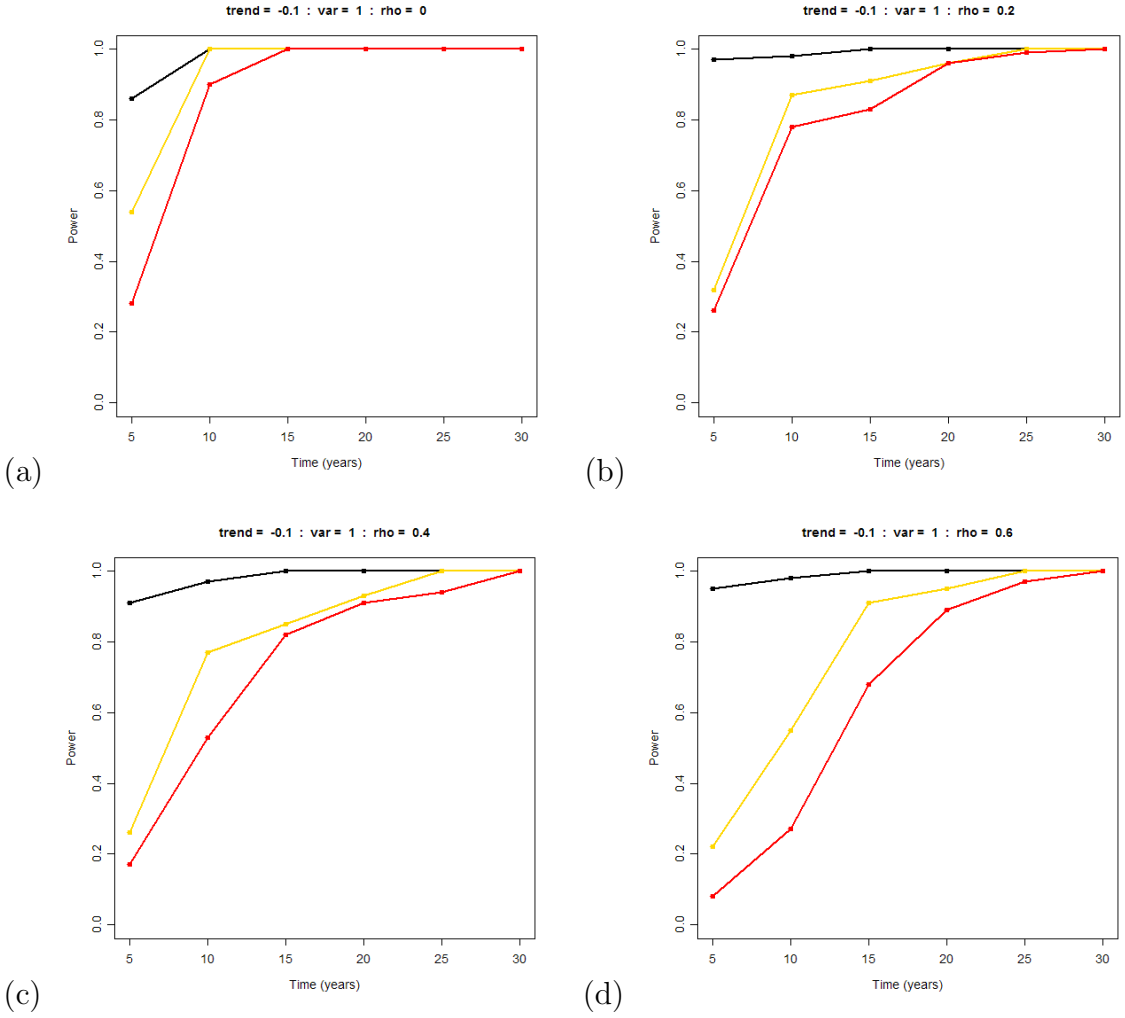


FIGURE 2.15: Simulation results showing how power to detect a fixed linear trend is affected by different strengths of autocorrelation

throughout) however, this feature of the data will not be taken into account when fitting nonparametric models to sampled subsets.

The model used to simulate the non-linear data is described below;

$$y_{1t} = \mu_1 + \beta_1 x_t + \gamma_a \sin\left(\frac{2\pi(\text{doy}_\text{t} - \theta)}{365}\right) + \gamma_b \cos\left(\frac{2\pi(\text{doy}_\text{t} - \theta)}{365}\right) + \varepsilon_1 \quad (2.4)$$
$$\text{where } 0 \leq t \leq t_i \text{ and } \varepsilon_1 = (\varepsilon_{t_1}, \ldots, \varepsilon_{t_i})$$

$$y_{2t} = \mu_2 + \beta_2 x_t + \gamma_a \sin\left(\frac{2\pi(\text{doy}_\text{t} - \theta)}{365}\right) + \gamma_b \cos\left(\frac{2\pi(\text{doy}_\text{t} - \theta)}{365}\right) + \varepsilon_2 \quad (2.5)$$
$$\text{where } t_{i+1} \leq t \leq t_j \text{ and } \varepsilon_2 = (\varepsilon_{t_i+1}, \ldots, \varepsilon_{t_j})$$

$$y_{3t} = \mu_3 + \beta_3 x_t + \gamma_a \sin\left(\frac{2\pi(\text{doy}_\text{t} - \theta)}{365} + \varepsilon_{3t}\right) \gamma_b \cos\left(\frac{2\pi(\text{doy}_\text{t} - \theta)}{365}\right) + \varepsilon_3 \quad (2.6)$$
$$\text{where } t_{j+1} \leq t \leq t_n \text{ and } \varepsilon_3 = (\varepsilon_{t_j+1}, \ldots, \varepsilon_{t_n})$$

$$\text{where } 0 < i < j < n, i = \frac{3n}{10} \text{ and } j = \frac{8n}{10}.$$
$$\varepsilon_m \sim N(0, \sigma_m^2 V) \text{ for } m = 1, 2, 3 \ ,$$
$$\varepsilon_m = \rho\varepsilon_{m-1} + Z_m \text{ and } \quad Z_m \sim N(0, 1)$$

Equations 2.4, 2.5 and 2.6 represent the equations of the linear segments. The different intercept values for each line enables the line segments to be connected. Here n is the number of years and t is the index of the daily observations, doy_t is the of day of year for observation t and x_t is the time corresponding to observation t in decimal year which has been calculated using the formula year $+\frac{(\text{doy}-1)}{365}$. For each of the simulations considered for the non-linear scenario $i = \frac{3n}{10}$ and $j = \frac{8n}{10}$.

Figure 2.16 is an example of a 20-year simulated monthly dataset consisting of three linear sections with a red line representing the underlying trend in each of the sections. The underlying nonlinear pattern is strong, the variance is fixed and the observations are uncorrelated. As well as the ratio of the section lengths remaining constant across the time periods considered, the direction of trend, variance and correlation remain the same in each of the models considered. The green vertical lines on this plot indicate the end points of the linear sections which there is a change in the trend. The first section has a negative linear trend, the second section covers a longer period of time and has a smaller negative trend and the third, shortest section has a positive trend. The reason for these choices is that this represents a fairly typical pattern in environmental determinands such as those of interest; where there has been a notable decline in the determinand in early

years of monitoring, a smaller decline as time goes on, perhaps due to the effects of measures of improvement which have been put in place, and finally a small increase in recent years. This pattern is similar to the non-parametric trend in OP data at Lake of Menteith which is shown in Figure 2.7(b). Table 2.4 below contains the

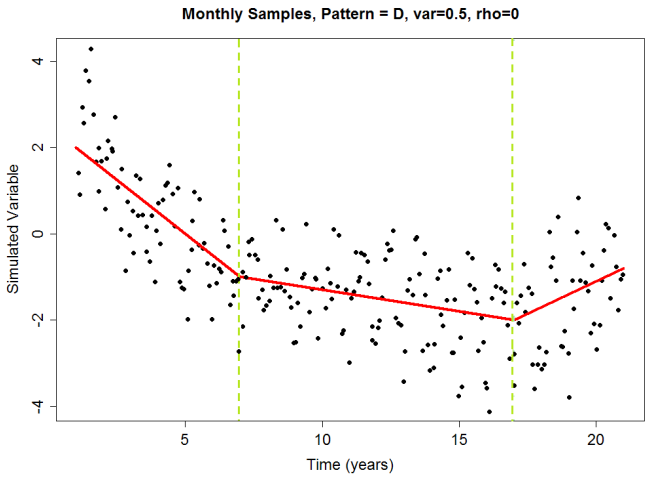


FIGURE 2.16: Example plot of simulated non-linear data (trend shown in red) 20 years of monthly observations.

different model and sampling conditions which were investigated within the non-linear trend simulations. The trend model conditions are described in sets of three, corresponding to the three linear sections which comprise the piecewise linear model used to simulate the data. As with the fixed linear simulation, a maximum of thirty years worth of daily values are generated. For each combination of model conditions, 6 datasets are simulated corresponding to each of the 6 lengths of time series considered. Following from this, from each of these datasets, 500 subsets are sampled at different sampling frequencies and a model was fitted to each subset.

The three sets of trend values correspond to three approximate 'strengths' of non-linearity in the data. The set of trend values labelled D. in Table 2.4 is designed to represent a 'strong' level of non-linearity in the data, while C. corresponds to a 'moderate' level, and B. corresponds to a 'weak' level. Set A. corresponds to no pattern in the data, and is designed to estimate the statistical size of the data, in terms of the ability of different sampling schemes to detect a non-linear pattern over time. An example of the underlying models for each of these non-linear trends is shown in 2.17. The red lines here represent the underlying model from which data are sampled.

Data Conditions: Non-linear Trend	
Number of simulations	500
Model Conditions	
Form of trend	non-linear piece-wise linear model (3 linear sections)
Ratio of Section Length	3:5:2
Magnitude of trend	A.(0,0,0), B.(-0.2, -0.05, 0.1), C.(-0.3, -0.1, 0.1), D.(-0.5, -0.1, 0.3)
Variance (σ^2)	1.5, 1, 0.5, 0.1
Correlation (ρ)	0, 0.2, 0.4, 0.6
Sampling Conditions	
Sampling frequency	annual, 6 times per year, monthly, fortnightly, weekly
Length of time series	5, 10, 15, 20, 25, 30 years
Sample Dates	unequally spaced, generated from a relevant Un(0,a) (a will be determined by specified sampling frequency)

TABLE 2.4: Conditions for Non-linear Trend Simulation

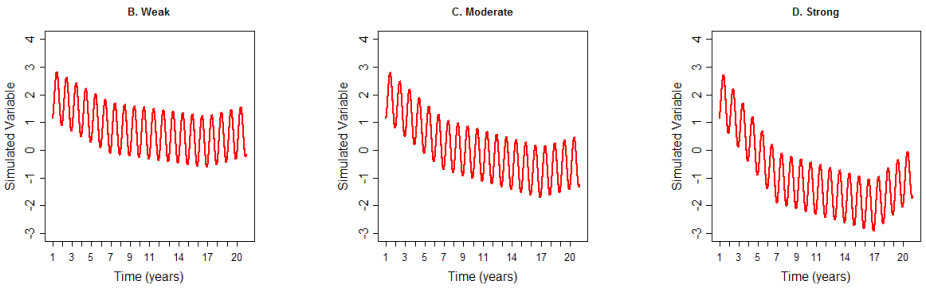


FIGURE 2.17: Examples of trends used in non linear simulation study

Assessing Power to Detect a Non-Linear Trend

After the data have been simulated and sampled, a nonparametric model is fitted to each subset using nonparmetric regression using the sm.regression function in R. To ensure the local linear regression model fit for each dataset is both appropriate and comparable AICc was used to automatically select the most suitable smoothing parameter value given the data. In the non-linear simulation the following hypotheses are tested;

$$H_0 : E\{y_t\} = \mu, \quad \text{constant mean} \tag{2.7}$$

$$H_1 : E\{y_t\} = g(x_t), \quad \text{some non-parametric relationship} \tag{2.8}$$

In order to obtain the power of detecting an underlying pattern in the data which was generated from the piecewise linear model, the proportion of subsets where the non-parametric model returned a significant p-value (corresponding to a rejection of the null hypothesis) is calculated. In addition to this first hypothesis test, the suitability of a linear relationship between the simulated determinand and time will also be tested. To assess the suitability of a linear model the following hypotheses are used

$$H_0 : E\{y_t\} = \mu + \beta x_t, \text{ linear model} \tag{2.9}$$

$$H_1 : E\{y_t\} = g(x_t), \text{ some non-parametric relationship} \tag{2.10}$$

The proportion of sampled subsets where both the null hypotheses (Equation 2.7 and Equation 2.9) are rejected was calculated. Approximate F-tests were used to compare models. From this, the probability that a non-linear model is a suitable fit to the data, given the data have been generated from an underlying model which was non-linear could be assessed.

2.4.1 Scenario 2 - Results

How does strength of 'non-linearity' affect power to detect a non-linear trend?

Figure 2.18 shows a set of results which demonstrate how power to detect statistically significant non-linear patterns alters as the underlying effect size changes. Each of the panels from (b)-(d) represents a different non-linear pattern and each line represents a different sampling frequency. Panel (a) shows the statistical size of the simulation; the probability of detecting a non-linear when there is in fact no pattern present. The underlying simulated dataset for all the results presented in Figure 2.18 had a fixed variance and independent samples.

It can be seen from 2.18 (a) that the probability of detecting a significant non-linear pattern in the data is less than 5% for all sampling frequencies other than annual sampling, when the probability is slightly higher for all lengths of time series considered. Annual sampling also performs poorly in terms of detecting a

non-linear pattern in the underlying data when the curvature in the pattern is weak or moderate, and, even when the non-linear pattern is strong, around 15 years of data are required before a power of around 0.8 is reached. With time series in excess of 10 years weekly, fortnightly and monthly sampling schemes are comparable and perform well in terms of their ability to detect a non-linear pattern.

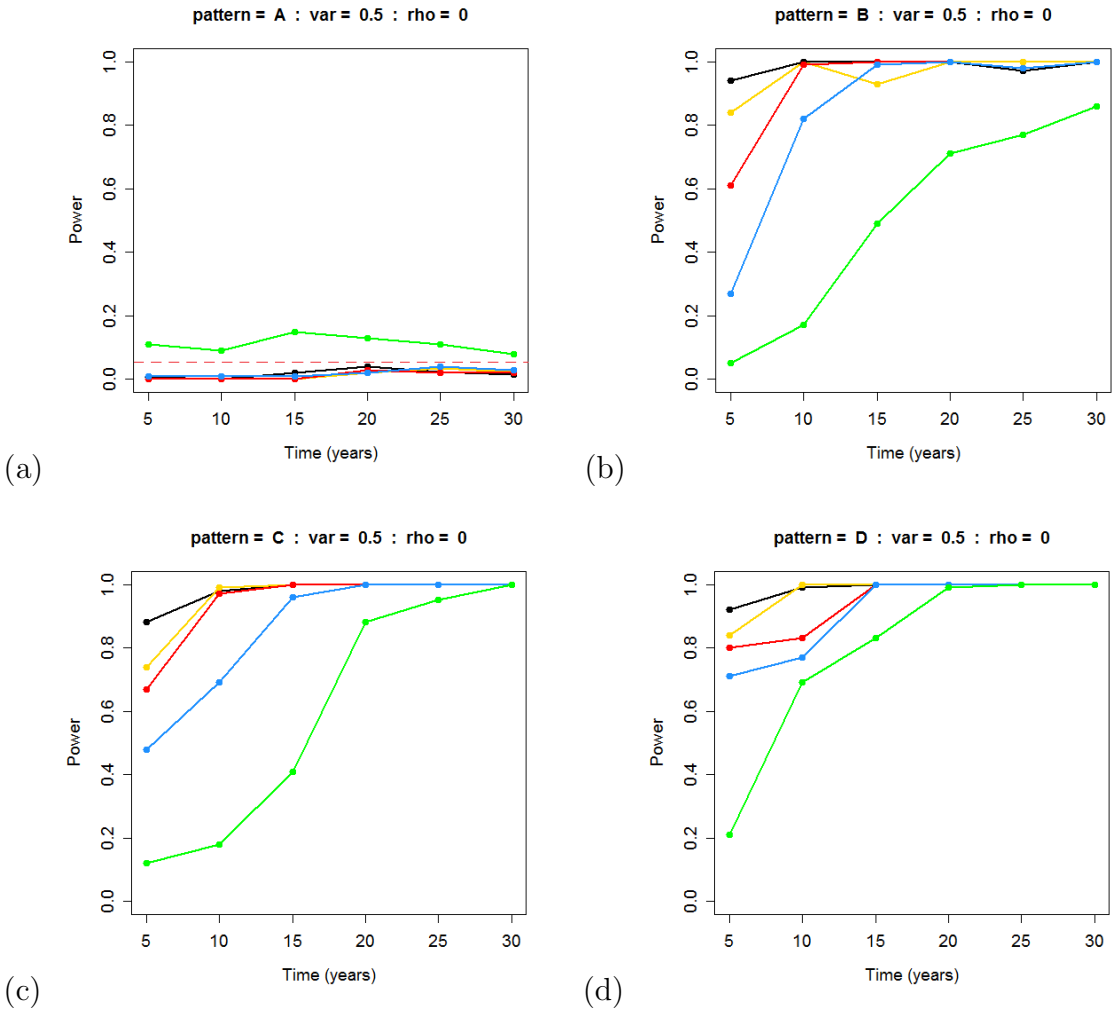


FIGURE 2.18: Simulation results showing how power and statistical size is affected by different magnitudes of non-linear pattern

How does variability affect power to detect a non-linear trend?

Figure 2.19 shows the effect of different levels of variability on power to detect a non-linear change in the simulated determinand. In the results presented the underlying data is uncorrelated, and the non-linear trend is the same for each

value of the variability considered. Each panel represents a different variability value, while each line represents a different sampling frequency. The four variability values investigated are 0.1, 0.5, 1 and 1.5, which in this case correspond approximately to coefficients of variation of 30%, 70%, 95% and 115%.

As with the fixed linear results it is clear that as the coefficient of variation increases the ability to detect a change decreases. When the coefficient of variation is less than 100%, monthly, fortnightly and weekly sampling all perform well in terms of power to detect a moderate non-linear pattern in the data and reach an acceptable level of power with time series of around 10 years or longer. With the same level of variance, annual sampling and samples collected at a frequency of 6 times per year require around 20 years or longer before a power around 0.8 or greater is achieved.

How does correlation affect power to detect a non-linear trend?

Figure 2.20 shows the effects of different levels of correlation in the sampled data on the power to detect a non-linear trend. Each panel represents a different level of correlation and each line represents a different sampling frequency. Variance values are fixed at 0.5 throughout, corresponding to a coefficient of variation of around 70%, and a moderate strength of non-linear pattern is considered.

In general, these results are similar to the fixed linear case; correlation in the data has little effect on weekly sampling in terms of the power to detect a non-linear pattern, but a more notable effect on monthly and fortnightly sampling. When the correlation coefficient is 0.4, about 15 years of monthly samples would be required to detect a moderate non-linear pattern with a power close 0.8, whilst with a higher correlation coefficient around 25 years of data would be required.

2.5 Scenario 3 - Varying Seasonal Component

For the third simulation scenario the ability to detect a change in the seasonal pattern under different sampling conditions was considered. Again, the effects of variability, different sizes of change and the strength of autocorrelation on the ability to detect the underlying change was investigated. Model 2.11 below was used to generate data with a varying amplitude seasonal component. The trend,

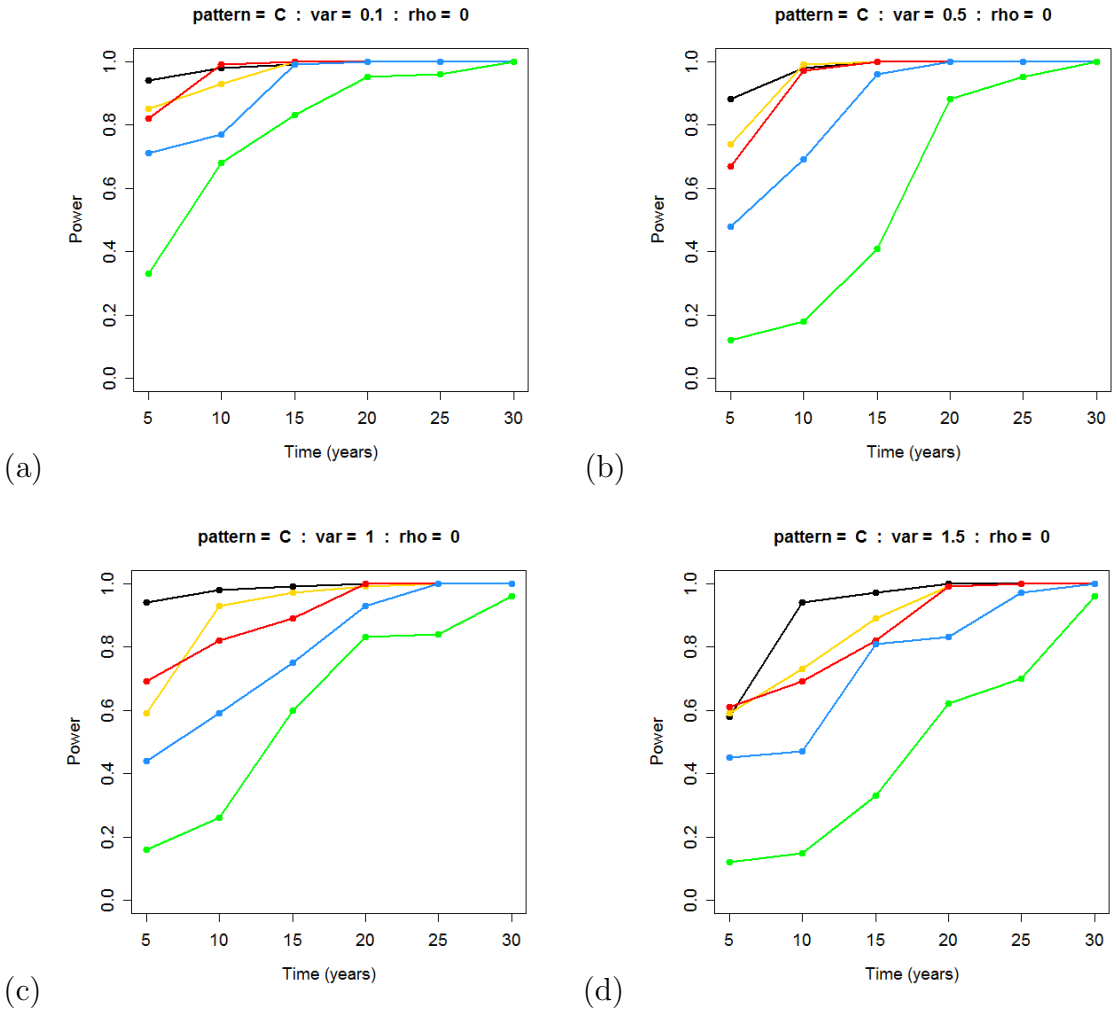


FIGURE 2.19: Simulation results showing how power to detect a non-linear trend is affected by different levels of variability

variance and the phase of the seasonal component remained constant over the time period for which data were generated.

$$y_t = \alpha + \beta x_t + \gamma_a(r - x_t)\sin\left(\frac{2\pi \text{doy}_\text{t}}{365}\right) + \gamma_b(r - x_t)\cos\left(\frac{2\pi \text{doy}_\text{t}}{365}\right) + \varepsilon_t \quad (2.11)$$

$$\text{where } \varepsilon_t = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n) \sim N(0, \sigma^2 V)$$

$$\text{for } t = 2, ..., n \ , \ \varepsilon_t = \rho\varepsilon_{t-1} + Z_t \text{ and } \quad Z_t \sim N(0, 1)$$

In this model the amplitude of the seasonal pattern reduces by a specified percentage over the time period considered; r is a constant which reverses decimal

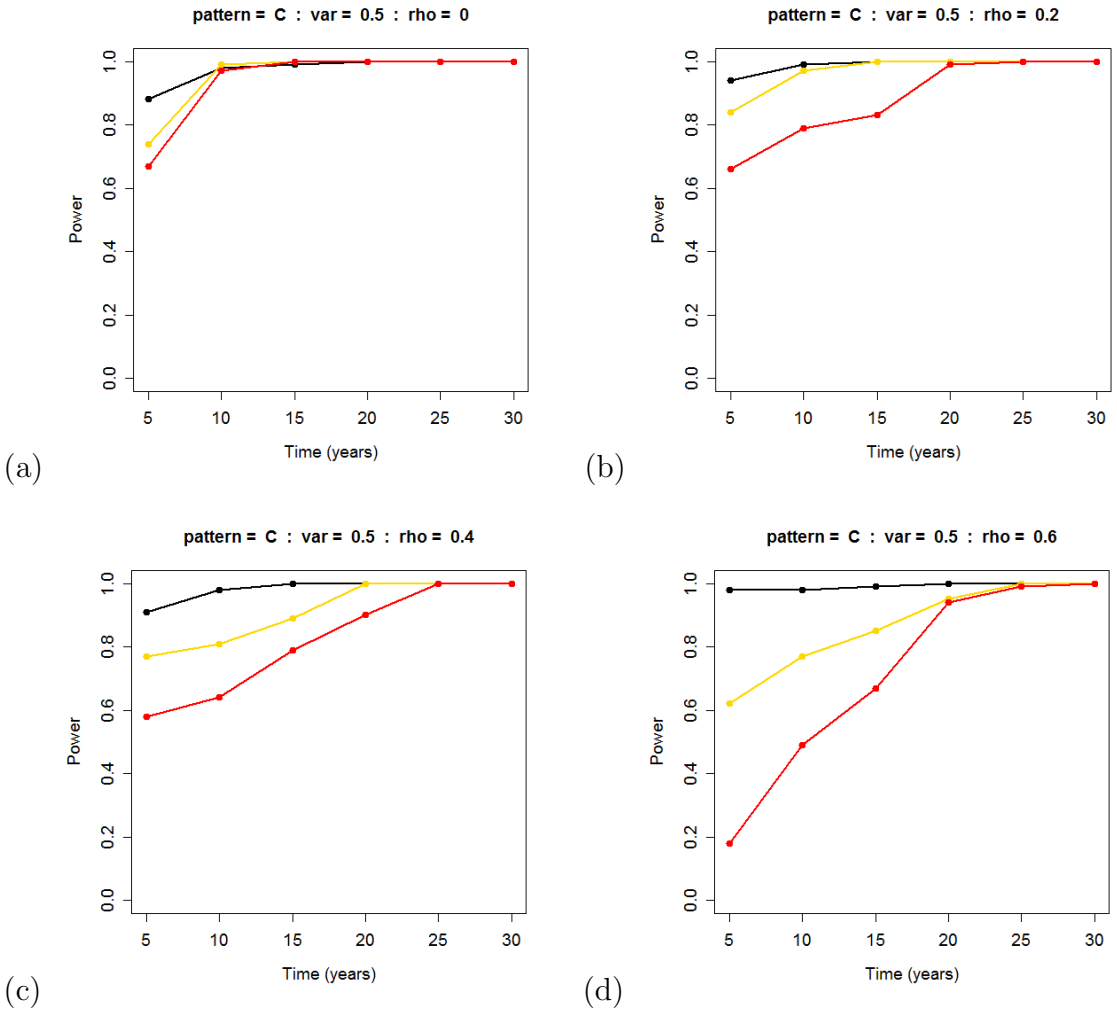


FIGURE 2.20: Simulation results showing how power to detect a non-linear trend is affected by different strengths of autocorrelation

year. The implication of this reduction in amplitude is that samples recorded in the summer and winter months become increasingly similar in terms of their observed value as time progresses. Table 2.5 contains a summary of the different model and sampling conditions that were considered in the varying amplitude simulation scenario. The same values of the slope, variability and correlation as in the fixed linear scenario were investigated.

An example of the underlying models for each of these reductions in amplitude over a 20 year period is shown in 2.21. The red lines here represent the underlying model from which data are sampled. While it could be argued that a 20% reduction is not particularly noticeable and will inevitably be difficult to detect, it is important to consider changes of this magnitude since even small changes in the

Data Conditions: Changing Seasonal Pattern	
Number of simulations	500
Model Conditions	
Form of trend	fixed linear
Seasonal Pattern	varying amplitude of seasonal term
	0%, 20%, 40% and 60% reductions
Magnitude of trend	-0.1,
Variance	0.1, 0.5, 1, 1.5
Correlation	0, 0.2, 0.4, 0.6
Sampling Conditions	
Sampling frequency	monthly, fortnightly, weekly
Length of time series	5, 10, 15, 20, 25, 30 years
Sample Dates	unequally spaced, generated from a relevant Un(0,a)
	(a will be determined by specified sampling frequency)

TABLE 2.5: Conditions for Varying Amplitude Simulation

% indicates the reduction in the amplitude over time

seasonal patterns can have large resulting effects on water ecosystems (Carvalho and Kirika, 2003).

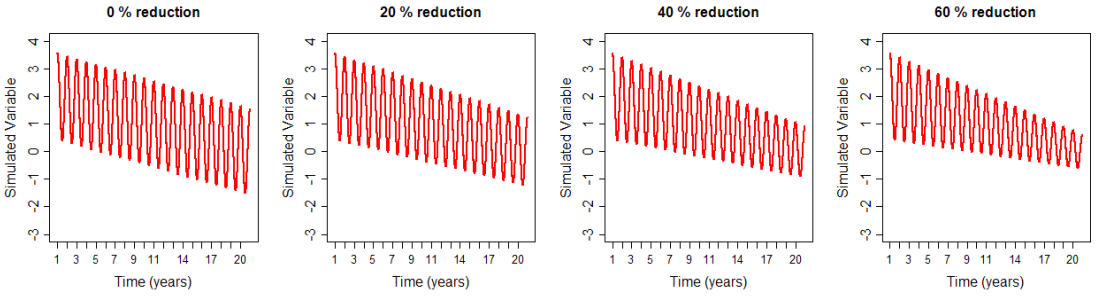


FIGURE 2.21: Examples of trends used in varying amplitude simulation study

Assessing power to detect a changing seasonal pattern

Two different models were fitted to each of the 500 simulated datasets for each set of conditions; an additive model of the form;

$$y = \mu + g_1(\text{year}) + g_2(\text{dayofyear}) + \epsilon, \epsilon \sim N(0, \sigma^2) \tag{2.12}$$

and a bivariate model of the form;

$$y = g(\text{year}, \text{dayofyear}) + \epsilon, \epsilon \sim N(0, \sigma^2). \tag{2.13}$$

For both models the two explanatory variables of interest used within the models are year (in the form of decimal year as described previously) and the day of year. Including these as covariates enables the trend over time as well as the seasonal pattern within each year to be examined. The additive model assumes that there is a constant seasonal pattern over time, or in other words that the pattern within a single year is the same across all years considered. Using the additive model, an estimate of the trend over time and an estimate of the seasonal pattern within each year are obtained separately. Both these estimates are non-parametric smooth functions which are found using local linear regression.

In contrast to the additive model, the bivariate model is more flexible and allows the estimate of the seasonal pattern within a single year to be different across the years. Given it is known that the underlying model from which the data have been sampled does in fact have a different seasonal pattern in each year, this could be said to be the 'correct' model. However, the question of interest here is how often the change in the seasonal pattern can be detected under different sampling and model conditions and so it is of interest to assess how often the bivariate model would be chosen over the additive model. In order to compare these two models an approximate F-test is carried out. To ensure that fair comparisons can be made between the pairs of models the smoothing parameters for each model have to be equivalent. After a sensitivity analysis was conducted using several simulated datasets it was decided that smoothing parameters should be selected so that in each model, 10% of the total number of observations contributed to the kernel estimate at each target observation of interest.

Ideas as to what magnitudes of change in amplitude should be considered in this simulation study were obtained by assessing changes in the amplitude of the seasonal pattern in observed datasets. There was no statistically significant evidence in a change in seasonal pattern in the Linlithgow Loch or Lake of Menteith data considered earlier and so an additional dataset from stations on the River Tweed in the South of Scotland were used to estimate realistic changes in the amplitude of a seasonal pattern for water quality data. A subset of a longer time series was taken and the ratio of the amplitudes of the seasonal pattern at the start and end of the time period was computed. A series of different subsets were

taken and the amplitude ratio was calculated for each subset. These ratios were then used to inform the magnitudes of change in amplitude used in the simulation study. While this may be viewed as a rather simplistic approach as it assumed a constant decrease across the time period, in many subsets the rate of change in amplitude appeared reasonably steady.

2.5.1 Scenario 3 - Results

The results of our simulation study highlight a difficulty in the detection of relatively small changes in the amplitude of a seasonal pattern over time. Figure 2.22 shows the power of detecting three different magnitudes of change in the seasonal pattern over time under different sampling frequencies. The power of detecting a changing seasonal component when it is not there, equivalent to the statistical size for the varying seasonal scenario, is shown in Figure 2.22 (a). A red dashed line on panel (a) indicates the 5% level. The magnitudes of change shown in panels (b), (c), and (d) correspond to 20%, 40% and 60% reductions in the amplitude of the seasonal pattern over the time period considered. In all of the results shown in 2.22 the underlying data are independent, have a constant variance corresponding to a coefficient of variation of 70%, and have a fixed linear trend corresponding to a 10% reduction in the mean level of the simulated determinand each year.

As could be expected, from panel Figure 2.22 (a) it is clear that for all sampling frequencies considered the statistical size is close to the 5% level. It can be seen from panel (b) that under all sampling frequencies, the power of detecting a 20% reduction in the amplitude of the seasonal component is extremely small for all lengths of time series considered. Even with 30 years of weekly sampling, it is unlikely that this magnitude of reduction in the seasonal component will be detected. With a change in amplitude of 40%, 20 years of weekly samples are required before a power of 0.8 is exceeded, while with fortnightly sampling it is only with a 30 year time series that the level of power to detect the change is adequate. There is some improvement in the power of detecting an underlying change in the seasonal component when the reduction in amplitude is large. For a 60% reduction, weekly and fortnightly sampling reach a level of power of around 0.8 or greater with time series lengths in excess of 5 years. Figure 2.22 (d) demonstrates that even with a 60% reduction in amplitude, monthly sampling fails to reach an acceptable level of power with time series shorter than around 25 years.

The inability of monthly sampling to detect large reductions in the seasonal pattern, even when the time series in question is lengthly, could be of concern in the context of environmental monitoring. Changes in the seasonal pattern over time in water quality determinands are a likely consequence of long-term changes in climate (Winder and Schindler, 2004) and this simulation study demonstrates that there is a real risk that underlying declines in the amplitude of a seasonal signal could potentially be missed by the type of monitoring programmes that are currently in place. A change in the seasonal pattern could be misinterpreted as non-constant variance. If this was the case, it is likely that the transformations typically employed to deal with instability in the variance over time, such as the log transform, would be inadequate.

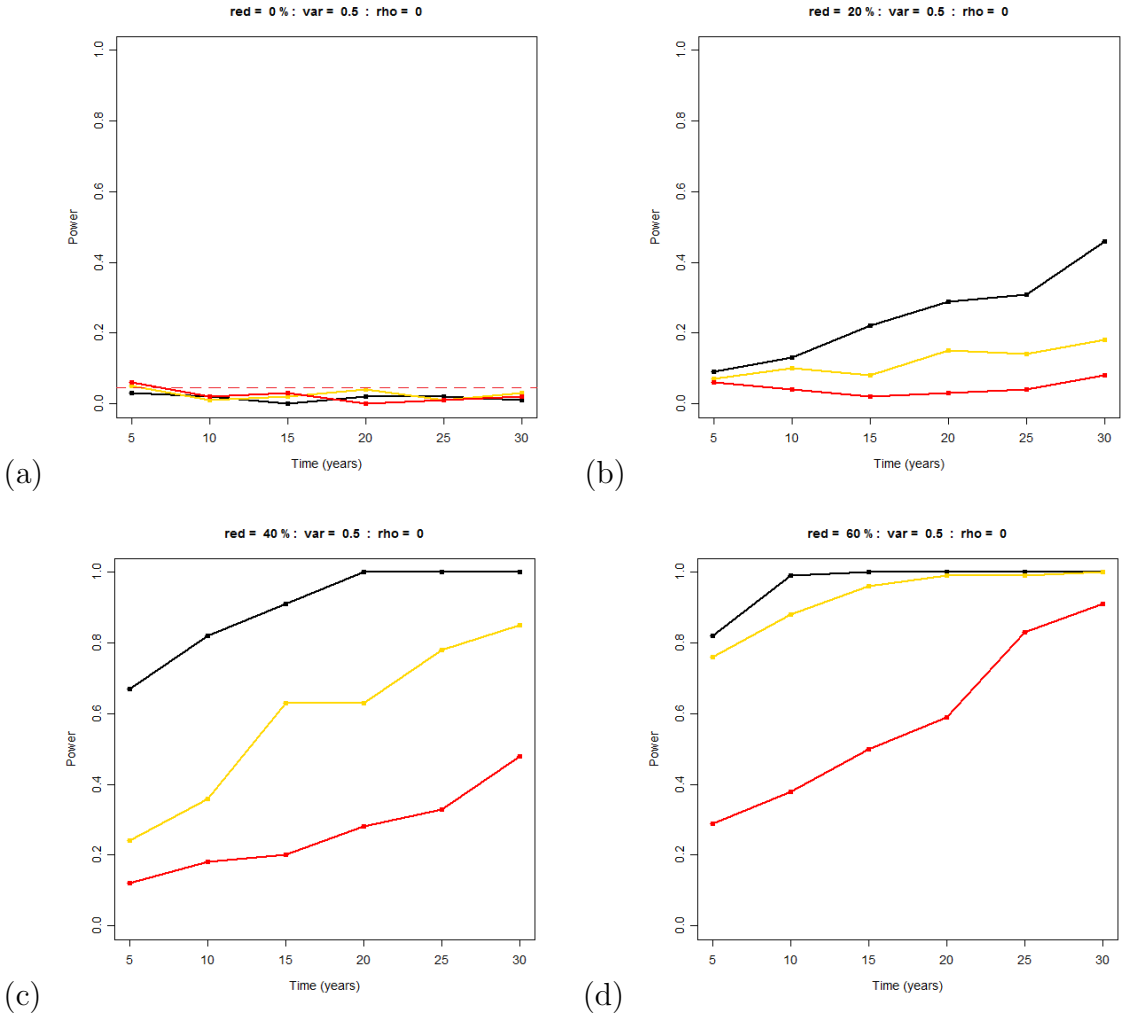


FIGURE 2.22: Simulation results showing power to detect different magnitudes of changing seasonal signals

The effects of variability on the power to detect changing seasonal patterns

The next step was to assess the effects of different levels of variability in the seasonal pattern. The variance values used in the varying amplitude scenario were again 0.1, 0.5, 1 and 1.5 which correspond to coefficients of variation of approximately 30%, 70%, 100% and 120% respectively. It is worth noting that the coefficient of variation in the underlying data for the results presented in Figure 2.22 is one of the smaller values considered within this study. The effects of variation on our ability to correctly identify a change in the seasonal component over time are presented in Figure 2.23. Each of the panels in this figure represents a different level of variability in the underlying data, while each line again represents a different sampling frequency. The magnitude of the linear trend and the magnitude of the reduction in the seasonal component amplitude is 60% in the four sets of results presented.

It is clear that the more variable the data, the weaker the power of detecting a changing seasonal pattern when it is, in fact, present. The magnitude of reduction considered here is 60% over the time period. While this is a substantial reduction, it is only for the smallest variability value in our simulation that monthly sampling reaches an acceptable level of power. In this case when the coefficient of variation is around 30%, after approximately 5 years all sampling frequencies investigated are comparable and perform well in terms of their power to detect this form and magnitude of change. When the CV is 70% or greater, monthly sampling does not reach a level of power in excess of 0.8 at any length of time series considered.

Fortnightly sampling also performs poorly in terms of power to detect this change in the amplitude of the seasonal pattern when the variability is relatively large, and it can be seen from 2.23 (d) that at the largest value of variation in the data investigated, corresponding to a CV of around 120%, even at the weekly sampling frequency, a time series in excess of 20 years is needed to detect this change in the seasonal signal.

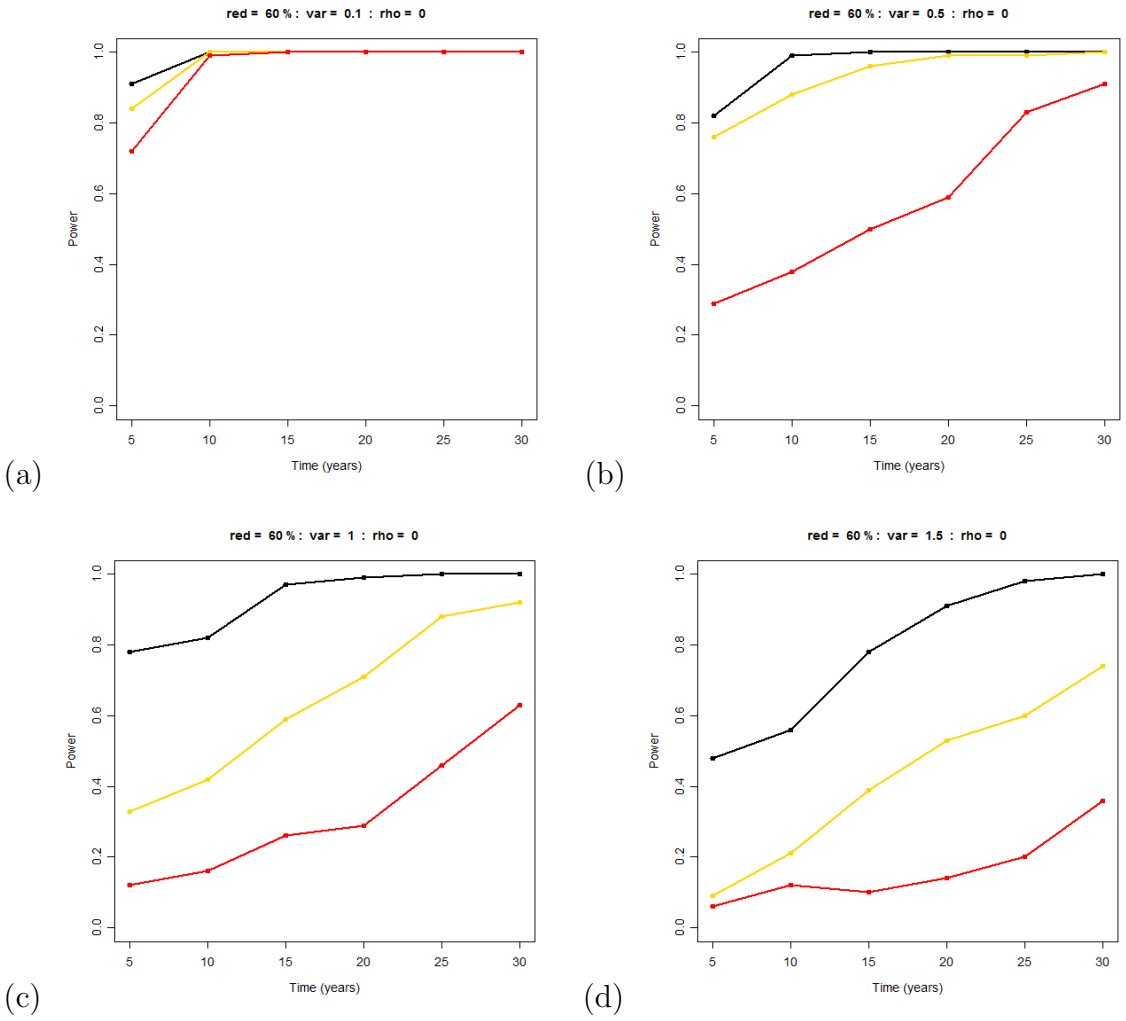


FIGURE 2.23: Simulation results showing how power to detect a changing seasonal signal is affected by different levels of variability

The effects of correlation on power to detect changing seasonal patterns

The effects of correlation on the ability to detect a changing seasonal component are summarised in the results shown in Figure 2.24. Each of the panels corresponds to a underlying dataset with a different strength of correlation. All results presented are based on data with the same strength of underlying linear trend, and the same constant variability (CV=70%), and a 60% reduction in the amplitude of the seasonal signal. Figure 2.24 indicates that correlation has a notable effect on the ability to detect a change in the amplitude of the seasonal pattern over time. The power of weekly sampling to detect a 60% reduction in the amplitude of the seasonal component exceeds 0.8 even after 5 years of samples which are independent. However, when there is a level of correlation in the samples which

corresponds to an AR(1) process with a correlation coefficient of 0.4, around 15 years of samples are needed. Further to this, when the correlation coefficient is 0.6, around 25 years of samples are required.

Monthly sampling and fortnightly sampling also see a decrease in the power to detect this form of change in the presence of correlated data. It can be seen from Figure 2.24 (b) that even when the level of autocorrelation in fortnightly samples is relatively low (ρ=0.2) around 20 years of samples are required before the power to detect a large change in the seasonal component is adequate. When the level of correlation is greater than this fortnightly sampling fails to reach a level of power in excess of 0.8.

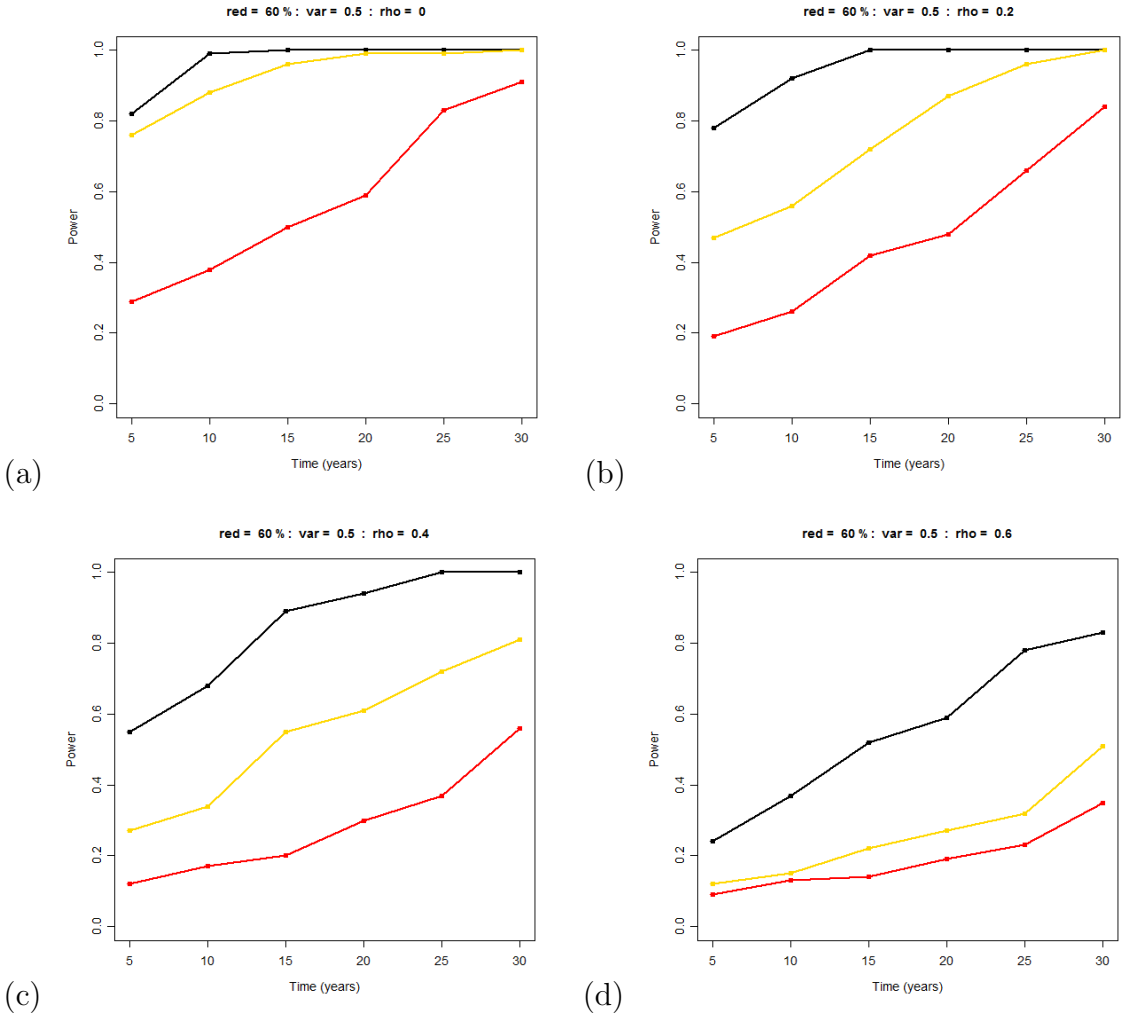


FIGURE 2.24: Simulation results showing how power to detect changing seasonal signal is affected by different levels of correlation

2.6 Summary

The simulation patterns chosen within these scenarios are designed to be indicative of the types of underlying patterns which may be of interest when investigating changes in water quality. The rationale behind this simulation study was to provide an insight into the likely effectiveness of the monitoring programmes which are commonly used at present, and to act as an uncomplicated, yet useful, guide to the relative power associated with different patterns and forms of underlying change. It is possible that any and all of these changes are happening in water quality determinands and so it is necessary to quantify how reliable the results of currently used sampling programmes are.

For all of the simulation scenarios, it can be seen that the power to detect change increases as the length of time period covered by the data increases, and as the sampling frequency increases. It is also apparent that high variability and high correlation have a negative effect on power to detect change, while the larger the magnitude of change, such as strength of linear trend or the reduction in seasonal amplitude, the higher the power. All of these features are well established, however, some of the particular results from the simulation study may be useful when placed in the specific context of the environmental monitoring programmes used to classify water bodies under the WFD. Many of the examples considered in the literature are concerned with power to detect changes in species populations rather than contaminants. Often these examples investigate the ability to detect changes which are greater in magnitude than some of the changes considered in contaminant monitoring. The smallest trend in the simulation study in this Chapter corresponds to a decrease in the concentration of the determinand of 2.5% per year, while the largest is a corresponds to a decline of 20%. Keizer-Vlek et al. (2012) considers the power to detect declines of 20% and 40% in populations of macroinvertibrates, while Nagelkerke and van Densen (2007) considers power to detect 15% per year reductions in fish populations. The sampling programmes considered in these examples are at an annual level.

The fixed linear scenario in the simulation represented the most basic case and so it is unsurprising that the linear trends were amongst the easiest to detect for all of the changes investigated. In general, after around 15 or more years of data, weekly and fortnightly sampling work well in terms of detecting fixed linear changes, even when those changes are relatively small. This is supported by the

findings of Howden et al. (2011). If the underlying trend in the data is strong, monthly sampling also reaches an adequate level of power with around 15 years of data. Annual sampling performs poorly, and if this sampling frequency is used, time series in excess of around 20 years are needed for even the strongest trends considered. Although there are examples of power analysis study for environmental data in the literature, very few look at anything other than constant linear trends over time. In the simulation study presented here, non-linear and changing seasonal patterns have also been explored. In terms of the non-linear simulations, weekly and fortnightly sampling again perform well with all three strengths of non-linear pattern considered. Annual sampling was again noticeable in its inability to detect underlying change, even when the time series of data are long. It has also been shown that for all of the scenarios considered, when there is strong correlation present, or when there is a high level of variability, the power of these different sampling schemes diminishes further.

Although the methods to estimate power to detect change are often different from our approach, the results of our simulation are comparable in cases where the magnitude of changes considered are comparable. For example, Field et al. (2007) considers the objective of detecting a decline of 30% over 10 years, and subsequently states that with up to 5 years of annual monitoring, there would be little chance of detecting changes of interest and consequently little increased confidence in resultant management decisions. It is also noted by the authors here that stopping monitoring programmes after 5 years, before adequate levels of power have been reached, would unnecessarily waste all the monitoring effort and resources invested up until that point. Howden et al. (2011) found that hydrological variability can mask trends in water quality datasets and identifies 12 years as the approximate minimum period required to consistently detect the true underlying trend.

While some analysis of water quality data has indicated that seasonal patterns are changing over time (Morton and Henderson, 2008), the results of the varying amplitude simulation indicate that even when there is a large change in the amplitude of the seasonal signal, corresponding to a 60% reduction, that monthly sampling often does not have a sufficient level of power to detect this change. Another finding of this simulation scenario is the extent to which power to detect changing seasonal patterns is hampered by the presence of autocorrelation in the data. Autocorrelation appears to have more of a detrimental effect on the power

to detect a varying seasonal pattern than any of the other forms of change considered. As noted, the findings of this simulation scenario are of particular interest as often there is evidence of non-constant variance in contaminant concentration data, which could in reality be a change in the seasonal signal. The ability to detect a varying amplitude is also important as changes in the seasonal pattern of a single determinand can have consequences on an entire ecosystem. For example, Winder and Schindler (2004) considers how a shift in seasonal patterns of nutrients can result in changes in the synchrony between trophic levels, and hence can cause perturbations to food webs if interacting species respond differently to shifting environmental conditions.

The results of the simulations in this Chapter may suggest that resources would be best used to sample at a reduced number of locations on a more frequent basis in order to detect change with a suitable level of power. This consequently raises the question of where the most suitable locations to implement increased monitoring should be. This will be investigated in the following Chapter where the question of how sites should be grouped for monitoring purposes will be considered.

Chapter 3

Functional Data Analysis

Until now, the main area of interest for this thesis has been investigating how frequently individual standing waters need to be monitored in order to detect different types of long-term change. However, as noted in Chapter 1, one of the features of the WFD is that lakes can be grouped together and the classifications of all members of the group can then be based on the classification of a single representative lake, enabling water quality to be predicted without monitoring. While water bodies are classified on the basis of levels of a range of chemical and biological determinands, currently, the groups of standing waters used for WFD monitoring by SEPA are based on typology which is derived from broad categories of alkalinity, altitude and depth. Often, the representative lake within each group is determined by logistics and ease of access for sampling purposes. There is some concern with regards to how reliable the current groups are, particularly as wrongly specifying either the groups, or the representative lake within each group, could potentially result in misclassification of all members.

The potential for inaccurate classification of lakes means there are two main aims when considering lake groupings. The first aim is to investigate the current group structure used by SEPA and to assess how well the existing groups capture differences between the lakes in terms of several chemical determinands which are used for WFD classification. The second aim is to look at alternative group structures which are based on different-statistical approaches applied to observed chemistry data. Furthermore, it is also of interest to explore if fewer, or indeed, more groups would be optimal in order to accurately represent the trends and variability in the water quality determinands of interest.

3.1 Available Data

In total there are approximately 104 standing waters across Scotland that SEPA classify within groups for the WFD. These lakes make up 30 distinct groups, with the number of lakes in a single group ranging from two to eight. In addition to the grouped lakes, there are several lakes which are classified on an individual basis. As mentioned previously, it is of interest to investigate similarities and differences between the water bodies in terms of chemical determinands in order to ascertain how suitable the current groups are. There is, however, limited chemistry data available on many of the lakes and in addition, monitoring data are often available on different lakes at different times. From 2007 onwards, when classification based on groups of lakes was first introduced, there is commonly only data available on the representative lake. Ideally, in order to ensure reasonable comparisons, a dataset is required where there are observations taken over a common period on all lakes within each group. For this reason, data from a subset of lakes were provided by SEPA. The dataset used throughout Chapters 3 and 4 consists of seven groups made up of 24 lakes. The time period covered by the data is from January 1996, when sampling procedures in SEPA laboratories were granted United Kingdom Accreditation Service (UKAS) accreditation, to December 2009. Even within this subset the number of samples, and the dates on which samples were collected, varies enormously from lake to lake. Data were available on the 3 different determinands of interest: Alkalinity (as CaCO3), Phosphorus (as P) and Chlorophyll$_a$. All measurements for these determinands are in micrograms per litre ($\mu g/L$).

The geographical locations and current groups used by SEPA are indicated on Figure 3.1. The different colours correspond to the different groups which are currently used by SEPA. For each of the 24 lakes considered, the lake name, SEPA location code and current SEPA group number is shown in Table 3.1. The representative lake for each of the groups are also indicated (marked by X).

Figure 3.2 provides a graphical representation of the quantity of data available at each lake for each of the three determinands and the time period that is covered by the observations.

While initially the intention was to compare lakes, both between and within groups, using the samples collected at each lake matched by date, it can be seen

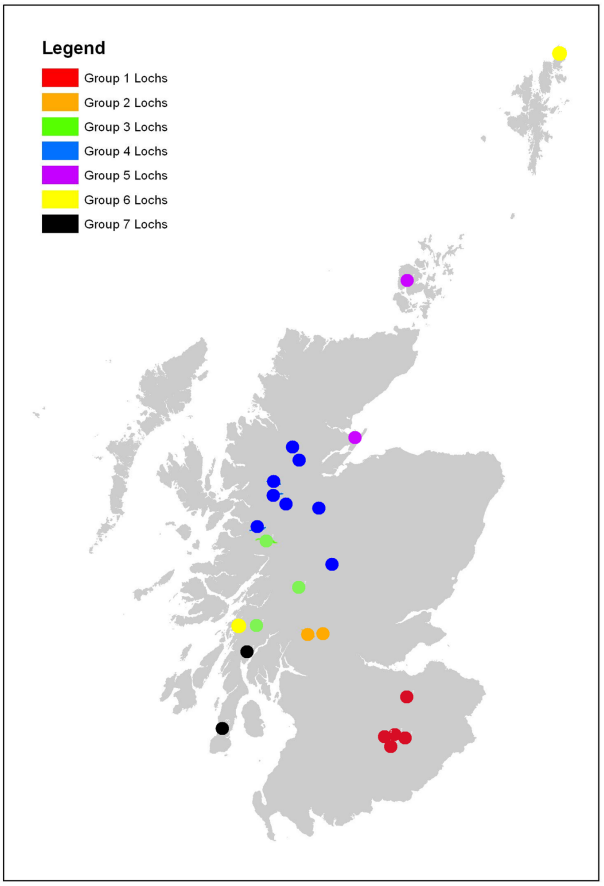


FIGURE 3.1: Map of Scotland with subset of lakes. Colours represent different SEPA groups for WFD classification.

from Figure 3.2 that the data were inconsistent in terms of both the quantity of data collected and the time period spanned. The lack of synchrony between sample dates across different lakes meant that it was difficult to compare lakes by comparing observations collected within a short period of one another, for example matching samples which were within seven days of one another. Comparing the samples by season (matching 3 month seasonal averages) was an option that was investigated, however, with only 4 years of data this would result in only 16 samples per lake. The black dashed lines on each plot in Figure 3.2 represent a subsection of the time period between early 2003 and late 2006 where it was thought there was a suitable quantity of data on most of the lakes. There were 21 of the lakes out of 24 which appeared to have a sufficient quantity of data over this common time period. This subset of data will be used throughout Chapters 3 and 4.

In view of this initial investigation of the available data, it quickly became clear that matching the samples at the lakes by date, or season, was not going

Lake	Name	Group	Rep	Location Code
1	Gladhouse Reservoir	1		7451
2	Talla Reservoir	1		9597
3	Fruid Reservoir	1		9598
4	St Marys Loch	1	X	9831
5	Megget Water	1		300344
6	Loch Katrine	2	X	7111
7	Glen Finglas Reservoir	2		10841
8	Loch Avich	3		103641
9	Loch Ba	3	X	126137
10	Loch Arkaig	3		233810
11	Loch Beinn a Mheadhoin	4		200307
12	Loch Mhor	4		200309
13	Loch Mullardoch	4		200311
14	Loch Monar	4		202909
15	Loch Glascarnoch	4	X	233763
16	Loch Quoich	4		233792
17	Loch Luichart	4		235710
18	Loch Garry	4		320796
19	Loch Eye	5		233768
20	Harray Loch	5	X	233892
21	Loch Tralaig	6		103642
22	Loch of Cliff	6	X	204025
23	Lussa Loch	7		103492
24	Loch Glashan	7	X	103388

TABLE 3.1: Table of Loch Grouping Details, current SEPA groups are shown and representative lakes (Rep) are identified using an X

to be an approach which would result in a quantity of data that was sufficient to reliably analyse any underlying relationships of interest. In addition to the problems associated with comparing lakes which are caused by the lack of data, it was also thought that comparing raw values alone may not be the approach best suited to capturing the maximum amount of information from the lakes. It may be more useful to compare patterns in the data over time - for example trends and seasonal patterns - and then investigate groupings by comparing the lakes on the basis of these more complex temporal features. This would allow the existing groups to be compared and, if necessary, re-structured, not only on the basis of the mean level of each determinand at each lake, but also on the basis of any common patterns over time. Potentially valuable information about the temporal dynamics of the three chemical determinands of interest could be lost if lakes are grouped only on mean values. Following from this, functional data analysis (fda)

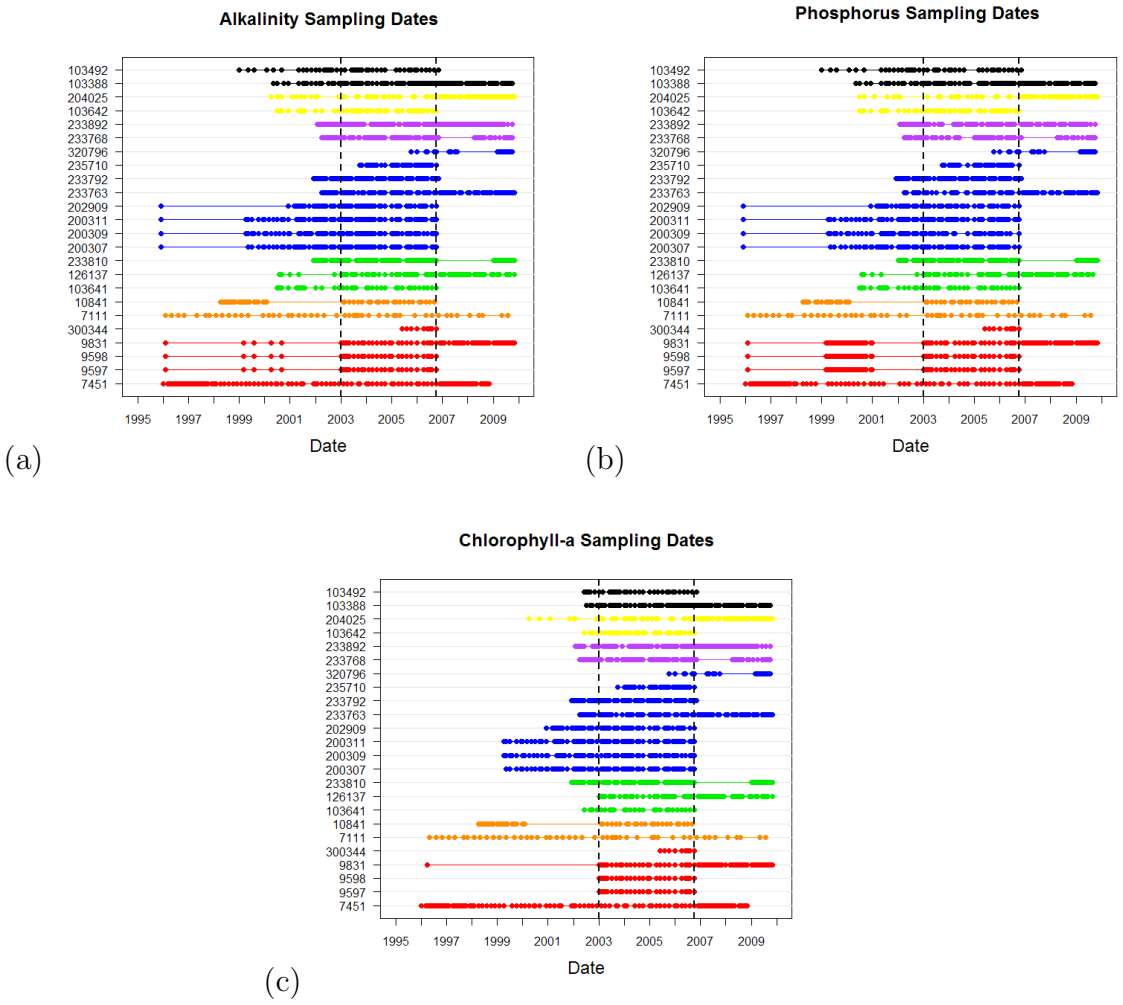


FIGURE 3.2: Plot showing sample dates for alkalinity, phosphorus and chlorophyll at Scottish Lakes

has been employed as this is an approach where the observed data are treated as a set of curves over time.

3.2 Functional Data Analysis (FDA)

It is natural to think of a time series of data as observations of a continuous function collected at a finite series of time points. In this context, this means the basic unit of interest can be thought of as a curve constructed from all observations collected from an individual, which here would be a lake. Functional data analysis describes analysis of data where the observations of interest are curves and in recent years functional equivalents to many standard statistical techniques have

been developed. Some of these techniques which are analogous to the standard approaches will be discussed in more detail later in this Chapter. Ramsay and Dalzell (1991) state that fda is a "natural sequel" to multivariate data analysis and although for both there is a finite set of observations available for each individual, the main difference is that for fda, these observations are viewed as discrete numerical representations of infinite-dimensional objects. Regarding the data in this way makes it easier to see if there are common long-term patterns in the data across individuals (lakes), and has a further advantage of overcoming some of the problems associated with irregularly spaced or sparse data, since the curves and not individual samples, become the objects of interest.

As it is unlikely that data will be in regular, functional form the first step is to estimate a smooth function of the observed data. While fitting non-linear functions will often capture more of the features of the data than a straightforward linear model, it is more difficult to compare and contrast the non-parametric smooth curves. Using splines, smooth curves can be fitted to samples from each individual, and subsequently these curves can be analysed. Ramsay and Silverman (1997) is a good reference for discussing basic functional data analysis techniques while Ramsay et al. (2009) provides details of applying these techniques in the `R` software package.

Estimation of functional data from potentially noisy, discrete observations is one of the main challenges in fda and has to be the starting point in any analysis. One popular technique which is widely used as a method for producing a smooth curve fitted to observed data is spline smoothing as discussed in Section 1.3.3. Ramsay and Silverman (1997) suggests that basis spline functions are more commonly used in fda than kernel approaches as they not only provide a large amount of flexibility, but are also a computationally efficient way to store information on functions which can potentially be constructed from a large number of data points. Using basis splines also allows functions to be expressed in such a way that matrix algebra can be used for most of any subsequent calculations.

3.2.1 Exploratory Functional Data Analysis

Assuming the estimated curves provide a good fit to the data then the original data can be discarded and the curves can be treated as the observations of interest. There are several standard statistical analysis techniques for which functional

equivalents have been developed and these can subsequently be applied to the estimated curves representing each individual. The techniques most commonly used in functional data include functional principal components analysis, functional regression (Henderson, 2006), functional linear discriminant analysis (James and Hastie, 2001) and functional cluster analysis. Functional cluster analysis is described in greater detail in Chapter 4 and a brief summary of functional regression (with a functional response) is provided in the following Section.

At a more basic level, there are also functional equivalents of summary statistics such as the mean and variance. Given a set of N curves $g_n(t)$ where $n = 1, ..., N$ measured at a set of time points, t, then the functional mean curve can be defined as the curve which is obtained by taking the mean at each time point,

$$\bar{g}(t) = \frac{1}{N}\sum_{n=1}^{N} g_n(t) \tag{3.1}$$

Similarly, the univariate sample variance curve is defined as the sample variance of the curves at each individual time point,

$$Var_g(t) = \frac{1}{N-1}\sum_{n=1}^{N}(g_n(t) - \bar{g}(t))^2$$

To summarize the dependence across different time values for a set of N curves, $g_1(t), ..., g_N(t)$, a covariance function can be defined. For all pairs of time points, t_1 and t_2,

$$Cov_g(t_1, t_2) = \frac{1}{N-1}\sum_{i=1}^{N}\{g_i(t_1) - \bar{g}(t_1)\}\{g_i(t_2) - \bar{g}(t_2)\}$$

The corresponding correlation function is then written as

$$Corr_g(t_1, t_2) = \frac{Cov_g(t_1, t_2)}{\sqrt{Var_g(t_1)Var_g(t_2)}}$$

In addition to summarising the dependence between time for the curves, cross-covariance and cross-correlation functions can also be defined to explore the relationships between multiple functional determinands which have been measured for each individual. Exploratory plots are also being developed as a tool in the initial analysis of functional data. Sun and Genton (2011*b*) have proposed a functional boxplot in order to visualize summary statistics of functional data as well

as identifying outliers.

3.2.2 Functional Regression Models

Further to functional summary statistics, it is possible to build functional regression models which are analogous to the techniques used in standard linear regression analysis. A linear model is said to be functional if any of the following hold,

1. the response variable is functional and the covariates are scalar;
2. both the response and one or more of the covariates are functional;
3. there is a standard scalar response with one or more functional covariates.

In the first of these cases, predicting a functional response using a set of scalar variables is known as functional multiple regression, while decomposing the variation in a functional response into functional effects using a scalar design matrix is known as functional analysis of variance (fANOVA). Functional multiple regression and fANOVA are equivalent to standard multiple regression and one way ANOVA respectively with the fundamental difference being that rather than estimating a set of regression coefficients, say β_j, a set of regression coefficient functions, $\beta_j(t)$, are estimated. Regression models with a functional response and scalar predictors are described with an example in Faraway (1997).

In the second case the model is more general and both the response and covariates are functional although scalar covariates can be included as constant functions over time. In the most straightforward situation, known as the concurrent model, both response and covariates are functions of the same argument, say time, and the model relates the value of the functional response to the value of the functional covariates only at the corresponding time points. The concurrent model is similar to varying coefficient models for standard data. Alternatively, it is possible for a model to be built where the argument for the response and covariate functions covers a different time period for each. In this case, a constraint has to be placed on this type of model so as to avoid backward causation and restrictions have to be imposed to ensure that at any time point, t, the time points $s < t$ cannot be used to predict the response value. This model is described as the historical linear

model in Malfait and Ramsay (2003). The third case is where there is a scalar response variable whose value is predicted by a set of independent variables, at least one of which is functional.

The different forms of functional regression models are explained in more detail in Ramsay and Silverman (1997). In this thesis it is of greatest interest to predict values of a functional response variable using a standard scalar design matrix (case 1 above) and so further details have been provided on the construction of an fANOVA model. Ramsay and Silverman (1997) has been used as the key reference in this section. Assuming there are N response functions (representing N different lakes), K groups and n_k individuals within each group (where $k = 1, ..., K$). Then the model, $g_{ik}(t)$ for the response at the i^{th} lake in the k^{th} group can be written as

$$g_{ik}(t) = \beta_0(t) + \beta_k(t) + \varepsilon_{ik}(t), \qquad \text{where } i = 1, \ldots, N. \tag{3.2}$$

In this model, $\beta_0(t)$ is the overall mean function across all N individuals (lakes) and $\beta_k(t)$ is the group effect which quantifies the departure from the overall mean corresponding to the k^{th} group. For any individual i in group k, the additional variation which is not explained by either the overall mean or the group effect is contained within the residual function ε_{ik}. If an individual can belong to only one group which is often the case then the additional constraint is required that $\sum_k \beta_k(t) = 0$ for all t so as to ensure that group effects can be uniquely identified. In order to fit this model, an appropriate $N \times K + 1$ design matrix $\mathbf{Z}$ is first defined. Each row of the matrix corresponds to a single lake, the first column consists entirely of ones to represent the overall mean, and the subsequent K columns correspond to each of the groups. The ij^{th} entry of $\mathbf{Z}$, can be written as

$$z_{ij} = \begin{cases} 1 & \text{if individual } i \text{ is in group } j; \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

If the vector of parameter functions to be estimated is written as $\beta = (\beta_{\mathbf{0}}, \beta_{\mathbf{1}}, ..., \beta_{\mathbf{K}})$ then the model described in Equation 3.2 can be written in matrix notation as

$$g(t) = Z(t)\beta(t) + \varepsilon(t)$$

where g is the N dimensional vector containing response functions and ε is a vector containing N residual functions. The constraint on the group effect functions can

also be expressed in matrix notation as $\sum_{j=2}^{K+1} \beta_j(t) = 0$. Assuming the same set of basis functions is used to represent each of the coefficient functions and that there are p_β basis functions, then the parameter vector β can be written as

$$\hat{\beta}(t) = B\theta(t)$$

where θ is a basis vector of length p_β, and B is a $K + 1 \times p_\beta$ matrix. As seen previously, the response functions can also be expressed using a basis function expansion. This time setting the number of basis functions to be p_g then the response function vector $g(t)$ can be alternatively written as

$$g(t) = C\phi(t)$$

where C is a $p_g \times K$ matrix in which the i^{th} row contains the expansion coefficients of function g_i where $i = 1, ..., N$. As in previous notation, $\phi(t)$ is a K dimensional vector containing the basis functions. It is possible for the same set of basis functions to be used in the expansion of both the response and the regression coefficient functions, in which case $p_g = p_\beta$ and $\phi = \theta$. In order to obtain an estimate for the vector of regression function coefficients, least squares can be used and so the fitting criterion we would aim to minimise is

$$\begin{aligned} SSE(\beta) &= \int [g(t) - Z\beta(t)]^T [g(t) - Z\beta(t)]\, dt \\ &= \int [C\phi(t) - ZB\theta(t)]^T [C\phi(t) - ZB\theta(t)]\, dt. \end{aligned} \tag{3.4}$$

While this expression assumes independent errors and constant variance it is possible to include a weight matrix to account for this. In order to include a roughness penalty when fitting the model a further adjustment has to be made to Equation 3.4. Although the functions fitted to the response variable have already been smoothed, Ramsay and Silverman (2003) suggests that in functional linear models the response functions should be smoothed very little, or even not at all, and instead smoothing should be carried out within the regression coefficient functions which are being estimated. The reason for this is that there may be important variability within the individual response functions which will play a role in the estimates of the regressions coefficient functions but could be overlooked if it has been lost by smoothing. Taking L to be any linear differential operator, for example, $\beta''(t)$, then a roughness penalty for β can be written as $\int [L\beta(t)]^2 dt$ and the

penalised integrated squared error becomes

$$SSE_{PEN}(\beta) = \int [C\phi(t) - ZB\theta(t)]^T [C\phi(t) - ZB\theta(t)]\, dt + \lambda \int [L\beta(t)]^2\, dt \quad (3.5)$$

where λ is a smoothing parameter. Further details of the model fitting procedure using least squares are provided in Ramsay and Silverman (1997). To ensure the estimates for each individual can be fairly compared, it is usual for the linear differential operator and the smoothing parameter used to both be kept the same for all of the curves. After fitting the model, confidence intervals can be constructed for the estimated functions by estimating an error covariance. Firstly, the residual for the i-th observation of the j-th curve can be written as

$$r_{ij} = y_{ij} - Z_j(t_i)\beta(t_i).$$

Next, by evaluating a fitted interpolating spline and the predicted model at a regular series of time points, a matrix of residuals can be constructed and written as r. Following this, an error covariance estimate can be found using the equation

$$\Sigma_e^* = \frac{1}{N} r r^T$$

Using this matrix, standard errors can be obtained and plotted for each of the estimated regression function coefficients.

3.2.3 Permutation Tests

In order to determine if there are any statistically significant differences between groups permutation tests can be used for functional hypothesis testing. Functional F-tests can be be used to test if there is any statistically significant relationship between functional variables while permutation t-tests can be used to test if there is any statistically significant differences between groups of functions.

Functional F-tests

A permutation F-statistic could be used in order to assess if there are any significant differences between the groups, or if a mean only model would be sufficient. For a set of N curves represented by the smooth functions $g_i(t)$, Ramsay and Silverman (1997) define the functional equivalent of the univariate F-test statistic

as:

$$F(t) = \frac{Var\left[\hat{g}(t)\right]}{\sum_{i=1}^{N}\sum(g_i(t) - g(t))^2} \tag{3.6}$$

where $i = 1, ..., N$ and $\hat{g}$ are the predicted values from a fitted fANOVA model such as that in Equation 3.2. This equation gives a function built from the series of point estimates at each of the time points, t. However, as with all hypothesis tests, in order to formally test the null hypothesis that there is no relationship between the functional variables a single test statistic is required, as well as a p-value which indicates the probability of observing a result as extreme, or more extreme, if the null hypothesis is true. Using the maximum of the test-statistic function, $F(t)$, as the test statistic, then a distribution of the test statistic under the null hypothesis can be obtained by calculating the test-statistic several times, each time using random permutations of curves. Shen and Faraway (2004) discuss an alternative approach for comparing nested functional regression models and propose the use of a functional F-statistic that is defined in terms of differences in the integrated residual sums of squares.

Assuming the N individuals form K distinct groupings then the null hypothesis for this test can be written as

H_0 : There is no difference between the K groups

and the corresponding alternative hypothesis as

H_0 : There is a difference between at least two of the K groups

The main idea behind the permutation test is that under the null hypothesis, for any given time, t, the pairing of the value of the i^{th} curve $g_{ik}(t)$ and the curve number i are entirely random. The procedure used to calculate the critical values is as follows

- Calculate the observed F-statistic function using Equation 3.6 and find the maximum of this, $\max\{F(t)\} = F_{obs}$
- Randomly re-label the curves with different curve numbers, but leave the grouping structure unchanged

- For the set of re-labelled curves calculate an F-statistic function ($F_{perm}(t)$), using Equation 3.6 at a fine grid of t time points, and find the maximum of this function
- Repeat this re-labeling procedure a set number of times, say $nperm$, and for each calculate the pointwise F-statistic function
- To find the pointwise 0.05 critical value of the null distribution, at each time point, t, calculate the 95th percentile of the $nperm$ F-statistic values corresponding to that time point
- To find the maximum 0.05 critical value of the null distribution, calculate the 95th percentile of the maxima of the $nperm$ permutations

The p-value corresponding to this test is the proportion of occasions where the maximum value of the permutation F-statistic function is greater than maximum of the observed F-statistic function, F_{obs}. The pointwise curve can be plotted alongside the observed test statistic curve in order to provide some indication of the time points at which the groups are least distinguishable.

Functional t-tests

Similarly to the functional F-test, a permutation t-test can be used to assess if there is any statistically significant difference between groups of functions. Assuming there are two distinct groups of curves, $g_1(t)$ and $g_2(t)$, with N_1 curves in group 1 and N_2 curves in group 2, then a t-test statistic function can be defined as

$$T(t) = \frac{|\bar{g}_1(t) - \bar{g}_2(t)|}{\sqrt{\sum_1^{N_1} Var\left[g_1(t)\right] + \sum_1^{N_2} Var\left[g_2(t)\right]}} \tag{3.7}$$

The maximum of the observed t-statistic function can be used as the test statistic and can then be compared to a relevant null distribution which is calculated from a set of permutations. Similarly to the functional F-test, this test is based on the idea that under the null hypothesis, for any given time, t, the pairing of the value of the i^{th} curve in the k^{th} group, $g_{ik}(t)$, and the group number k are entirely random. It should be noted that an assumption of the permutation t-test is that all groups of curves have the same variability. This assumption can be assessed informally by visual inspection of the curves. It may also be the case that prior

knowledge of the context of the data indicates whether or not this assumption is likely to hold. A null hypothesis that could be tested using a functional t-test is

H_0 : There is no difference between the mean of groups 1 and 2

and this would be set against the alternative hypothesis

H_1 : There is some unspecified difference between the mean of groups 1 and 2

A similar procedure to that outlined for the functional F-test can be used to estimate a distribution of the test statistic under the null hypothesis. Again a p-value is computed by calculating the proportion of occasions where the maximum value of the permutation t-statistic function is greater than maximum of the observed t-statistic function.

3.3 Application of FDA to the Lakes Data

The first step in applying FDA to the lakes data was to fit a smooth function to the observed sample values for each determinand at each lake. Ideally the observations for each of the lakes should cover the same time period, and have the same start and end date, to ensure that fair comparisons of temporal patterns across lakes can be made. For the Scottish lakes data, problems were encountered due to the sparsity and irregularity of the observed data. Although the irregular nature of the data at a single lake was not overly extreme, and there was roughly one observation per month over the time period considered, when looking at the sampling dates at all of the lakes together, the lack of consistently spaced observations caused concern.

To ensure fair comparisons of different curves, it is important that the same quantity of smoothing is applied to all lakes. Keeping the levels of smoothing consistent not only requires using the same smoothing parameter at each lake, but also requires that the knots are spaced so that there is the same quantity of data within each of the intervals defined by the knot placement. If there are different quantities of data between each of the knots, some basis coefficients will be estimated more accurately than others. In order to deal with the problem of

irregularly spaced observations, it was decided that missing data could be imputed at each lake by fitting a natural cubic interpolating spline to the data and subsequently extracting estimated values for the missing data. This ensures that there is a set of regularly spaced time points for each lake and hence enables fair comparison of the curves representing the different lakes. Following the imputation of any missing observations, smooth functions were fitted to the data using a cubic B-spline basis combined with the roughness penalty described in Equation 1.12. The B-spline functions used were order 4 and knots were initially placed at 3 month intervals - meaning there are 17 knots over the 4 year period considered, including the knots at the boundaries. Three months was chosen as a suitable interval as this time period represents one season in a year. Consequently, the relationship that states the number of basis functions is equal to the number of interior knots plus the order of the basis functions implies that 19 basis functions are being used to fit the functions.

A discussed in Chapter 1, there are clearly important choices to be made both as to the number of basis functions that should be used and what value should be selected for the smoothing parameter. GCV was initially considered to choose appropriate smoothing parameters for the lakes data, however this proved to be uninformative and plotting the smoothing parameter value against the GCV values computed produced a curve which was reasonably flat. After considering GCV, a sensitivity analysis was carried out in order to investigate the effect of different values of smoothing parameters on the fitted curves, and to determine a value that is suitable for the lakes data. In order to illustrate clearly the effect of the smoothing parameters, a subset of 3 lakes will be investigated in more detail. Figure 3.3 shows both a line plot of the log transformed phosphorus data (left) at 3 lakes and interpolating splines that were fitted to these data (right). The numbers on the plots represent the lake numbers (see Table 3.1). After any missing values were imputed from the interpolating splines, smooth functions were fitted to this data using a cubic B-spline basis with 3 month knots (19 basis functions). Figure 3.4 shows these smooth functions when 6 different smoothing parameters are used ranging from 1×10^{-6} to 0.1. As can be seen, using the smallest of the smoothing parameters (1×10^{-6}), there is evidence of under-smoothing. There are relatively harsh peaks and troughs, and a lot of local variation in the fitted function, although, as could be expected, this is not as severe as when the interpolating spline functions are used (Figure 3.3). While there is a vast difference in the relative sizes of the smoothing parameters 1×10^{-6}, 1×10^{-5} and 1×10^{-4},

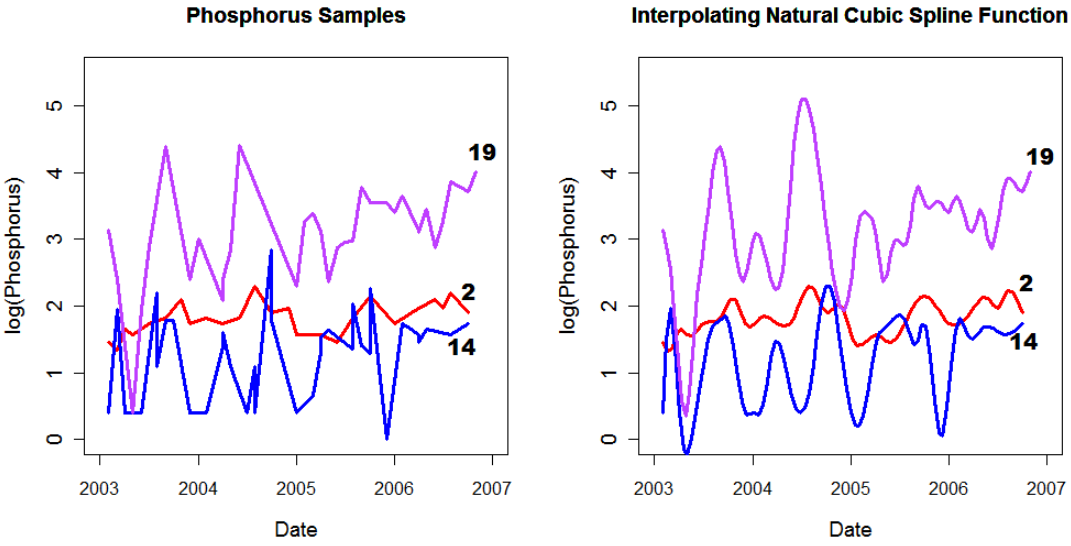


FIGURE 3.3: Plots of log (phorphorus $\mu g/l$) samples (left) and fitted cubic interpolating splines (right) at lakes 2, 14 and 19

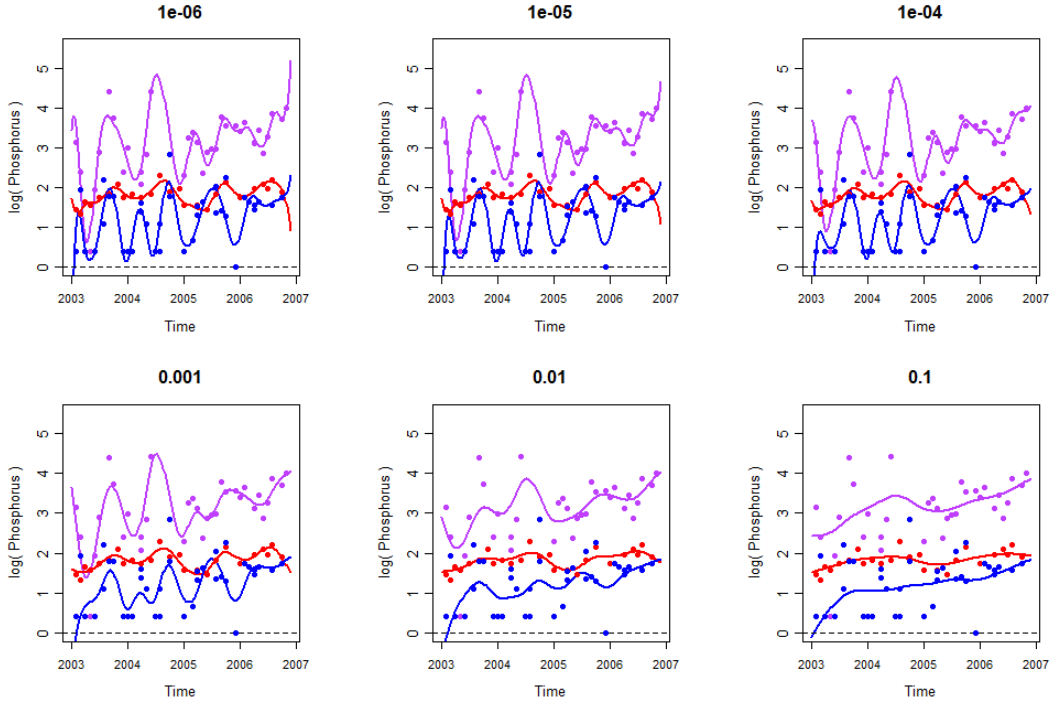


FIGURE 3.4: Plots of log (phorphorus $\mu g/l$) fitted spline functions at lakes 2, 14 and 19 with different smoothing parameter values

their effect on the level of smoothing in the fitting procedure is very similar and for all three values, the curves do not seem particularly smooth. In contrast, using the two larger smoothing parameter values considered here, there is evidence that the data may have been over-smoothed. The fitted curves are very flat and do

not accurately reflect some of the underlying patterns in the observed data. When the smoothing parameter equals 0.001, there appears to be a compromise between these two extremes with the key trends and potential cyclical features of the data being detected, while the localised random fluctuations are removed.

Using the value of 0.001 for the smoothing parameter, cubic spline functions were fitted to the log transformed alkalinity, phosphorus and chlorophyll data from the lakes. Although the sensitivity analysis has only been discussed for phosphorus, for all three of the determinands of interest this value of smoothing parameter ($\lambda = 0.001$) was found to be an appropriate choice. The estimated curves for each of the determinands are shown in Figure 3.5 with the different colours indicating the original SEPA groupings.

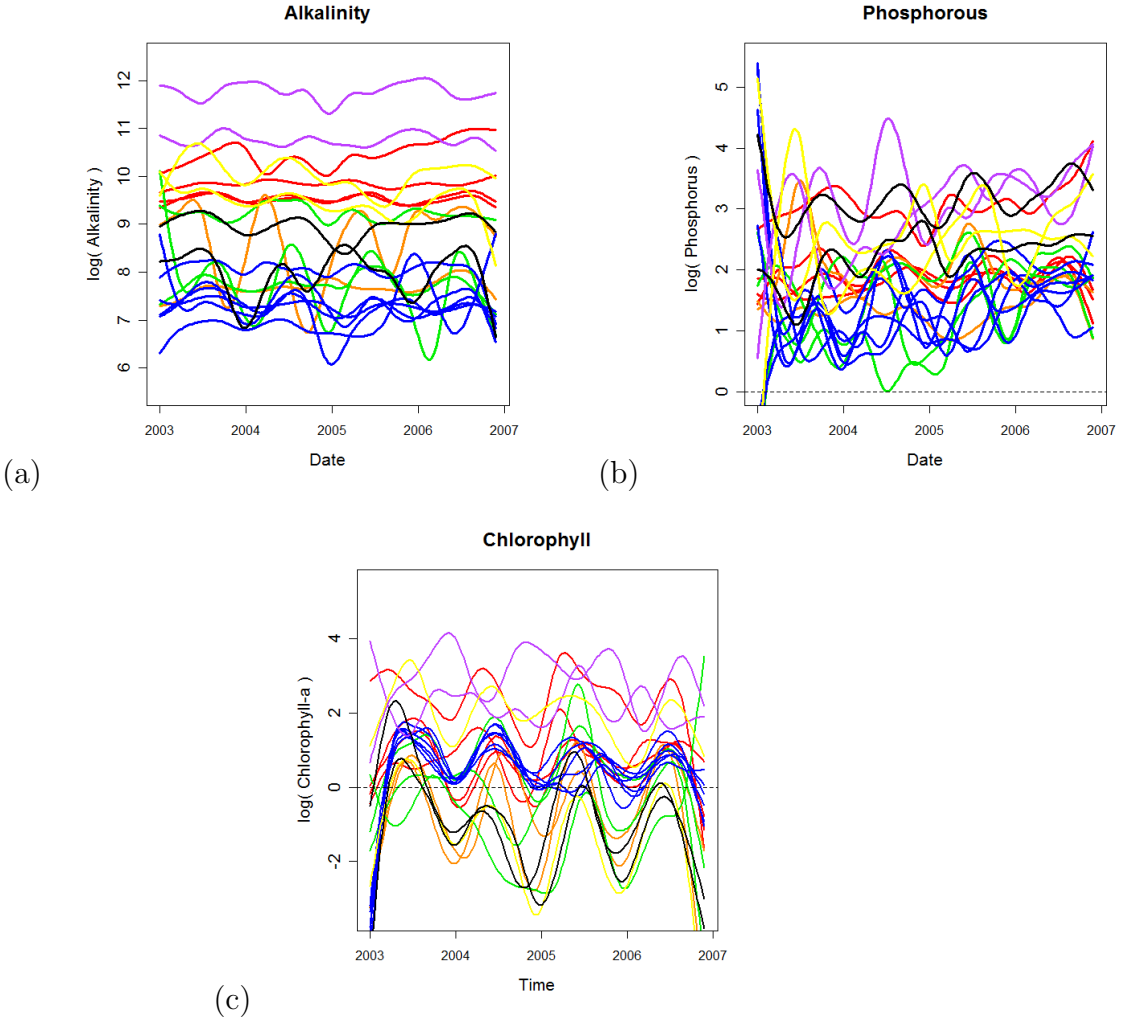


FIGURE 3.5: Fitted Spline functions for log(alkalinity $\mu g/l$), log(phosphorus $\mu g/l$) and log(chlorophyll $\mu g/l$)

The most apparent feature from Figure 3.5 is the huge degree of overlap in the curves and therefore in the existing groups for each of the determinands. It can be seen that there are similar patterns at some lakes, particularly in terms of mean level, within some of the currently used groups. For all three determinands there are groups of curves of the same colour that can be seen within Figure 3.5, for example, there are several blue (SEPA group 4) curves grouped closely together. The blue curves represent lakes which tend to be at the lower end of the scale for alkalinity and phosphorus and which have intermediate values for chlorophyll. For all three determinands it can be seen that the group consisting of the lakes represented by the purple curves (SEPA group 5) tend to have the highest values. Beyond looking at the mean levels of the curves, there is almost no apparent trend in any of the variables at any of the lakes although there is evidence of seasonality at some lakes. For chlorophyll, in particular, there seems to be a strong seasonal pattern in almost all of the lakes. Within the other two determinands the presence and strength of the seasonal signal is more variable.

Given the huge degree of overlap in the curves, the question is whether or not the current number of groups is optimal in terms of keeping within-group variation to a minimum. Following this Figure 3.6 shows the functional group means and the overall functional mean of all lakes (shown by the dashed black line) for each determinand separately. Again these plots highlight the large amount of overlap in the groups, particularly for phorphorus and chlorophyll. The alkalinity means, shown in Figure 3.6(a), appear to form three groups. It could be expected that alkalinity shows this greater degree of separation since the current SEPA groups are based on broad categories of alkalinity. The functional means imply strongly that there is little separability between the groups already in place in terms of mean levels of each determinand, however more formal techniques need to be used in order to determine if there is any scope to split the lakes based not only on the mean level, but perhaps on seasonality.

Representative Sites

Figure 3.7 shows the estimated curves for each lake (dashed lines) coloured by original SEPA grouping, with the representative lakes used by SEPA highlighted using the heavier, solid lines. As already discussed there is considerable overlap in the curves for different groups and so it is unsurprising there is overlap in some

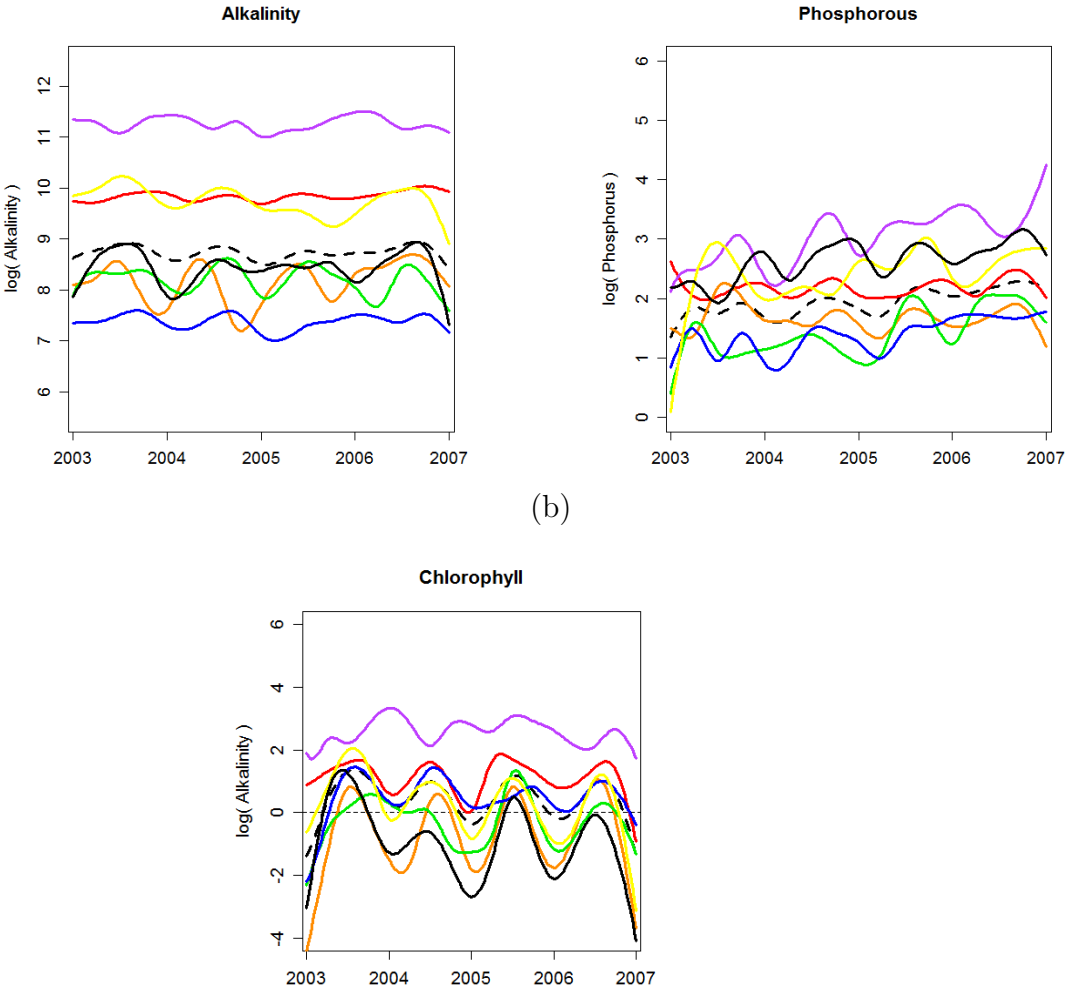


FIGURE 3.6: Functional Group and Overall Means for log(alkalinity $\mu g/l$), log(phosphorus $\mu g/l$) and log(chlorophyll $\mu g/l$)

of the representative curves. It is difficult to assess visually if these curves are representative of all members of the group. At present the representative sites are selected for a number of reasons such as ease of access to the site. It was thought an alternative method of selecting the representative site may be to select the site which is closest to the group mean curve shown in 3.6. The metric used to measure the proximity of each curve to it's relevant group mean curve was functional distance which is described in the Chapter 4, Equation 4.1. Figure 3.7 again shows the estimated curves for each lake (dashed lines) coloured by original SEPA grouping, with the representative sites as selected by using the minimum functional distance highlighted by the heavier solid lines. The representative lakes have been selected separately for each determinand using the minimum functional distance approach and so the representative lake within each group may not be

the same for all determinands. As can be seen, for several of the groups the representative site determined by the functional distance approach is the same as that currently used by SEPA.

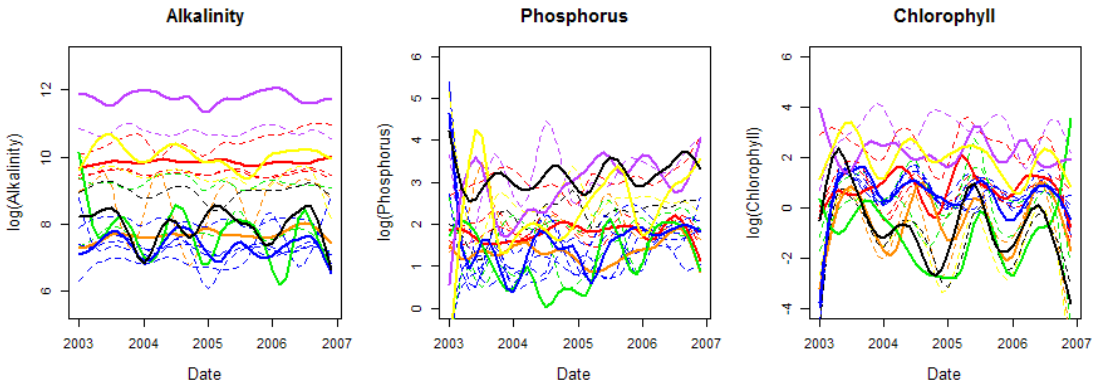


FIGURE 3.7: Plot of estimated functions for each lake (dashed lines) with SEPA representative lakes highlighted (solid lines)

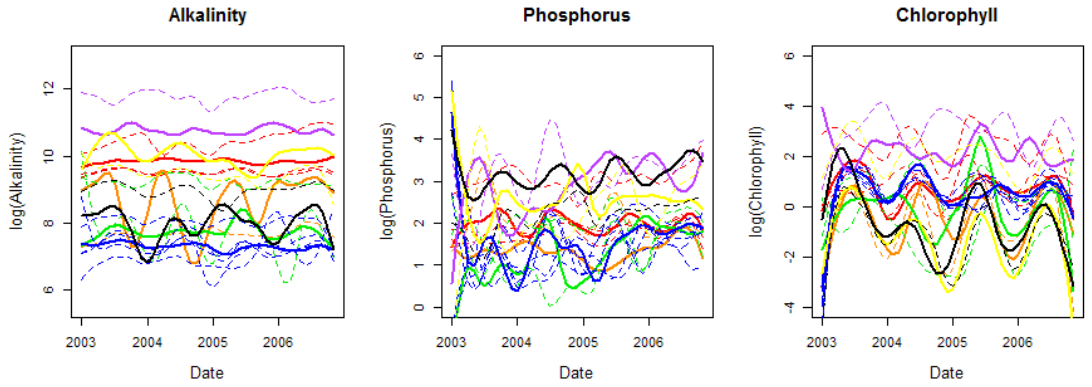


FIGURE 3.8: Plot of estimated functions for each lake (dashed lines) with representative lakes obtained using minimum functional distance approach highlighted (solid lines)

While looking at the functional means for each of the groups enables us to obtain an informal impression of how similar or different the current groups are, it is of interest to explore if any perceived differences are statistically significant. Functional linear regression can be used to estimate a group effect for each of the current groupings. The variance in the functional response (i.e. the functions fitted to the determinand values over time at each lake) will be decomposed into functional group effects by fitting a linear model where a categorical variable (the grouping structure) will form the design matrix.

Functional Regression

In order to determine if the groups of lakes that are currently used are statistically distinct we can estimate a functional regression coefficient for each of the groups and then determine the significance of these coefficients by constructing confidence intervals. Furthermore, permutation tests can be used to test the null hypothesis that there is no difference between the groupings. At this stage, each chemical determinand is considered separately. An example of functional regression will be discussed for alkalinity, although permutation test results will be presented for all three chemical variables.

A functional ANOVA model was fitted using unsmoothed functions of alkalinity, which were obtained by fitting natural cubic interpolating splines to the alkalinity data at each lake, as the response curves. Using the existing SEPA grouping structure, a standard design matrix was constructed where each row corresponded to an individual lake and each column corresponded to one of the parameters (coefficient functions) being estimated. Following this design, the number of regression coefficient functions that are estimated is equal to the number of groups plus one, which in the case of the alkalinity data is eight. The first regression function corresponds to the overall average while the subsequent seven functions relate to each of the group effects. It should be noted that a smoothing parameter value of 0.001 (the same as that used when fitting penalised cubic spline functions to the data shown in Figure 3.5) was used in the estimation of the regression coefficient functions. As proposed by Ramsay and Silverman (2003), smoothing was only imposed on the estimates of the regression coefficients and unsmoothed curves were used as the response functions to ensure that any potentially important patterns in the alkalinity data were not lost by smoothing the data twice. The model was fitted using the `fda` package in `R` (Ramsay et al., 2010).

Figure 3.9 shows the estimated regression coefficient functions for both the overall mean function and each of the groups with confidence bands plotted. The functional regression coefficient from a particular group gives an indication of how far away the curves in that group are from the mean level, meaning that if the dashed line at zero is outwith the confidence bands shown then that group is significantly different from the mean. From Figure 3.9, it can be clearly seen that groups 1, 4 and 5 appear to be different from the mean level at all time points. Group 6 is just above the mean level except at a couple of points, Group 7 overlaps

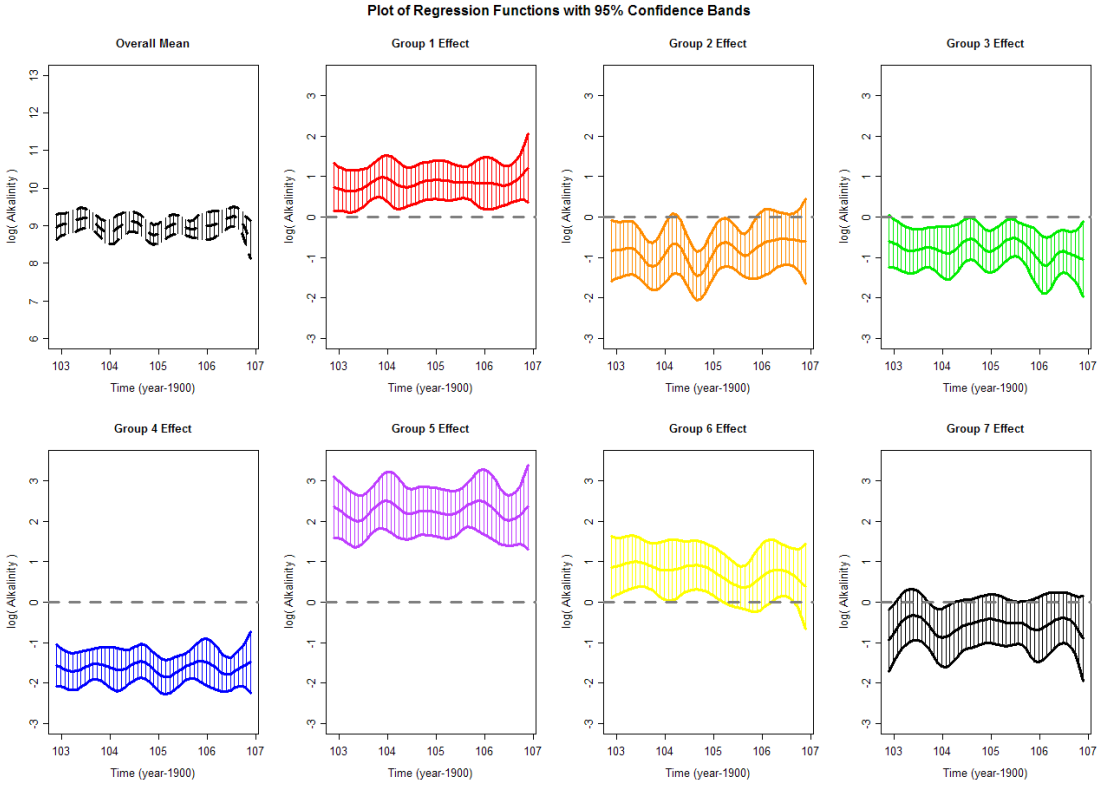


FIGURE 3.9: Estimated regression coefficient functions for log(alkalinity) mean and group effects (with 95% confidence bands)

the zero line, while groups 2 and 3 are just below the mean level with the exception of a few time points. The confidence bands on the estimated coefficient functions are wide due to the fact that there are very few observations in many of the groups. As well as this model, a second model was fitted using the penalised smooth curves as the response in order to ensure that using the unsmoothed functions as the response curves did not allow random fluctuations in the data to dominate the estimates of the regression coefficients function. The results of this model are not presented as the only difference was that the regression coefficient functions which were estimated using the penalised curves were slightly flatter than when the unpenalised curves were used and there was a less prominent seasonal pattern. The same differences were observed between the groups and the overall mean.

In order to test if the differences in the group effects were significant, two types of permutation test were used. The first was a functional F-test which tested the null hypothesis that there are no differences between the mean functions for all of the groups against the alternative that there are some unspecified differences between at least two of the groups. The procedure used for this permutation test

is outlined in Section 3.2.3. A plot of the observed F-test statistic function over time for alkalinity is shown in Figure 3.10, where the red curve represents the value of the observed test statistic over time as found from Figure 3.6. Also on this plot, the red dashed line represents the pointwise critical value and the red dotted line represents the maximum critical value, both at the 5% significance level. It is difficult to distinguish between the pointwise and the maximum critical value lines on this particular plot as they are so close. This test is based on a null distribution which has been constructed using 500 random permutations of the curve labels. As the red line lies above the pointwise critical level at all time points it is clear that there is sufficient evidence to reject the null hypothesis, and conclude that there are statistically significant differences between the groups in terms of the alkalinity levels. The p-value corresponding to this test (also included on the plot) is less than 0.001 which further indicates a highly significant result.

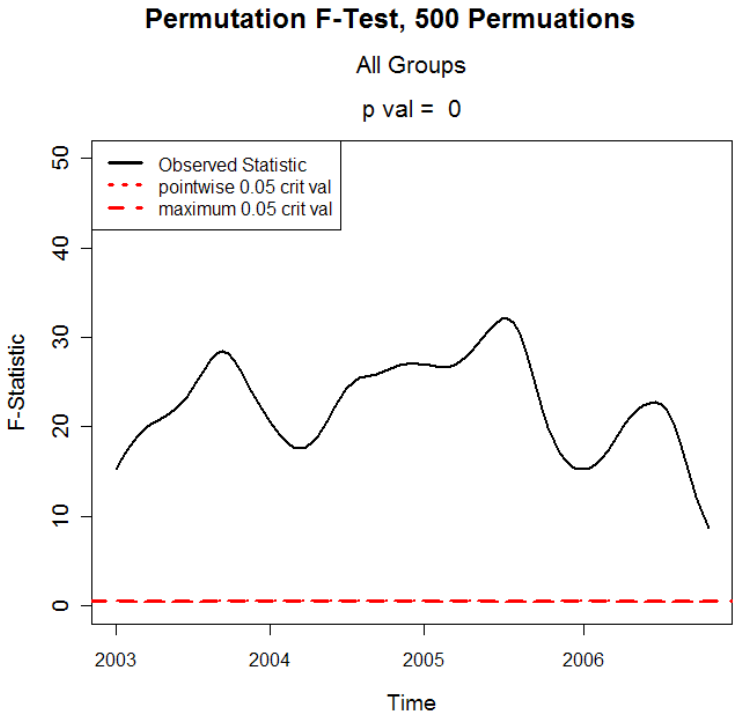


FIGURE 3.10: Plot of functional F-test for log(alkalinity $\mu g/l$) model

For all three determinands, the p-values of the functional F-tests are highly significant (< 0.001) and therefore imply there are clear differences between at least some of the SEPA groups in terms of their mean function. This is unsurprising, as although there was a huge amount of overlap between the current groups there were several pairs of groups where the mean level was quite different. To find

out more specifically where the differences lie, a functional t-test was next carried out for each pair of groups. The permutation t-test procedure which was used is outlined in Section 3.2.3. Associated with each pair of groups there is a p-value which corresponds to the test of the null hypothesis that there is no difference between the mean functions of two groups, against the alternative hypothesis there is some unspecified difference. These p-values are displayed in Table 3.2. To account for the 21 multiple comparisons a Bonferroni correction has been applied to the quantile of the null distribution that the observed t-statistic is compared to.

- Alkalinity
  The functional t-tests show that there are 13 out of the 21 pairs where there is a significant difference between the mean functions, so there does appear to be some level of separation between the current SEPA groups. Notably, there are significant differences between the mean of lake in group 5 and the mean of all other groups. There is evidence of this from Figure 3.5(a) where it can be seen that the group 5 lakes (shown in purple) appear to have a much higher mean level than all other lakes.

- Phosphorus
  For phosphorus there seems to be less separability between the mean functions of the groups compared to alkalinity and chlorophyll. Again, this is consistent with the earlier impressions obtained from Figure 3.6(b) which showed a huge deal of overlap in the group means. Group 5 again seems to be the most distinct of the groups as its mean function is different to the mean function of 3 of the 6 other groups.

- Chlorophyll
  There are 9 pairs of groups which differ significantly in terms of their mean Chlorophyll level. The mean function for group 6 is not significantly different from any other group. Looking back to the plot of group means shown in Figure 3.6(c) it can be seen that the mean of the group 6 lakes (shown in yellow) is almost directly in the center of all other group means.

3.4 The Effects of Correlation

Until now all analysis has been based on the assumption that the monthly observations from each of the lakes were independent and so the effects of temporal correlation do not have to be accounted for. To assess if there is any correlation through time for each variable, a first step was to produce autocorrelation function (ACF) plots for each lake and see if there was any correlation structure present in the observed data. The sample autocorrelation function was calculated for each variable at each lake over the 4 year time period of interest between 2003 and late 2006. These plots show, for approximately monthly data, the lag correlations after extracting the deterministic components (trend and seasonality) of the model, where necessary, using first and twelfth order differencing transformations. All three of the chemical variables were investigated and for all of them there appeared to be very little evidence of significant correlation at the majority of the lakes, The ACF plots again highlighted the differences between the lakes in terms of the seasonal patterns present, while at some lakes there is clear evidence of seasonality at others there is very little structure present.

Figure 3.11 shows an example of ACF plots for log alkalinity at a single lake (Lake 23, Lussa Loch). From the plot on the left hand side which is the sample autocorrelation function of the original data there appears to be a clear decreasing pattern which is indicative of a trend in the data. After first order differencing has been applied and the ACF is re-calculated for the transformed data, it is clear most of the structure has been removed. The ACF plot of the de-trended data indicates that there is no significant correlation over time. This was typical of the situation at the majority of the lakes. Although the available Scottish lake groups

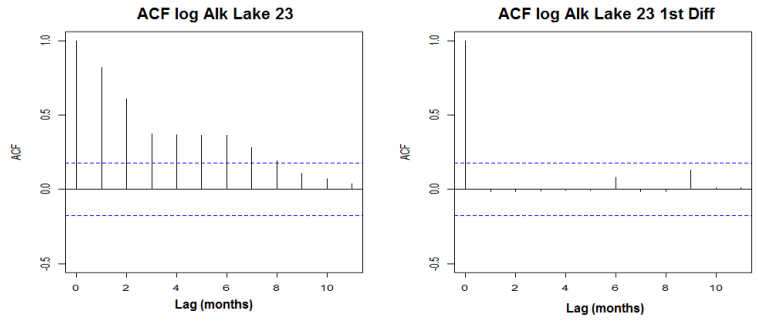


FIGURE 3.11: Plot of Autocorrelation Function for log(alkalinity $\mu g/l$) at Lake 23

Determinand	**Functional t-test p-value** 500 permuations 5% significance level Bonferroni corrected significance level = 0.24% 21 multiple comparisons					
Alkalinity						
		G1	**G2**	**G3**	**G4**	**G5**	**G6**
	G2	< 0.001					
	G3	< 0.001	0.56				
	G4	< 0.001	0.27	0.04			
	G5	< 0.001	< 0.001	< 0.001	< 0.001		
	G6	0.34	< 0.001	0.19	< 0.001	< 0.001	
	G7	< 0.001	0.70	0.58	0.04	< 0.001	< 0.001
Phosphorus						
		G1	**G2**	**G3**	**G4**	**G5**	**G6**
	G2	0.22					
	G3	0.15	0.80				
	G4	0.01	0.22	0.67			
	G5	0.10	< 0.001	< 0.001	< 0.001		
	G6	0.55	< 0.001	< 0.001	< 0.001	0.37	
	G7	0.06	0.34	0.18	< 0.001	0.70	0.66
Chlorophyll						
		G1	**G2**	**G3**	**G4**	**G5**	**G6**
	G2	< 0.001					
	G3	0.03	0.22				
	G4	0.07	< 0.001	0.21			
	G5	0.32	< 0.001	< 0.001	< 0.001		
	G6	0.89	0.67	0.81	0.97	0.29	
	G7	< 0.001	< 0.001	0.20	< 0.001	< 0.001	0.68

TABLE 3.2: Table of Functional t test p-values

data contains very little evidence of temporal correlation, because correlation is an extremely prominent feature of many environmental datasets (Morton and Henderson, 2008) it continues to be of interest to investigate what the effects of correlated errors would be on our ability to identify differences between groups of lakes. In order to explore these effects a simulation study was carried out. If correlation is present in the data the effective sample size of the dataset will decrease, and hence the size of the errors corresponding to parameters estimated from that dataset will increase. The main question of interest is whether or not the presence of correlation in functional data affects our ability to distinguish between groups of lakes. With a single dataset, functional permutation F-tests and t-tests can be used to determine if there any differences between the lakes or if there are differences between groups of lakes, however these tests do not take into account any correlation in the errors. For this reason, a bootstrapping procedure was used where initially a large correlated dataset was simulated and then re-sampled several times. To each of the re-sampled sets of simulated data, permutations tests were applied. For several different strengths of correlation, the distribution of the results of the permutation tests were summarized and compared. A detailed outline of how the study was carried out is provided below.

Step 1: Generating the correlated dataset

The first stage in the simulation study was to use the existing Scottish lakes data as a basis to generate a realistic correlated dataset. The log transformed alkalinity data were used to do this. Both the interpolating splines which were fitted to the observed data, and the penalised regression splines fitted to the grid of complete monthly values (shown in Figure 3.5) were evaluated at a fine grid of daily time points. The measurement errors for each lake was subsequently calculated by finding the difference between the observed (interpolating spline) and fitted (penalised regression spline) values. The result of this was that for each lake there were a set of estimated daily values from the regression spline and an estimated error variance which was obtained using the variance of the residuals corresponding to that lake. Figure 3.12 shows two plots which demonstrate the steps taken in calculating the error variance for lakes 1, 9, 12, 20 (see Table 3.1 for details of these lakes). This is an illustrative set of lakes and only these four are shown so that the plots are clearer. The plot on the left hand side show the interpolating splines (using dashed lines) and smoothed splines (using solid lines)

while on the right hand side the plot shows the magnitude of the residuals at each time point for each lake. Although the estimated daily data were used in

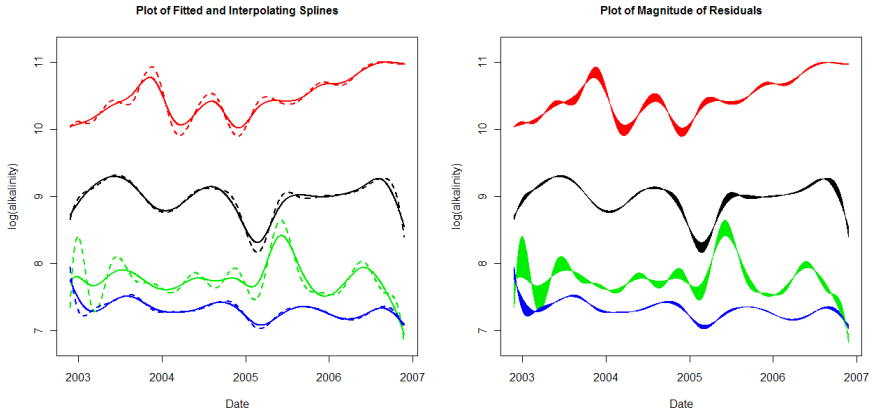


FIGURE 3.12: Plot of fitted and interpolating splines and plot of estimated daily residuals for log alkalinity at lake numbers 1, 9, 12 and 20

order to calculate the residuals and then the error variance for each lake, the next step was to simulate a sampling frame which could then be applied to this larger, daily dataset. For the 4 year period for which daily data were estimated using the smoothing splines, a single set of sample dates were chosen by randomly selecting a single date within each calendar month using a relevant uniform distribution. Following this, the estimated data values corresponding to these dates were then extracted from the large daily dataset. As a result of this procedure, there was a grid of randomly selected monthly samples available for each of the lakes of interest which were assumed to represent only the deterministic components of the process. The smoothing splines which were fitted to the original data were assumed to have removed random fluctuations and therefore were thought to only capture the more long-term features such as the trend over time and the seasonal pattern. Monthly samples were selected as this is the sampling frequency which is currently the most commonly used for standing waters by SEPA. In addition, random samples within each month were chosen since in practise, at each lake, samples are very unlikely to be collected on the same date in every month.

The next stage in generating the correlated data was to use a pre-specified correlation coefficient value and each of the lake specific error variances to generate a set of monthly random errors. These errors were simulated with an AR(1) structure implying that the correlation decreases exponentially as time lags increase. After generating the error terms these were then added onto the daily estimated

values obtained from evaluating the fitted smoothing splines at the selected sample dates, thus producing a monthly dataset based on the original observed data but with random errors that were correlated through time.

Step 2: Applying FDA to the simulated data

Using the simulated monthly dataset for each of the lakes with correlated errors, smoothing splines were fitted to the monthly values and then a fANOVA model was fitted. As with the observed data, the original SEPA grouping structure was used to construct the design matrix. A functional permutation F-test and t-test were then also used in order to assess if there were any statistically significant differences between the lakes and between groups of lakes. Since the standard errors are not incorporated in either the t or F-test statistic (see Section 3.2.3), comparing the results of permutation tests for simulated datasets with different strengths of correlation coefficient cannot indicate what, if any, effect temporal correlation has on our ability to distinguish between groups of lakes. However, the process of generating datasets outlined above was repeated and for each set of simulated monthly values the tests were applied. It was then possible to summarize the results of multiple permutation tests, for both tests, and to compare how often a significant result was obtained for each of the different values of correlation coefficient.

Firstly, 500 sets of different sampling dates were generated and the corresponding data values were extracted from the smoothing splines. Along with simulating 500 sets of independent data with no correlated errors to use as a control set, 500 datasets were simulated for correlation coefficients 0.2, 0.4 and 0.6. In order to ensure the comparisons of the results for each strength of correlation were as fair as possible, the same 500 sampling frames was used for each correlation value. While looking at the percentage of occasions where significant results were obtained gives an indication of the effect of correlation, it was thought it may be of greater interest to examine the distribution of p-values obtained. As stated previously, the p-values in each of the tests correspond to the proportion of occasions where the maximum value of the permutation test statistic function is greater than maximum of the observed test statistic function. For each of the 500 datasets simulated for a single correlation coefficient, the permutation test statistic function was based on 500 permutations.

For all of the 500 simulated datasets corresponding to the independent data and each of the correlation coefficient values of 0.2, 0.4 and 0.6, the functional F-test p-values were less than 0.001. This is unsurprising as the F-test tests the null hypothesis that there is no difference between any of the 7 groupings defined by SEPA. It is clear from looking at the observed data for alkalinity that there are differences between at least some group means and although there is correlation present in the simulated data, its effect will be relatively small in comparison to the size of the differences between the groups. It could therefore be expected that all tests would identify this difference as being statistically significant.

Following from this, the functional t-test assesses the null hypothesis that there is no difference between two groups of lakes. In order to investigate the effect of correlation on the ability to detect a difference between the mean of two groups, it is important to consider two groups where there is evidence of some difference. For this reason, the results of the tests which compare the mean alkalinity of the group 5 lakes to the mean alkalinity of all other lakes (those lakes not in group 5) will be explored. Histograms of the 500 p-values corresponding to this test are shown in Figure 3.13. The red line on each plot indicates the 5% significance level. In addition, Table 3.3 shows the percentage of significant p-values for this test when different strengths of correlations were used to generate the data.

Group 5 vs All Other Lakes Permutation t test results	
Strength of correlation (ρ)	% of significant p-values
0	76.2
0.2	66.6
0.4	70.6
0.6	64.8

TABLE 3.3: Percentage of significant t-test values for the difference between Group 5 and all other lakes

It could be expected that because temporal correlation effectively reduces the amount of information available at each lake it will become harder to differentiate between the groups and so the percentage of significant p-values will decrease. While from Table 3.3 there is no clear decrease in the percentage of significant results as correlation in the underlying data increases, it can be seen that the independent data has the highest proportion of occasions when a difference between the two groups was detected. Figure 3.13 shows that for the independent data,

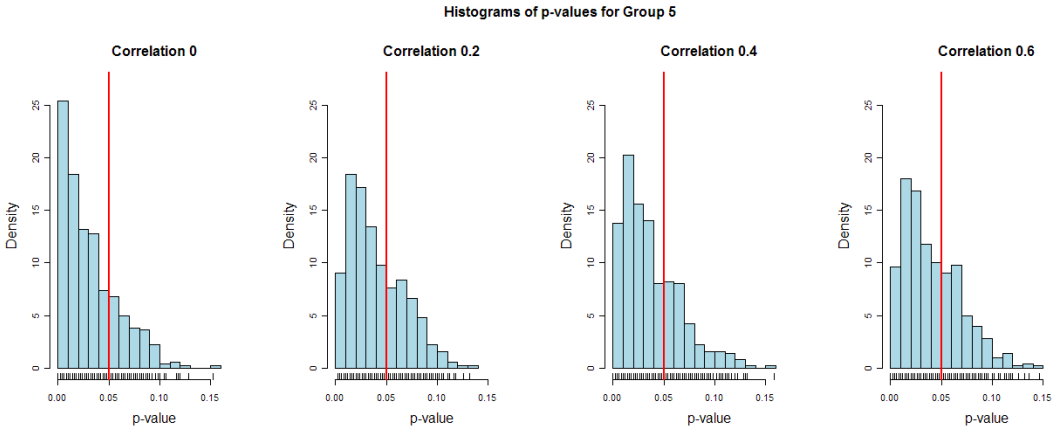


FIGURE 3.13: Histogram of p-values for results of permutation t-tests applied to 500 simulated dataset with temporal correlation (Group 5)

more of the p-values calculated were close to zero in comparison to the corresponding proportion when correlation was present.

In summary, while it does not appear that the presence of temporal correlation in the data has a marked effect on the ability to distinguish between different groups of lakes, there is some evidence from the above simulation study to suggest that there may be a limited effect. If there is strong autocorrelation in the data the potential effect of this should be taken into account when interpreting the output of functional permutation tests. This study has looked at monthly observations since this is the sampling frequency of interest for the Scottish lakes data however, in other contexts if the observations were more frequently collected it is likely the correlation may have a bigger impact.

3.5 Summary

Fitting smooth curves using splines, and more specifically using penalised regression splines, is both a computationally efficient and flexible way of estimating the true functions underlying the observed data. In addition, although not ideal, using natural cubic interpolating splines to first obtain a regular grid of data enables several of the problems associated with comparing lakes where there are irregularly spaced sample dates and different quantities of data available. The functions fitted using this approach appear to provide a good fit to the data.

Exploratory functional data analysis of the curves fitted to the lakes data has highlighted the huge degree of overlap of the group mean functions of all variables considered. This is unsurprising given there a relatively small number of observations and a relatively large number of groups. Alkalinity appears to show most distinction between the SEPA groups, however this could again be expected as the groups already used are primarily based on broad categories of alkalinity and so it could be expected this would drive any differences between the existing groups. For both phosphorus and chlorophyll there is less evidence of differences between the group means looking only at the estimated curves for each lake. Functional regression proved to be a useful tool as it can be used to estimate group specific effects which summarise the group data and can be compared to one another. There is however a great deal of uncertainty associated with each of these group effects when calculated for the Scottish lake data due to the fact the estimates are based on relatively few observations. While the permutation F-tests prove for all determinands that there are at least some differences between the current groups, the results of t-tests reinforced that not all of the current groups were distinct from one another and that fewer groups may be sufficient. SEPA Group 5 appears to be the most distinct from the other current groups, particularly in terms of alkalinity.

All of the initial exploratory investigations of the lakes suggests that the number of groups required to accurately capture the variability of the lakes could be reduced, potentially with some of the current groups being combined to form new, larger groups of lakes. The exact way in which the groups should be re-structured and to what extent the number of groups should be altered still has to be investigated more thoroughly. In order to do this functional cluster analysis has been considered.

Chapter 4

Functional Clustering of Water Quality Data

While Chapter 3 considered how a functional data approach could be used both to fit curves to observed data for an individual lake, and how pre-defined groups of these curves could be compared, one of the key ideas of this thesis is to explore alternative statistical approaches to how groups of functional data objects can be determined. For standard multivariate data, cluster analysis is a technique which is used to determine group structures in data where there are multiple determinands measured on each individual. For functional data, there are also multiple observations collected on each individual, it is just that these are values of the same determinand collected over a period of time. This Chapter will first discuss briefly the main idea behind cluster analysis and will then consider how clustering techniques for standard multivariate data have been further developed for functional data.

Cluster analysis is an automatic technique which is used to classify individuals or objects into mutually exclusive groups (called clusters) based on similarities of measurements which have been collected on these objects. Objects within each cluster are more similar to one another than objects which are assigned to different clusters and, since the primary aim of cluster analysis is to find group structures in the data, no group structure is defined before cluster analysis techniques are applied. It is a widely used technique and is applied to data from a broad range of different fields such as bio-informatics, social sciences and data mining. There are several different standard methods that can be used to cluster individuals. The

most commonly used of these are agglomerative hierarchical techniques (Ward, 1963), where each individual object initially forms a cluster, then on the basis of a measure of similarity these clusters are merged to form larger groups, and k-means (MacQueen, 1967). K-means is an iterative partitioning procedure where the number of groups is first specified, and then objects are moved from group to group, until the within-group sums of squares is minimised. Hartigan (1975), Mardia et al. (1980) and Kaufman and Rousseeuw (1990) provide a comprehensive introduction to different forms of standard clustering.

In addition to these methods for grouping, model based clustering procedures such as that described in Banfield and Raftery (1993) and Fraley and Raftery (1998) are a popular choice. A review of model based clustering provided by Fraley and Raftery (2002) states that while there has been extensive research into non-probabilistic clustering techniques, such as hierarchical and k-means, these methods fail to address several key questions such as how many clusters are optimal and how outliers are treated - these are problems which model based approaches often overcome. Model based clustering views the data as arising from a finite mixture of probability distributions with a single component distribution representing each different cluster. Since the models are constructed within a statistical framework, standard model comparison criteria can be used to compare different models. Comparing mixture models with different numbers of component distributions can consequently determine the optimal number of clusters. Moreover, the probability that any particular individual falls into a given group can be calculated. These are some of the features of model based clustering which make it an attractive choice.

Clustering Functional Data

Following from standard clustering methods, a range of techniques are also being developed for clustering functional data. The majority of approaches to functional clustering can be classed as dimension reduction methods, where a functional data object is first estimated for each individual using a finite dimensional basis, and then individuals are grouped by applying some clustering method to the basis coefficients that define these smooth functions. As well as filtering, another approach to functional clustering is to split the time interval over which the function is estimated into discrete sections, resulting in a dataset for each curve which can

subsequently be clustered. Both filtering and regularization have several potential drawbacks which need to be considered. One problem with the regularization approach is that this method results in data for each individual which are both high dimensional and autocorrelated, although James and Sugar (2003) state that a regularization constraint can be imposed in order to account for this and prevent unstable within-cluster covariance estimates. A further issue with regularization is that it cannot be applied if the data are irregularly sampled, since different individuals may not have observations at a common set of time points. The filtering approach also has problems when the data are sparse or irregularly sampled. If there are not a regular set of observations for each individual, then some curves, and hence some basis coefficients, will be estimated more accurately than others, meaning that not all individuals are directly comparable in terms of their variability.

Both non-probabilistic and model based functional clustering approaches have been developed, the majority of which use the filtering method as the first step. Hierarchical clustering methods for functional data are outlined in Henderson (2006) while Abraham et al. (2003) applies k-means methods to estimated spline coefficients in order to cluster sets of curves. Ignaccolo et al. (2008) applies a similar non-hierarchical technique of k-mediods (Kaufman and Rousseeuw, 1987), to cluster air quality data from different stations. Garcia-Escudero and Gordaliza (2005) also explores functional clustering of air quality networks using a variation on the k-means approach. Model based functional clustering approaches are being developed and are often applied in order to establish patterns in gene expression data such as in Luan and Li (2003) and Chudova et al. (2004). Other examples of functional clustering of gene expression data are provided in both McNicholas and Murphy (2010) and Shaikh et al. (2010). McNicholas and Murphy (2010) propose using mixtures of multivariate Gaussian distributions with a modified Cholesky-decomposed covariance structure in order to explicitly account for the serial correlation that is potentially present in longitudinal data. Shaikh et al. (2010) proposes a similar approach for clustering of longitudinal data using mixtures of Gaussian distributions with a modified covariance structure which can accommodate incomplete data series. A model based functional clustering method which is of particular interest in the context of the Scottish lake data is discussed in James and Sugar (2003), where a clustering model is proposed for functional data which have been sparsely or irregularly sampled. More recently, Chiou and Li

(2007) developed a model based approach which uses eigenfunctions rather than B-spline basis functions in order to initially define the functional data.

In this thesis, two of the approaches mentioned above will be discussed in more detail and implemented using the Scottish lake group data. Initially, the hierarchical method of clustering lakes from Henderson (2006) will be explored and then used to provide a visual representation of similarity amongst the lakes. After this, the more formal, model based approach outlined in James and Sugar (2003) will be investigated and applied.

4.1 Hierarchical Clustering

With standard agglomerative hierarchical clustering techniques the primary aim is to partition a set of N objects into clusters, where each individual object initially forms a cluster, then on the basis of a measure of similarity, these N clusters are merged iteratively to form progressively larger groups. The hierarchy of clusters can then be summarised in a dendrogram.

In order to measure similarity, the distance between pairs of observations are quantified by a metric. Common metrics used to measure the distances between two individual points include Euclidean or squared Euclidean distance and maximum dist ance. In addition, a linkage criterion is also required to determine how the clusters are formed. Possible linkage criteria include complete linkage, where the distance between two clusters is computed as the distance between the two farthest elements in the two clusters, single linkage, where the distance between two clusters is computed as the distance between the two closest elements and average linkage where the distance between two clusters is equal to the distance between the cluster means. To apply hierarchical clustering to a set of points, a distance matrix D is first calculated which contains the distance between all possible pairs of points. The i, j^{th} entry of D is the distance between points i and j as determined by whichever metric has been chosen. Although calculating the distance between pairs of functional objects seems slightly more difficult to compute, Henderson (2006) states that the idea of measuring distances is easily transferable from pairs of points to pairs of curves and defines a method of computing a functional distance matrix as follows.

Let the i^{th} and j^{th} curves, $g_i(t)$ and $g_j(t)$, be expressed as a linear combination of basis functions with coefficient vectors $\mathbf{c}_i$ and $\mathbf{c}_j$ respectively. The distance between the curves can then be written as

$$d_{ij} = (\mathbf{c}_i - \mathbf{c}_j)^T W (\mathbf{c}_i - \mathbf{c}_j) \tag{4.1}$$

In the above expression $W = \int \phi(t)\phi^T(t)dt$ is a matrix which is of equivalent form to the roughness penalty matrix (denoted by R in Equation 1.17). It is a symmetric square matrix of order P, where P is the number of spline basis functions. For each set of basis functions, W can be evaluated using numerical integration, if necessary, and the functional distance matrix D with entries d_{ij} as defined above can be computed. Standard algorithms for hierarchical clustering can then be applied to the functional distance matrix.

Hierarchical clustering is straightforward to implement and it provides an exploratory picture of the data, however, the results in the form of a dendrogram are often difficult to interpret. Another potential limitation of a hierarchical approach is that there is no way of quantifying uncertainty in the cluster partitions and the results obtained can often be sensitive to the choice of linkage criterion and distance metric selected. Furthermore, selecting the number of clusters from inspection of a dendrogram is somewhat subjective as different people will identify different groupings in the data. While there is often no clear visual indication as to what number of groups is optimal in terms of minimising the between-group homogeneity visually, there are techniques available that can be applied to investigate the most appropriate number of clusters. One such approach is the gap statistic which is discussed later in Section 4.3. Model based approaches have the additional advantage that model selection techniques can often be used to answer the question of how many clusters to choose. It was felt worthwhile to investigate both the hierarchical techniques for functional clustering and the more formal, model based approach in order to obtain as full an understanding of any group structure in the Scottish lakes data as possible.

4.2 Model Based Clustering

With model based clustering the assumption is that the observations are generated according to a mixture distribution with a fixed number of components. James

and Sugar (2003) and Fraley and Raftery (1998), who provide a clear description of how mixture models for cluster analysis are constructed and fitted, have been used as key references in this section.

Assume there are N observations, $x_1, ..., x_N$ and these can be classified into at most G clusters. Then let $f_k(x_i|\theta_k)$ be a multivariate normal density function of observation x_i from the k^{th} component which is parameterized by θ_k and let $\mathbf{z}_i = (z_{i1}, ..., z_{iG})$ be the, initially unknown, cluster membership vector for the i^{th} observation. If observation x_i belongs to cluster k then $z_{ik} = 1$, otherwise $z_{ik} = 0$. There are two ways in which the model representing the clusters can then be constructed. The first is using the classification likelihood approach where the likelihood to be maximised is written as

$$L_{class}(\theta_1, ..., \theta_G, \mathbf{z}_1, ..., \mathbf{z}_1|x_1, ..., x_N) = \prod_{i=1}^{N} f_{z_i}(x_i|\theta_{\mathbf{z}_i}) \tag{4.2}$$

When the identity covariance matrix (multiplied by a constant scalar variance) is used within each of the component multivariate normal densities in Equation 4.2 then the solution (when estimated using the Classification EM algorithm) is equivalent to that of the k-means approach. The second approach, known as the mixture likelihood approach, views the cluster membership vectors as being multinomial random variables, rather than being parameters which have discrete values. Assuming $\mathbf{z}_i$ is multinomial with parameters $\pi_1, ...\pi_G$ then the probability that x_i belongs to cluster k can be written as π_k and parameter estimates can be obtained by maximising the likelihood

$$L_{mix}(\theta_1, ..., \theta_G, \mathbf{z}_1, ..., \mathbf{z}_1|x_1, ..., x_N) = \prod_{i=1}^{N} \sum_{k=1}^{G} \pi_k f_{\mathbf{z}_i}(x_i|\theta_{\mathbf{z_i}}) \tag{4.3}$$

Maximisation of both the classification likelihood and the mixture likelihood require an iterative procedure such as the Expectation-Maximisation algorithm (Dempster et al., 1977). Within the EM algorithm, in the context of cluster analysis,

the $\mathbf{z}_i$'s are considered to be missing. Starting with initial values of $\mathbf{z}_i$, the EM algorithm alternates between updating the maximum likelihood estimate of the parameters conditional on the current $\mathbf{z}_i$'s and updating the $\mathbf{z}_i$'s with their expected value conditional on the current parameter estimates. These two steps iterate until some convergence criteria are satisfied.

K-means clustering

To obtain the initial estimates of cluster membership within the EM algorithm in model based clustering, K-means clustering is often used. As before, assume there are N observations, $x_1, ..., x_N$ and these can be classified into at most G clusters. The k-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Let $\bar{x}_k$ be the mean of cluster k where $k = 1, ..., G$. The squared error between $\bar{x}_k$ and the the points x_i in cluster k is defined as

$$\sum_{x_i \in k} ||x_i - \bar{x}_k||^2$$

The aim of K-means is to minimize the sum of the squared error over all G clusters. Since the squared error decreases as the number of clusters increases, the number of clusters has to be fixed in advance. After selection of an arbitrary initial partition, new partitions are formed by assigning each individual to its closest cluster centre, and then updated cluster centres are computed based on this new partition. These steps are repeated until cluster membership stabilizes. For functional data each individual, and the cluster centres, can be defined in terms of the sets of basis coefficients which define the curves. Krzanowski and Lai (1988) states that although convergence of the squared error to a global optimum is not guaranteed, checks on the worth of the final solution can be made by repeating the computations several times with different random starting paritions.

4.2.1 Model Based Functional Clustering

Like standard model based clustering approaches, the functional clustering model (FCM) proposed by James and Sugar (2003) not only enables individuals to be partitioned into distinct groups, but also provides a confidence in classification by quantifying the uncertainty in the partition. In addition, one of the most attractive features of this particular model is that it accounts for sparse and irregularly spaced data, which is a clear problem not only in the Scottish lakes data, but also with other environmental data-sets.

In order to fit the FCM a filtering procedure is first used in which a finite set of basis functions is employed to estimate functional data objects (curves) for each

individual. In contrast to other non-hierarchical functional clustering methods, such as Abraham et al. (2003), the estimated basis coefficients are not treated as fixed effects, but are instead modelled as random effects. This is the key difference between the FCM and other model based functional clustering methods. The advantage of treating the individual effects as random is that a regular fine grid of data is not required for each individual since strength can effectively be borrowed across individuals, assuming the total number of observations over all individuals is large. Treating the individual effects as random effects also ensures the model is more efficient, since the number of parameters will be less than if each individual effect is estimated using different fixed parameters. An application of the model is provided in Pastres et al. (2011) where the univariate FCM is implemented using data from water quality monitoring stations at sites in Venice.

A detailed description of the FCM is now given. Let there be N individuals and let m_i denote the number of observations for individual i. Then the function which represents the i^{th} individual curve at time points $t = (t_{i,1}, ..t_{i,m_i})$ can be written as

$$Y(t_{ij}) \quad = \quad g_i(t_{ij}) + \epsilon_{ij}, \text{ where } i = 1, ..., N \text{ and } j = 1, ...m_i. \tag{4.4}$$

Here $g_i(t_{ij})$ is the true value of the i-th curve at time t_{ij} and ϵ_{ij} is the corresponding measurement error. Dropping the time index notation then Equation 4.4 can be written more simply as $\mathbf{Y_i} = \mathbf{g_i} + \boldsymbol{\epsilon_i}$. It is assumed $\boldsymbol{\epsilon_i} \sim N(0, \sigma^2)$ and are independent. Following this, $\mathbf{g}_i$ can be expressed as a linear combination of P natural cubic spline basis functions as discussed in Section 1.3.3, so that

$$\mathbf{g}_i = s^T \boldsymbol{\eta}_i \tag{4.5}$$

where s is a P dimensional spline basis vector. The vector of spline coefficients, $\boldsymbol{\eta}_i$ can be modelled using a Gaussian distribution. Assuming there are G distinct groupings of lakes and that site i belongs to group k where $k = 1, ..., G$ then $\boldsymbol{\eta}_i$ can be further defined as

$$\boldsymbol{\eta}_i = \boldsymbol{\mu}_k + \boldsymbol{\gamma_i}, \tag{4.6}$$

In this equation it is assumed $\boldsymbol{\gamma}_i \sim N(0, \Gamma)$ and $\boldsymbol{\gamma}_i$ is independent of $\boldsymbol{\gamma}_j$ for $i \neq j$. This parametrization splits the spline coefficients into a random group effect, $\boldsymbol{\mu}$, and an individual effect, $\boldsymbol{\gamma}$. Consequently, $\mathbf{g_i}$ can itself be expressed as the sum of

a group effect and an individual effect as

$$\mathbf{g_i} = \mathbf{g_k} + \mathbf{s}^T\boldsymbol{\gamma_i}$$

where $\mathbf{g_k}$ is the mean of the k-th group and can be written as $\mathbf{g_k} = \mathbf{s^T}\boldsymbol{\mu}_k$ using Equations 4.5 and 4.6. A further re-parameterisation of the cluster mean spline coefficients is also written as follows

$$\boldsymbol{\mu_k} = \boldsymbol{\mu_0} + \boldsymbol{\Lambda\alpha_k} \tag{4.7}$$

In this expression $\boldsymbol{\mu_0}$ is a P dimensional vector which represents the overall mean for all lakes, $\boldsymbol{\alpha}_k$ is a h dimensional vector and $\boldsymbol{\Lambda}$ is a $P \times h$ matrix where $h \leq min(P, G-1)$. The additional parameterisation of the cluster means using h is another of the key differences between the FCM and other functional clustering approaches. James and Sugar (2003) state that there are two main advantages of formulating the cluster means in this way. The first being that setting $h < G-1$ reduces the number parameters to be estimated and therefore makes the model more straightforward computationally. This is of limited benefit when the number of clusters is already fairly small. Moreover, a further benefit of the cluster means being structured in this way is that low-dimensional projections of both estimated curves and cluster centres can be produced. Plotting these low dimensional representations enables clusters of individuals to be more easily identified than simply looking at curves alone. A detailed outline of how low-dimensional projections for each curve are constructed is provided in James and Sugar (2003), Section 3.1.

The functional clustering model can then be expressed as

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{S}_i(\boldsymbol{\mu}_0 + \boldsymbol{\Lambda\alpha}_k + \boldsymbol{\gamma}_i) + \epsilon_i \\ \epsilon_\mathbf{i} \sim \mathbf{N}(\mathbf{0}, \sigma^\mathbf{2}\mathbf{I}) &\text{ and } \boldsymbol{\gamma_i} \sim \boldsymbol{N(0, \Gamma)} \end{aligned} \tag{4.8}$$

where $i = 1, .., N$ and $\mathbf{S}_i$ is the spline basis matrix for the i-th curve evaluated at time points $t_{i,1}, ..., t_{i,m_i}$. It is assumed that all lake effects have a common covariance structure represented by Γ. In Equation 4.8 it is clear that the spline coefficients representing each of the curves are the sum of an overall mean effect, represented by the parameter $\boldsymbol{\mu}_0$, a cluster mean effect, represented by $\boldsymbol{\Lambda\alpha}_k$, and an individual effect, which is modelled as being a random effect, and is represented by $\boldsymbol{\gamma}$.

If curve i comes from cluster k is written as $\Psi_{i,k}$, then conditioned on the event, $\Psi_{i,k}$, $\mathbf{Y_i}$ has a normal distribution which can be written as

$$\mathbf{Y_i}|\Psi_{i,k} \sim N(\mathbf{S}_i(\boldsymbol{\mu_0} + \boldsymbol{\Lambda\alpha_k}), \Sigma_i), \tag{4.9}$$

where $\Sigma_i = \sigma^2 I + \mathbf{S}_i \Gamma \mathbf{S}_i^T$. Writing the probability of the $\Psi_{i,k}$ as π_k then the set of model parameters that are required to be estimated in order to fit the functional clustering model are $\boldsymbol{\alpha_k}, \Gamma, \boldsymbol{\mu_0}, \boldsymbol{\Lambda}, \sigma^2$ and π_k where $k = 1, ..., G$. In terms of the cluster membership probabilities, it is required that $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{G} \pi_k = 1$. Furthermore there are two identifiability constraints which must be imposed on the parameters which define the cluster means, $\boldsymbol{\Lambda}$ and $\boldsymbol{\alpha}_k$, in order to ensure these parameters are not confounded. The constraints are that

$$\sum_k \boldsymbol{\alpha_k} = 0 \tag{4.10}$$

and

$$\boldsymbol{\Lambda}^T \mathbf{S}^T \Sigma^{-1} \mathbf{S} \boldsymbol{\Lambda} = I. \tag{4.11}$$

where $\mathbf{S}$ is the matrix of basis function values evaluated on a grid of time points encompassing the full range of the data and $\Sigma = \sigma^2 I + \mathbf{S}\Gamma\mathbf{S}^T$. James and Sugar (2003) state that the consequence of placing this constraint on the $\boldsymbol{\alpha}_k$'s is that $\mathbf{S}^T\boldsymbol{\mu}_0$ can be interpreted as the overall mean curve. The authors also comment that there are several possibilities of constraint that could be placed on $\boldsymbol{\Lambda}$, however, the reasons for using the one described in (Equation 4.11) is so that low dimensional graphical representations of the curves can be produced.

4.2.2 FCM Fitting

While James and Sugar (2003) describe the two possible ways of fitting the model outlined in Equation 4.8 (the classification likelihood approach and the mixture likelihood approach) only the mixture likelihood approach will be considered in this thesis. The reason for this is that in the mixture approach each individual is assigned a probability of originating from each cluster and so these probabilities can be used to provide a confidence in classification for each of the lakes. The complete distribution for the FCM can be written as $f(\mathbf{Y}, \mathbf{z}, \boldsymbol{\gamma})$, however, since

the cluster memberships, $\mathbf{z_i}$, and the site effects, $\boldsymbol{\gamma}_i$ are assumed to be independent, this can be re-written as

$$f(\mathbf{Y}, \mathbf{z}, \boldsymbol{\gamma}) = f(\mathbf{Y}|\mathbf{z}, \boldsymbol{\gamma}) f(\mathbf{z}) f(\boldsymbol{\gamma}).$$

It follows that the log-likelihood to be maximised will be the sum of the log-likelihood of each of the individual component distributions, $f(\mathbf{Y}|\mathbf{z}, \boldsymbol{\gamma})$, $f(\mathbf{z})$, and $f(\boldsymbol{\gamma})$. The log-likelihoods for each of these three distributions (including additive constants which can later be removed) are now shown,

1. The cluster memberships, $\mathbf{z}_i$'s are each distributed as multinomial(π_i), so the likelihood can be written as,

$$L_1 = \prod_{i=1}^{N} \prod_{k=1}^{G} \pi_k^{z_{ik}},$$

and the log-likelihood is then,

$$\ell_1 = \sum_{i=1}^{N} \sum_{k=1}^{G} \left\{ \binom{G}{z_{ik}} + z_{ik} \log(\pi_k) \right\}. \tag{4.12}$$

2. The site effects, $\boldsymbol{\gamma}_i$'s, are each distributed $N(0, \Gamma)$, so the likelihood can be written as,

$$L_2 = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\Gamma}} \exp\left(\frac{-(\gamma_i)^2}{2\Gamma} \right)$$

and the log likelihood is then,

$$\begin{aligned} \ell_2 &= \sum_{i=1}^{N} \left(-\frac{1}{2} [\log 2\pi + \log|\Gamma|] - \frac{{\gamma_i}^2}{2\Gamma} \right) \\ &= -\frac{1}{2} \sum_{i=1}^{N} \left(\log 2\pi + \log|\Gamma| - {\boldsymbol{\gamma}_i}^T \Gamma^{-1} \boldsymbol{\gamma}_i \right) \end{aligned} \tag{4.13}$$

3. The mixture likelihood for all clusters, $\mathbf{Y}_i|\boldsymbol{\gamma}_i, \mathbf{z}_i \sim N(\mathbf{S}_i(\boldsymbol{\mu}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_k + \boldsymbol{\gamma}_i), \sigma^2)$ which is conditional on the cluster memberships and the site effects, can be

written as

$$L_3 = \prod_{i=1}^{N}\sum_{k=1}^{G} z_{ik}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\exp\left\{\frac{-(Y_i - \mathbf{S_i}(\boldsymbol{\mu_0} + \boldsymbol{\Lambda\alpha_k} + \boldsymbol{\gamma_i}))^{\mathbf{2}}}{2\sigma^2}\right\}$$

and the log likelihood is then,

$$\ell_3 = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{G} z_{ik}\left([\log 2\pi + \log\sigma^2] - \frac{1}{\sigma^2} \times ||Y_i - \mathbf{S}_i(\boldsymbol{\mu}_0 + \boldsymbol{\Lambda\alpha}_k + \boldsymbol{\gamma}_i)||^2\right) \tag{4.14}$$

Treating the $\mathbf{z_i}$'s as missing data, the full FCM log-likelihood to be maximised via the EM algorithm, can then be written as the sum of the expressions (Equations 4.12), (4.13) and (4.14). Spline coefficients for each of the individuals are first estimated using least squares fits to the observed data from each lake. Clearly, if there are very few observations for each of the individual curves, the standard least squares approach can break down and no solution can be obtained. To account for this a ridge term can be used where a small value is added to each of the observed values (Hoerl and Kennard, 1970). Following the estimation of the curves by adding a ridge term, a K-means procedure is applied to the B-spline coefficients that define the curves. The initial cluster membership probabilities are then calculated by finding the proportion of curves that are within each cluster. Finally, after the FCM log-likelihood outlined above has been maximised, each site is allocated to the cluster which has the highest corresponding probability of cluster membership.

4.2.3 Multivariate Functional Clustering Model

A multivariate extension of the FCM is also suggested by James and Sugar (2003) where functional data for several determinands of interest can be used to form clusters. This is similar to standard clustering approaches where there are several determinands measured on each individual, and is particularly useful as there are often situations where there are several functional data determinands which clusters would ideally be based on. In multivariate functional clustering there are multiple curves for each individual; one curve corresponding to a function over time for each determinand.

Using the same notation as before, let there be N individuals with J determinands measured on each and let m_{ij} denote the number of observations for individual i and determinand j, where $i = 1, ..., N$ and $j = 1, ..., J$. This set up allows for there to be different numbers of observations for both different individuals and for the different determinands. It is a realistic possibility that not all individuals will be measured at common time points for all determinands. Following this, let $\mathbf{Y}_{ij}$ represent the vector of observations for individual i, and determinand j at time points $t_{ij,1}, ..., t_{ij,m_{ij}}$. James and Sugar (2003) then state that the functional clustering model can be generalised and can be written for multiple determinands. For a particular site i and determinand j the response function can be written as

$$\mathbf{Y}_{ij} = S_{ij}\boldsymbol{\eta}_{ij} + \boldsymbol{\epsilon}_{ij} \text{ where } \boldsymbol{\epsilon}_{ij} \sim N(0, I\sigma_{ij}^2), i = 1 \dots N, j = 1 \dots J$$

In this model formulation, the way in which the error variance $\boldsymbol{\epsilon}$ is defined allows a different error vector for each determinand. Similarly to the standard setting, the spline coefficients, $\boldsymbol{\eta}_{ij}$, can be written as the sum of cluster mean and a random individual effect as

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\mu}_{\mathbf{z}_{ij}} + \boldsymbol{\gamma}_{ij}$$

It is assumed $\boldsymbol{\gamma}_i = (\boldsymbol{\gamma}_{i1}, \dots, \boldsymbol{\gamma}_{iJ}) \sim N(0, \boldsymbol{\Omega})$ and each $\boldsymbol{\gamma}_{ij} \sim N(0, \Gamma_j)$. For each cluster the cluster mean effect is formed from a concatenated vector of the cluster means for each determinand as can be parameterised using the same structure as in Equation 4.7. Hence, for any cluster k, the cluster mean effect $\boldsymbol{\mu}_k$ can be written as

$$\boldsymbol{\mu}_k = (\boldsymbol{\mu}_{k1}, ..., \boldsymbol{\mu}_{kJ}) = \boldsymbol{\mu}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_k$$

As with the single determinand model, one of the key reasons for this further parameterisation is so that data from each individual can be projected onto low dimensional space to enable any clustering to be visually identified from plots of these projected values. Using the multivariate extension of this model simply means that the projected data points for each individual will be based on the estimated curves from a set of determinands, rather than just one. Combining the above equations for all determinands, the functional clustering model for an

individual i can now be expressed as

$$\mathbf{Y_i} = \mathbf{S_i}(\boldsymbol{\mu_0} + \boldsymbol{\Lambda\alpha_k} + \boldsymbol{\gamma_i}) + \boldsymbol{\epsilon_i}$$
$$\text{where } \boldsymbol{\epsilon}_i \sim N(0, \mathcal{R}) \text{ and } \boldsymbol{\gamma}_i \sim N(0, \boldsymbol{\Omega})$$

Above, the data vector for each lake is formed from the corresponding data for each of the variables, $\mathbf{Y_i} = (\mathbf{Y_{i1}}, ..., \mathbf{Y_{iJ}})$. Assuming P spline basis functions are used to represent the functions corresponding to a single determinand, $\boldsymbol{\Lambda}$ is a matrix of size $JP \times h$, $\boldsymbol{\mu}_0$ is a vector of length JP while $\boldsymbol{\alpha}_k$ is a vector of length h.

The error vectors can be written as $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}_{i1}, ..., \boldsymbol{\epsilon}_{iJ})$ and the error covariance for site i, can be denoted by $\mathcal{R}$, and is a block diagonal matrix constructed from $I\sigma_1, ..., I\sigma_J$. The individual effect variance covariance matrix, $\boldsymbol{\Omega}$ and the spline basis matrix for the i^{th} individual, $\mathbf{S}_i$, are also both assumed to have block diagonal structures. The spline basis matrix $\mathbf{S}_i$ constructed from $\mathbf{S}_{i1}, ..., \mathbf{S}_{iJ}$. Each block $\mathbf{S}_{ij}$ is the spline basis matrix for the i^{th} and j^{th} variable curve evaluated at time points $t_{ij,1}, ..., t_{ij,m_{ij}}$.

4.3 Model Selection

One of the main difficulties associated with cluster analysis is identifying how many clusters are most appropriate given the data. Several approaches have been proposed to address this question, and common examples of earlier approaches, which can be applied to both non-probabilistic and model based clustering are provided in Calinski and Harabasz (1974),Hartigan (1975) and Krzanowski and Lai (1988). Calinski and Harabasz (1974) suggest a criterion based on the ratio of within and between cluster variation while the authors of the latter examples propose methods which are based on criteria which involve within-cluster sums of squares. Techniques such as these are summarised and reviewed in Milligan and Cooper (1985). One benefit of using model based clustering techniques is that model selection criteria such as AIC and BIC can often be used to determine the best model, and hence, the number of clusters which is most appropriate given the data. For example, after calculating BIC (Equation 1.23) for models with different numbers of mixture components corresponding to different numbers of clusters, the model which minimises BIC is selected. One of the drawbacks of using BIC is that for each potential number of clusters, and therefore each potential model, that is

to be compared, the maxmimum likelihood has to be calculated. With mixture models such as the FCM, maximum likelihood is obtained using the EM algorithm and so while a single model can be fitted relatively quickly, it is computationally expensive to fit the model repeatedly.

In view of the computational intensity of using some of the model selection criterion approaches to choosing the statistically optimal number of clusters, James and Sugar (2003) discuss their own approach for assessing the number of clusters when partitioning functional data. The authors suggest using the 'distortion function' which can be defined as

$$d_K = \frac{1}{P}\min_{c_1,\ldots,c_k} \mathrm{E}[(\boldsymbol{\eta}_i - \mathbf{c_{z_i}})^{\mathbf{T}}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\eta_i} - \mathbf{c_{z_i}})] \tag{4.15}$$

where $\boldsymbol{\eta}_i$ is the estimated spline basis coefficients for site i, $\mathbf{c_{z_i}}$ is the closest cluster centre to site i, P is the number of spline basis functions and Γ is the between-site covariance matrix. This expression is equivalent to the average mahalanobis distance between each curve and its nearest cluster centre. Sugar and James (2003) propose that substituting the identity matrix in place of Γ produces reasonable results, in which case Equation 4.15 becomes the average squared euclidean distance between each curve and its closest cluster centre. For each K, d_K is calculated using the estimated spline coefficients obtained using least squares and the observed data. Cluster membership and the resulting cluster centres are obtained by applying k-means to these spline coefficients. Following this, Sugar and James (2003) state that, assuming the distribution of the $\boldsymbol{\eta}_i$'s is a mixture of G P-dimensional clusters, and that the clusters are identically distributed with covariance Γ and finite fourth moments in each dimension, then under suitable conditions, there exists a set of real valued numbers $\mathcal{Y} > 0$ such that the jump defined as

$$\mathrm{jump}_k = d_k^{-\mathcal{Y}} - d_{k-1}^{-\mathcal{Y}} \tag{4.16}$$

will be maximised when k=G. There is no exact way of specifying what choice of $\mathcal{Y}$ is optimal, however, it is proposed that a suitable choice is to select $\mathcal{Y}$ so it is equal to one half of the effective number of dimensions in the data. For the functional clustering model there are three parameters which have to be optimised; the number of clusters, G, the value for parameterisation of the cluster means, h, and the number of spline coefficients P. The most crucial of these is clearly the choice of the number of clusters. James and Sugar (2003) suggest using the jump

method for first selecting the number of clusters and subsequently using BIC to determine the values of h and P.

Another popular approach for selecting the number of clusters is the gap statistic proposed by Tibshirani et al. (2001) which compares the change in within-cluster dispersion between the observed data and a null reference distribution that is generated using the observed data. A brief description of how the gap statistic method is applied will now be given. The first step in this approach is to apply the clustering method chosen to the observed data. Suppose the data are written as x_{ij} where $i = 1, ..., n$ are independent observations and $j = 1, ..., p$ is the number of features measured on each. If the data are then split into K clusters, then let $\mathcal{C}_k$ denote the indices of individuals in cluster k, and n_k be the number of individuals in cluster k. Following this, the within-cluster dispersion for cluster k can be defined as the sum of squared distances between all pairs of points which fall in this particular cluster. This can be written as,

$$D_k = \sum_{i,i' \in \mathcal{C}_k} d_{i,i'}^2 \tag{4.17}$$

While there are several potential choices for how distance between two points can be defined, Euclidean distance is probably the most simple and popular choice. Using Euclidean distance, for a fixed value of k, the within-cluster homogeneity can be measured by calculating the within-cluster sum of squares defined as

$$W_K = \sum_{k=1}^{K} \frac{1}{2n_k} D_k \tag{4.18}$$

Although this is the initial step in calculating the gap statistic, plotting W_K versus K has itself been used to determine the number of clusters. It is clear that as the number of clusters increases, W_K will decrease monotonically, however, the value of K at which W_K begins to "flatten markedly" indicates the number of clusters where there has been the largest increase in goodness of fit. Using this curve to choose the value of K is called the 'elbow' or the 'L-curve' method and while it provides a straightforward approach to determining the number of clusters, Tibshirani et al. (2001) highlight some of its deficiencies which the gap statistic attempts to overcome. The key drawbacks of the 'L-curve' approach include that there is no reference distribution with which to compare the W_K versus K curve, and that the differences between W_K for different values of K are not normalised and so it

is unreasonable to compare them. For this reason, to calculate the gap statistic, a reference distribution is needed which assumes there is no clustering structure in the data. A reference distribution is created either by generating data from a uniform distribution with limits chosen using the observed data. Alternatively, a principal components method which also uses a uniform distribution to generate reference data is suggested as this also takes into account the shape of the data and makes the procedure rotationally invariant. Clear details of the principal components approach for generating the reference distribution are given in Tibshirani et al. (2001).

Using one of these approaches a number of reference data sets, say B, are calculated and for each, the same clustering techniques that were applied to the observed data are used. For each reference set and each potential number of clusters, the within-cluster dispersion W^*_{Kb} is calculated where K is the number of clusters and $b = 1, \ldots, B$ represents the index of the reference set. The gap statistic can then be defined as

$$\text{Gap}(K) = \frac{1}{B}\sum_{b}^{B} \log(W^*_{Kb}) - \log(W_K)$$

Logs are taken in order to normalize the within-cluster dispersions being compared. The gap statistic for any given number of clusters is the difference between the average within-cluster dispersion from the B reference sets and the observed within-cluster dispersion. The largest gap corresponds to the number of clusters where there is the biggest gap between the within-cluster homogeneity of the observed data, which is assumed to have a clustered structure, and the reference data, which is assumed to have no clustering. However, in order to account for simulation error in the B reference data-sets, a tolerance of one standard error is used. Therefore, the estimated value of K, $\hat{K}$, is chosen using the rule

$$\hat{K} = \text{ smallest } K \text{ such that } \text{Gap}(K) \geq \text{Gap}(K+1) - (sd_{K+1}\sqrt{(1+1/B)})$$

where for each K, sd_K is the standard error of the B reference data sets. The standard deviation, sd_K, of the B reference data sets is

$$sd_K = \left[\frac{1}{B}\sum_{b}^{B}\{\log(W^*_{KB}) - \hat{l}_K\}^2\right]^{1/2}$$

where $\hat{l}_K = (1/B)\sum_b^B \log(W^*_{KB})$ is the average within-cluster dispersion of the reference data for each number of clusters k.

A more recent development in choosing the number of clusters based on the gap statistic approach is provided in Yan and Ye (2007) where a weighted gap statistic is suggested. Using the same notation as before, the sum of pairwise distances in (Equation 4.17) can be modified to be $\overline{D_k} = D_k/(2n_k(n_k-1))$ and the weighted within parameter dispersion can be defined as

$$\overline{W_K} = \sum_{k=1}^{K} \frac{D_k}{2n_k(n_k-1)}$$

Again if Euclidean distance is used, while D_k is the average squared distance between all pairs of points in cluster k, $\overline{D_k}$ is the average squared distance between observations in cluster k and the cluster mean. Yan and Ye (2007) state that it is easy to show that modified within-cluster dispersion $\overline{D_k}$ is an unbiased estimate of the population variance associated with cluster k. Consequently, the reason for weighting in this way is that $\overline{W_K}$ is thought to be more robust than W_K in terms of measuring the within-cluster homogeneity while taking into account the variations in the observed data and the reference data. However, there needs to be a sufficient number of observations within each of the clusters so that the within-cluster variation can be reasonably estimated. The weighted gap statistic can be written as $\overline{\text{Gap}(K)}$ and is applied in exactly the same way as the standard gap statistic approach with W_K replaced by $\overline{W_K}$.

In addition to the weighted gap statistic, Yan and Ye (2007) also suggest an alternative method to selecting the K that is most appropriate. While Tibshirani et al. (2001) use the "one standard error approach", Yan and Ye (2007) suggest that the standard gap statistic has a tendency to overestimate the number of clusters required in some situations and so propose a stopping rule based on the differences between successive gaps. The rule is called the DD-weighted gap method and aims to choose the estimated K, $\hat{K}$, in such a way that $\hat{K}$ provides a better fit than $\hat{K}-1$ clusters, but an additional cluster provides little extra benefit. It is suggested that a two step approach is taken to estimating the true number of clusters with the first stage being that the one standard error approach is used to test the null hypothesis that there is any clustering structure in the data at all (by comparing $\overline{\text{Gap}(1)}$ to $\overline{\text{Gap}(2)} - s_2$) and subsequently applying the stopping rule by choosing

K which maximises

$$DD\overline{\text{Gap}_n}(K) = D\overline{\text{Gap}_n}(K) - D\overline{\text{Gap}_n}(K+1)$$

where $D\overline{\text{Gap}_n}(K) = \overline{\text{Gap}_n}(K) - \overline{\text{Gap}_n}(K-1)$. Further details of this stopping rule approach are provided in Yan and Ye (2007).

In order to apply the gap statistic and the weighted gap statistic method to functional data, estimated spline coefficients for each curve can be regarded as being representative of observed data. For each of the n individuals, rather than having data on a set of different determinands for each individual, as would be the case with standard multivariate data, for functional data there will be a set of basis coefficients for each individual. Another consideration is that because there are potentially K clusters for the observed data, and B reference distributions simulated, the clustering procedure will need to be carried out $(B+1) \times K$ times. This will be time consuming if the FCM is fitted each time via the EM algorithm and so, as with the jump statistic approach, a reasonable alternative for using this method with the FCM would be to use the initial k-means algorithm to define the clusters each time. Further to this, problems may arise in generating the reference distribution using the principal components approach as singular value decomposition of the observed data matrix is required for this. Assuming as before that the data are sparse or irregularly sampled, then the data matrix will be incomplete and hence any missing data will need to be filled in. This can be done, as described in Section 3.3, by fitting natural cubic interpolating splines to the data for each individual, evaluating these functions at any points where data are missing data and then imputing these values.

4.4 Application of Hierarchical Functional Clustering to Lake Data

Hierarchical clustering, as outlined in Section 4.1, was applied to the Scottish lake data. As an initial step, this method has been applied to the log transformed data separately for each of the three chemical determinands of interest. This will result in a different group structure being obtained for alkalinity, phosphorus and chlorophyll separately. While a different group structure for each determinand is

not ideally what is desirable, since final classification of the lakes according to the WFD is based on a range of different determinands, it provides a starting point in enabling us to construct a statistically based group structure as an alternative to the current SEPA group structure which has not been based on observed data. A multivariate functional clustering approach will be investigated later in Section 4.5.

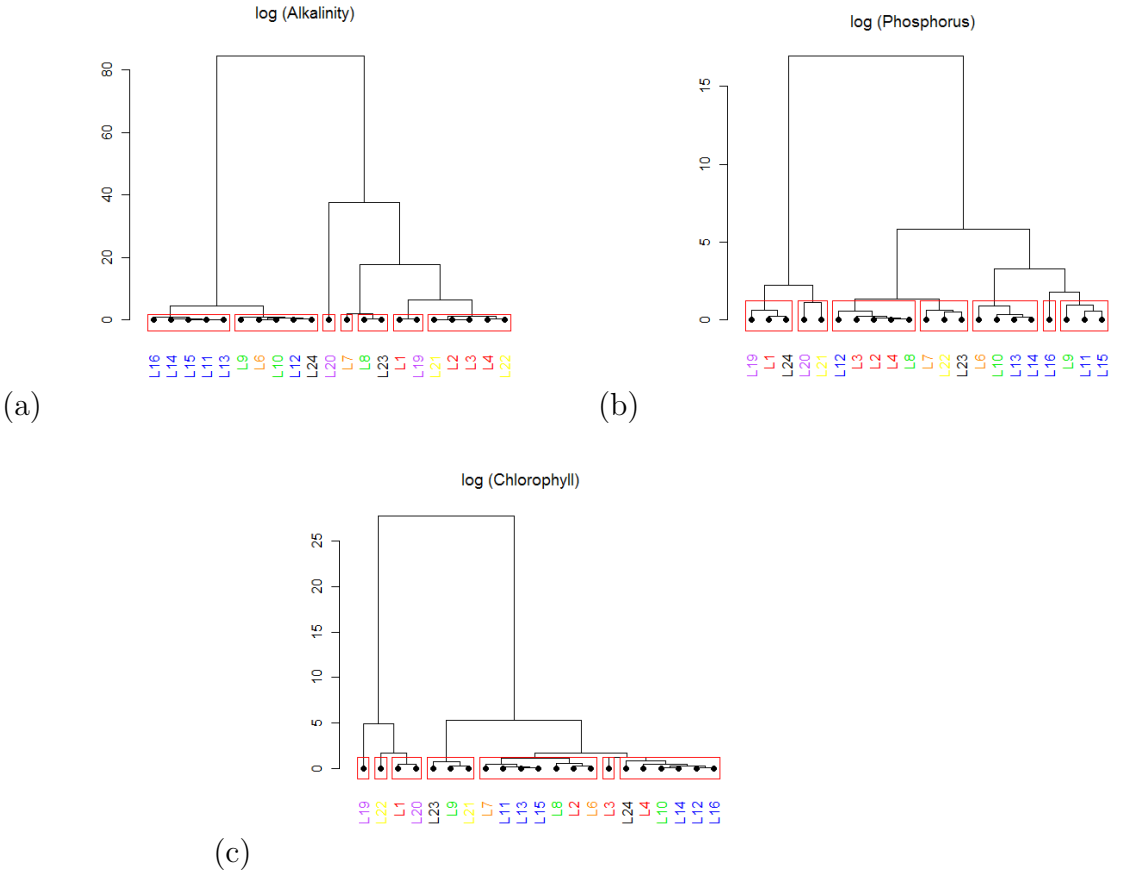


FIGURE 4.1: Dendrograms showing results of Hierarchical Functional Clustering for Scottish lakes data each cut to indicate seven groups.

To implement the hierarchical clustering procedure to the data the first step was to obtain estimated spline coefficients for the functions at each lake. For each of the three determinands, smooth functions were fitted to the log transformed data at each lake using the same procedure as that described in Section 3.3. Next a distance matrix was calculated using the estimated spline coefficients and the distance metric in Equation 4.1. Complete linkage was then used in conjunction with the distance matrix in order to produce a hierarchical clustering structure for each determinand. Alternative distance metrics were also explored, including single linkage and average linkage, however it was found that for all of the metrics

investigated there was a considerable degree of overlap between the hierarchical clusters produced and the original groups currently used by SEPA. Due to the chaining nature of its steps, single linkage is often used to identify observations which are outliers rather than a set of clusters. For the Scottish lakes data single linkage identified Lake 20 as being distinct from the other lakes in terms of alkalinity, while Lake 19 was identified as being distinct in terms of Chlorophyll.

For each of the log transformed determinands a dendrogram which summarises the clustering structure was produced. These dendrograms are shown in Figure 4.1. On each dendrogram, each node represents a single lake with the colours of the nodes corresponding to the original SEPA groups as shown in Figure 3.1 and the numbering being consistent with the lake numbers assigned to each lake in Table 3.1. Panel (a) corresponds to alkalinity, (b) correspond to phosphorus and (c) corresponds to chlorophyll. The dendrograms in Figure 4.1 are cut to show seven different groupings to enable comparison of the existing SEPA groups with the partitions determined using the hierarchical clustering. For alkalinity, from Figure 4.1 (a), it can be seen that there is some agreement between the groups specified using hierarchical functional clustering, and the groups which are currently used by SEPA. For example, five of the six lakes which are currently in SEPA group 4 (shown in blue) would continue to be in the same group using this new structure. Similarly, three of the four lakes which are currently in SEPA group 1 (red) and both of the SEPA group 6 (yellow) lakes are grouped together using hierarchical clustering. One notable lake is Lake 20, Harray Loch, which is currently in SEPA group 5. It was found that if either 3,4,5,6 or 7 groups are specified for alkalinity, Harray Loch is always identified as being separate from the rest. In the initial functional data analysis of the alkalinity data, the functions corresponding to the group 5 lakes (lakes 19 and 20) displayed a markedly higher mean level than the rest of the groups (see Figures 3.5(a) and 3.6(a)). The higher concentration at Lake 20 may be explained by the geographical location of this site on Orkney and so the surrounding environment may be quite different to that on the mainland. This lake is one of only two which are not situated on the Scottish mainland.

As before, the reasonable agreement between the SEPA groups and those identified using the functional hierarchical approach for alkalinity could be expected as although no observed chemical data are used for the current groups, they are partly based on broad categories of Alkalinity. If we look at the scale on the dendrograms for each determinand, which show functional distance, it can be seen

that the range of distances for alkalinity is far greater than that of phosphorus and chlorophyll. This is further evidence that there is more of a group structure in alkalinity than the other two determinands. Phosphorus and chlorophyll have no influence on the formation of the current SEPA groups and when comparing the 7 groups determined using hierarchical clustering for these determinands to the SEPA groups it can be seen there are more differences between the new and old group definitions. For Chlorophyll there is also one lake (Lake 19) which is identified as being different from the others if 4,5,6 or 7 groups are specified. As with Loch Harray, Lake 19, Loch Eye, is currently in SEPA group 5 and again appears to be distinct from the rest of the lakes due to a higher mean level. The geographical location of the lakes could again provide a potential reason for this difference. Unlike any of the other locations considered Loch Eye is situated on the North East coast of Scotland.

In terms of investigating the statistically optimal number of clusters for hierarchical clustering, the gap statistic was used and a range of different numbers of clusters were considered. The maximum number of groups which was considered in the comparisons was 10 so that a large range of possible values for number of groups could be explored. It is acknowledged that 10 is a fairly large number given that there are only 21 lakes, and given that the exploratory analysis of the functional data indicated that even the 7 groups currently used by SEPA seemed to be too many. However, investigating up to 10 groups covers a large range of potential group structures and allows there to be scope for lakes to be in a 'group' on their own. This is especially important as there appeared to be a few lakes which seemed to be different from the rest in the exploratory analysis of the curves.

Since the principal components method of simulating reference distributions requires the singular value decomposition of the observed data matrix, a complete set of data with no missing values was needed. In order to obtain a complete set of observed data interpolating splines were fitted to the data before estimates of the missing values were obtained by evaluating these functions. Using the principal components approach and the complete data matrix, 500 sets of reference data were generated. The hierarchical clustering procedure was applied to each reference distribution and the dendrogram was cut at a range of different numbers of clusters. The within-cluster sum of squares was next computed. Then, to form the expected reference distribution curve, for each number of clusters, the mean and standard deviation of the 500 reference distribution based sums of squares was

calculated. Although neither Tibshirani et al. (2001) nor Yan and Ye (2007) give a clear indication of how many reference data-sets should be used, 500 seemed to be sufficient as it produced results which were consistent when the calculations were repeated.

Figure 4.2 displays plots of the within-cluster dispersion against the number of clusters for each of the three determinands (panels (a), (c), and (e)), and the corresponding plots of the gap statistic against the number of clusters (panels (b), (d) and (f)). Before even calculating the gap statistic, using just the L-curves, which suggests that the correct number of clusters corresponds to the point where the curve markedly flattens out, indicates that for alkalinity 3 groups is most appropriate. For the other two variables the L-curves are less clear about the number of clusters that is is optimal. Figure 4.2(b), (d) and (f) display the gap statistic with bars representing one standard error. The red line segments on these plots can be used to help visually identify the optimal number of clusters suggested by the gap statistic; the number of clusters selected corresponds to the first instance where the gap is greater than the immediately succeeding gap minus one standard error (where the red line has a negative gradient). It is evident that the gap statistic method suggests 3 groups is best for alkalinity and chlorophyll, while 2 groups is best for representing the phosphorus curves.

Figure 4.3 shows a set of 3 dendrograms corresponding to each of the three determinands. These are the same dendrograms as shown in Figure 4.1 cut to depict the statistically optimal number of groups as determined by the gap statistic. Table 4.1 shows the original SEPA groups, the groups formed using hierarchical functional clustering with the gap statistic, and a set of clusters made up of the cross product of the three individual determinand clusterings.

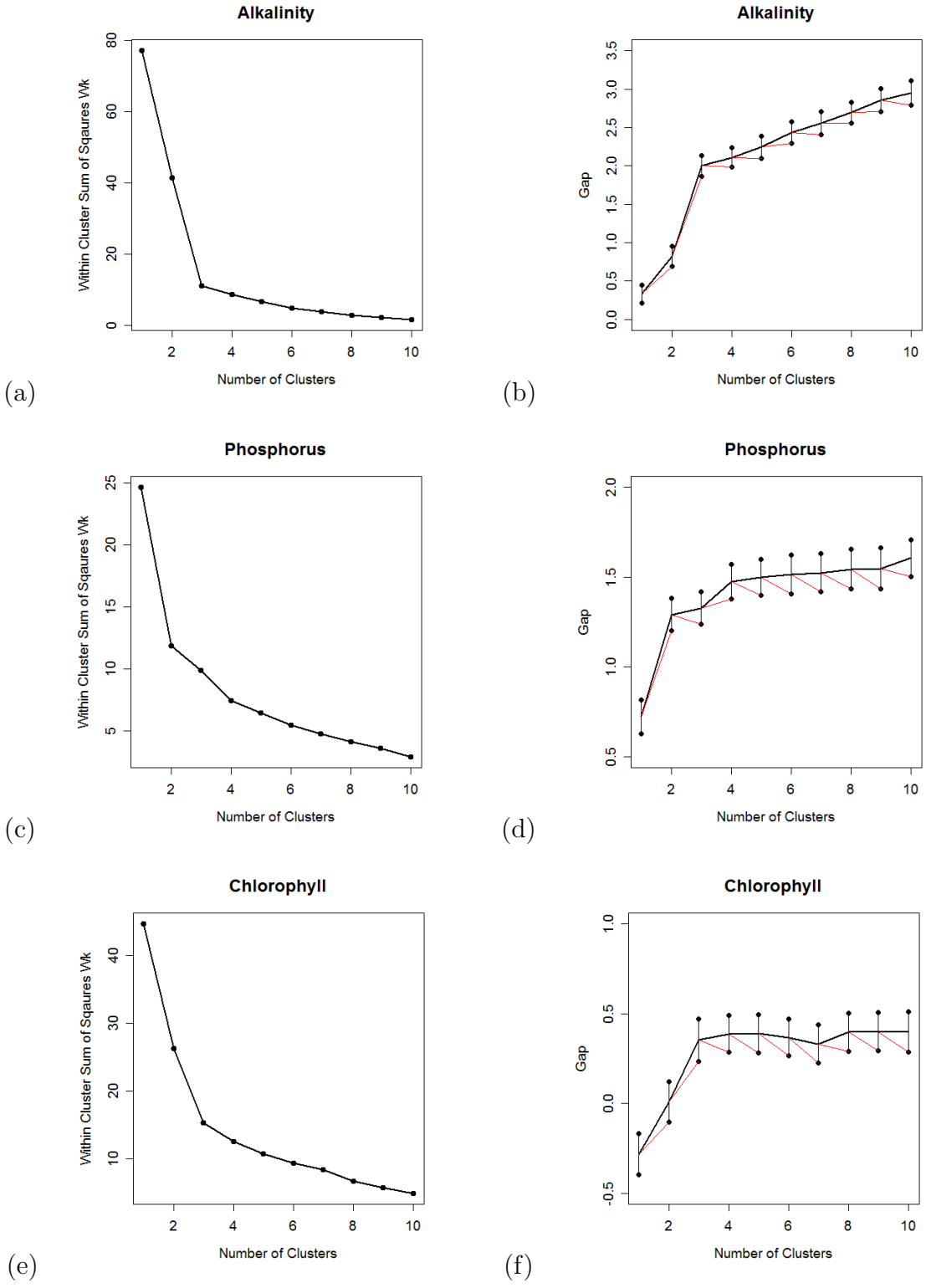
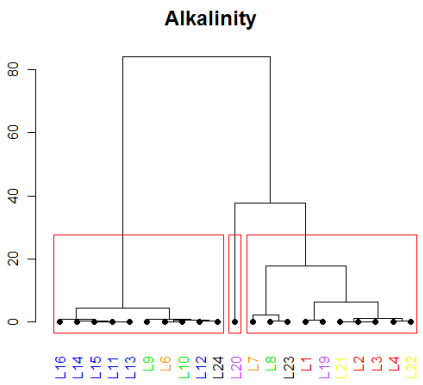
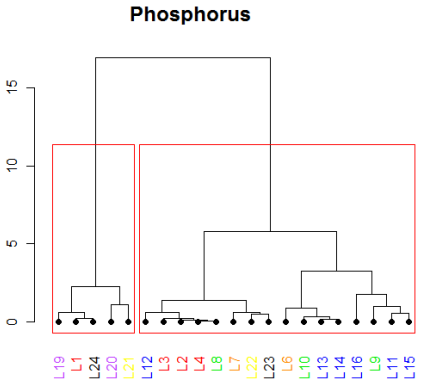


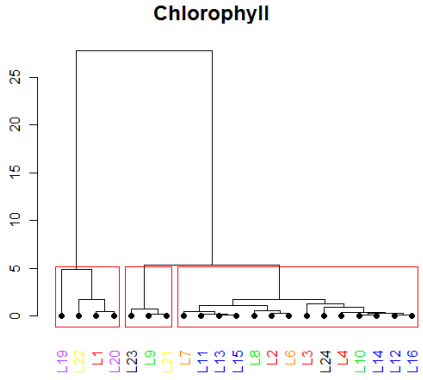
FIGURE 4.2: L-curves and gap statistic plots corresponding to hierarchical functional clustering for Scottish lakes data. Panels (a) and (b) correspond to log(alkalinity), (c) and (d) to log(phosphorus), and (e) and (f) to log(chlorophyll)



(a)



(b)



(c)

FIGURE 4.3: Dendrograms showing results of Hierarchical Functional Clustering for Scottish Lakes data cut to show the statistically optimal number of groups as determined by the gap statistic.

Lake	**Name**	**Alk** G=3	**Phos** G=2	**Chl** G=3	**SEPA group**	**cross prod**
1	Gladhouse Reservoir	3	1	1	1	A
2	Talla Reservoir	3	2	3	1	B
3	Fruid Reservoir	3	2	3	1	B
4	St Marys Loch	3	2	3	1	B
6	Loch Katrine	1	2	3	2	C
7	Glen Finglas Reservoir	3	2	3	2	B
8	Loch Avich	3	2	3	3	B
9	Loch Ba	1	2	2	3	D
10	Loch Arkaig	1	2	3	3	C
11	Loch Beinn a Mheadhoin	1	2	3	4	C
12	Loch Mhor	1	2	3	4	C
13	Loch Mullardoch	1	2	3	4	C
14	Loch Monar	1	2	3	4	C
15	Loch Glascarnoch	1	2	3	4	C
16	Loch Quoich	1	2	3	4	C
19	Loch Eye	3	1	1	5	A
20	Harray Loch	2	1	1	5	E
21	Loch Tralaig	3	2	2	6	F
22	Loch of Cliff	3	2	1	6	G
23	Lussa Loch	3	1	2	7	H
24	Loch Glashan	1	1	3	7	I

TABLE 4.1: Table of groups based on hierarchical functional clustering

For alkalinity there are two larger groups, each consisting of 10 lakes, and a single lake which forms a group. The lake identified as forming a group by itself is again Harray Loch (Lake 20) which has been discussed before as being distinct from the other lakes in terms of the mean alkalinity concentration observed. One of the larger alkalinity groups is primarily comprised of lakes which form SEPA groups 3 and 4, while the other contains lakes from SEPA groups 1 and 6. The phosphorus groups comprise of one smaller group containing 5 lakes, and one group of the remaining 16 lakes. All of the SEPA group 2, 3 and 4 and 6 lakes, and 3 of the 4 group 1 lakes are all contained within the larger group. For chlorophyll, the structure of the largest group, which contains 14 lakes, is similar to that of the larger phosphorus groups. It is clear that even with the smaller number of groups, for all determinands, much of the existing SEPA group structure has been preserved within the new groups.

The cross-product clusterings indicate that there are nine distinct combinations of classifications according to the individual determinands. Five lakes have

classifications which are unique whist there are 3 groups of more than one lake; one consisting of 2 lakes, one made up of 5 lakes and a larger group consisting of 8 lakes. All six SEPA group 4 lakes fall within the largest group of cross-product classifications. The five lakes which have a unique cross-product classification are situated close to the coast and so potentially weather or land use information may explain why these lakes are distinct from the others. While hierarchical clustering provides a good first step into investigation of the lakes group structure, a more formal approach was next taken to look at alternative groupings via the application of the functional clustering model.

4.5 Application of Model Based Functional Clustering

The functional clustering model (FCM) was applied to the Scottish Lake Data. Unlike the hierarchical clustering approach, there is no requirement to have a complete set of data which is common across all lakes for the FCM and hence only the observed log transformed data were used. As with the hierarchical approach, the first step was to fit models separately to each of the three chemical determinands of interest alkalinity, phosphorus and chlorophyll, using `R` code which accompanies James and Sugar (2003).

Within the model fitting procedure there are several parameters which need to be specified. These include the ridge regression parameter. James and Sugar (2003) suggest using a small value for the ridge parameter of around 0.001 rather than zero as the least squares estimation of the spline coefficients can break down if the dataset is incomplete. This value seemed to work well for the Scottish lake data and so in all FCMs presented in this thesis, the ridge parameter was set as 0.001. In addition there are a further three parameters for each of the FCMs which can be optimised. Obviously the parameter of most interest from these is the number of clusters, G, although decisions also need to be made concerning P, the number of spline functions used to represent the data as well as h, the number which should be used within the parameterization of the cluster means. James and Sugar (2003) suggest using different approaches to estimate these values; the jump method to select what number of clusters is appropriate, and then BIC to choose P and h. Pastres et al. (2011) who apply the FCM to water quality data

in Venice Lagoon use only BIC to select the parameters used. The selection of parameters for the Scottish lakes data will now be discussed.

Choosing P, G and h for the Scottish Lakes Data

As BIC is based on the likelihood of the fitted model, while the jump statistic is based on the k-means clustering of the data, BIC was at first thought to be the most accurate option and was initially used to optimise parameter values. Ideally BIC values for all possible combinations of P, G and h can be calculated simultaneously however this proved to be highly computationally expensive. The BIC was calculated for several models for values of G ranging from 1 (no grouping) to 10 while h values were considered from 1 to 9 (since $h \leq min(P, G-1)$ and for these data $P > G$). Considering this range of group numbers meant that for each each determinand, BIC values were calculated for 55 different models in order to optimise only h and G. Choosing the number of splines using BIC also would substantially increase this number of potential models yet again. It was felt very little would be gained by choosing the number of splines using BIC and so an alternative approach was used in order to avoid an excessive amount of computation. The number of spline functions was chosen from visual inspection of the functions estimated for each lake using the fitted FCM model. This is similar to the approach taken when selecting the choice of smoothing parameter when fitting spline functions to the data in Section 3.3 and requires the model to be fitted only once for each potential value of P.

There are several potential choices of P which all seem sensible, and as with the initial fitting of the functions to the observed data in Section 3.3, the aim was to choose the number of basis functions so that the main features of the data were captured without being too smooth. For each of the single determinand models, using 15 basis functions seemed to be an appropriate choice that balanced the responsiveness of the data and local variability. Although the number of basis functions used to estimate the functional data here is less than the number used to estimate the earlier functional data, the functions continue to be comparable in terms of the smoothness of the fit. There is also less data here since the dataset is irregular and missing values have not been imputed.

To illustrate the choice of 15 basis functions, Figure 4.4 shows the functions fitted to alkalinity data for selected lakes (1, 20 and 24) using both penalised

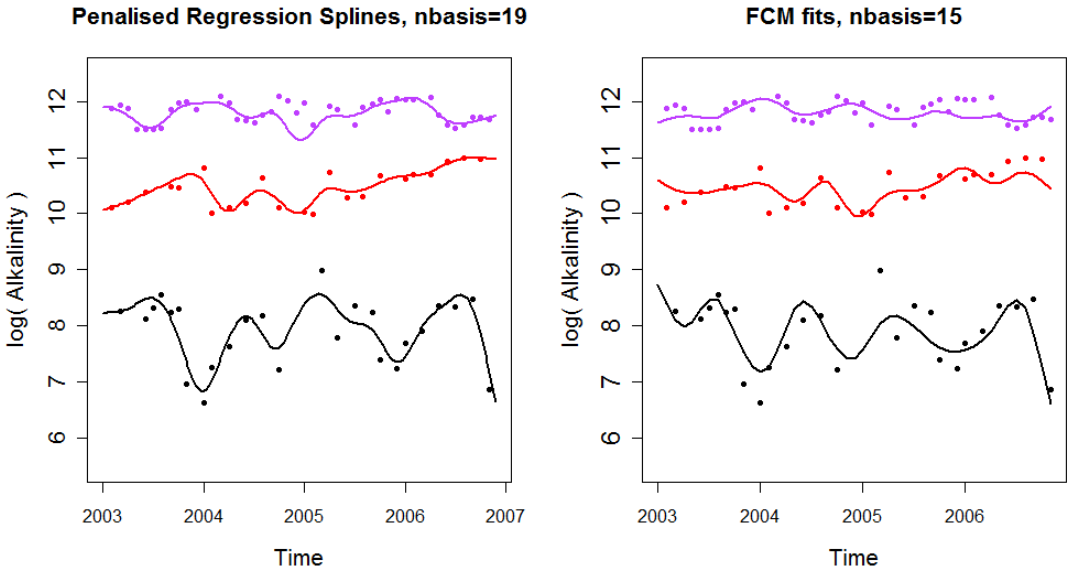


FIGURE 4.4: Comparison between fitted log(alkalinity) functions for penalised regression splines (left) and FCM (right) (lakes 1, 20, 24)

regression splines with 19 basis functions, and the functions fitted using the FCM model with 15 basis functions. The observed data are also shown on these plots. It can be clearly seen here, that while the two sets of curves are slightly different because of the two different methods of fitting (regression splines with a penalty and regression splines with the addition of a ridge term) they are very similar in terms of the smoothness of the curve. Both sets of curves appear to provide a good fit to the observed data without being excessively locally variable. It should be noted that the specific lakes shown here are purely illustrative and only these curves are shown to allow easier comparison across the two fitting methods.

Following the selection of the number of basis functions, Figure 4.5 displays plots of the BIC values for each of the determinands when the model was fitted using each combination of G and h, and with 15 set as the number of spline basis functions. The y-axis shows the BIC values, the x-axis the number of groups, G, and each line represents a different value of h. The BIC for each model was obtained using Equation 1.23 where the likelihood was calculated using (Equation 4.9) and n_p, the total number of parameters used in the calculation of BIC, was

given by

$$1 + P + (P \times h) + \frac{P \times (P+1)}{2} + hG - 1 + (G-1) \tag{4.19}$$

Table 4.2 contains a summary of the contribution of each of the parameters to be estimated to the total number of parameters, n_p. This takes into account the constraints imposed by the model on the group mean parameters, $\boldsymbol{\alpha}_k$, (detailed in Equation 4.10) and the probabilities π_k. Both $\boldsymbol{\Lambda}$ and Γ were matrices with unknown elements to be estimated and so the size of these matrices was included in our evaluation of n_p.

Parameter	**Contribution to n_p**
σ^2	1
$\boldsymbol{\mu}_0$	P
$\boldsymbol{\Lambda}$	$P \times h$
$\boldsymbol{\Gamma}$	$\frac{P(P+1)}{2}$
$\boldsymbol{\alpha}_k$	$h(G-1)$
π_k	$G-1$

TABLE 4.2: Summary of number of parameters (n_p) used in BIC calculations for the univariate models

For alkalinity, BIC is minimised when $h = 1$ and $G = 3$, for phosphorus when $h = 1$ and $G = 2$ and for chlorophyll when $h = 1$ and $G = 5$. While a minimum BIC value has been obtained for each determinand it can be seen from the plots that the minimum value appears to be only just smaller than surrounding values. A rule of thumb for comparing BIC values has been provided by Raftery (1995) who developed a set of rules for interpreting the difference between pairs of models. This rule of thumb states that a difference of between 0 and 2 BIC units is "weak" evidence of a difference, between 2 and 6 is "positive" evidence, between 6 and 10 is "strong" evidence and more than 10 units is "very strong". For alkalinity, the optimal model which has $h = 1$ and $G = 3$, has a BIC value of 817.07 while the closest model, in terms of BIC, has $h = 1$, $G = 2$ and has a value of 815.5. Going by this rule of thumb there is weak evidence that 3 groups is better than 2 groups for alkalinity. Similarly, for phosphorus, there is a relatively small difference between the BIC value for the 'optimal' model, where $G = 2$ and $h = 1$, and the model where $G = 3$ and $h = 1$ of around 5.5 units. Although there is some evidence for the model with 2 groups, it is again unclear this is a much better option than

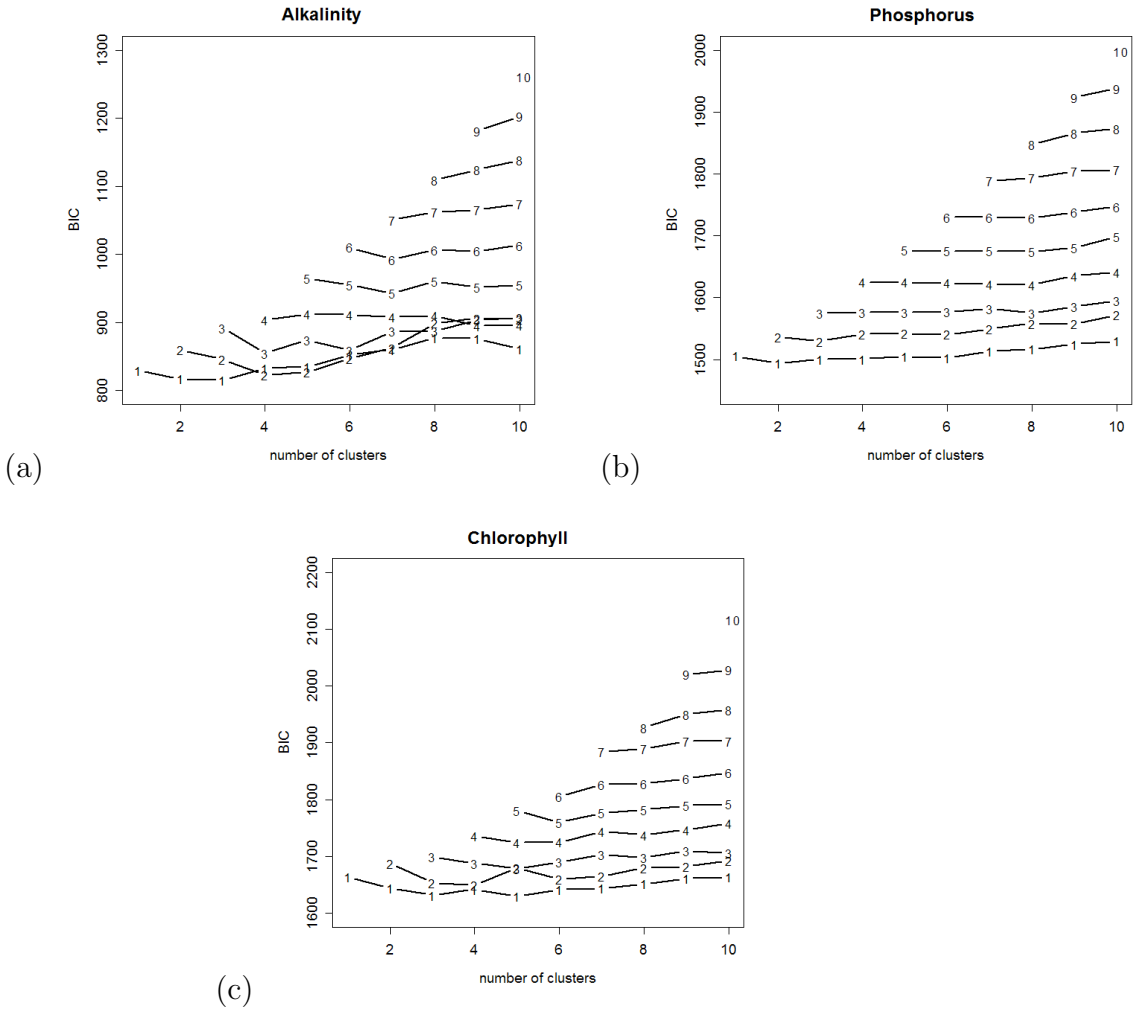


FIGURE 4.5: BIC for alkalinity, phosphorus and chlorophyll CM

the model with 3 groups. For chlorophyll, going with this rule of thumb the best choice is slightly more apparent and the model with the minimum BIC has a score more than 10 units smaller than any other model considered.

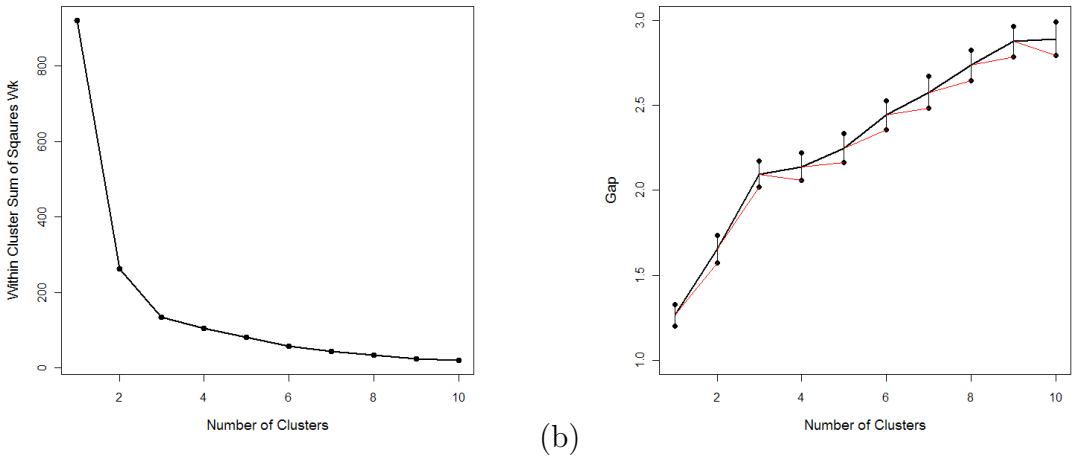
While selecting the number of groups using BIC utilises the full fitted model and is thought to be a reliable approach to choosing the number of clusters, for this particular dataset the results produced do not always indicate a single model as being markedly better than the rest. In light of this it seemed sensible to investigate alternative methods of choosing the statistically optimal number of clusters, namely the jump statistic and the gap statistic. For the jump statistic method the distortion function in (Equation 4.15) was calculated for each number of clusters ($G = 1$ to 10). The estimated spline coefficients for the curves used in this method were obtained from fitting the curves using regression splines plus a ridge term, as in the initial stage of fitting the FCM, and then the k-means algorithm was applied

to these to obtain estimates of cluster membership and hence cluster centres. As suggested by Sugar and James (2003) the k-means procedure was used rather than the full FCM in order to avoid unnecessary and excessive computation and the identity matrix was used rather than the lake covariance matrix, Γ as this matrix is only computed while fitting the full FCM. After calculating the distortion for each number of groups the jump statistic was next calculated and the optimal number of groups was indicated by where the jump statistic was maximised. It quickly became clear that the jump statistic method of estimating the most appropriate number of clusters was unreliable for the Scottish lakes data with the most concerning problem with this method being the high variability in the results obtained. After repeating the calculations multiple times using the same original data, it could be seen that the results produced, in terms of the number of clusters identified as being best, varied greatly each time. Furthermore, the jump statistic was also highly sensitive to the transformation power used. Although Sugar and James (2003) suggest that the transformation power should be equal to half the effective number of dimensions of the data, there continued to be uncertainty surrounding the choice of transformation power which was most appropriate for this datset.

Following this, the gap statistic was calculated for the same range of numbers of clusters for each determinand separately. As with the jump statistic, in order to avoid an unnecessary level of computation, k-means was used on estimated spline coefficients to provide the group structure. The functional distance between the curves, as used within the hierarchical approach, was computed to determine the differences between the observations, and the principal components method of generating a reference distribution was used to ensure the data generated to form the null distribution took into account the shape of the observed data. The principal components approach with the complete data matrix was used as before and 500 sets of reference data were generated. To each reference distribution the k-means clustering procedure was applied over a range of different number of clusters and the within-cluster sum of squares was computed.

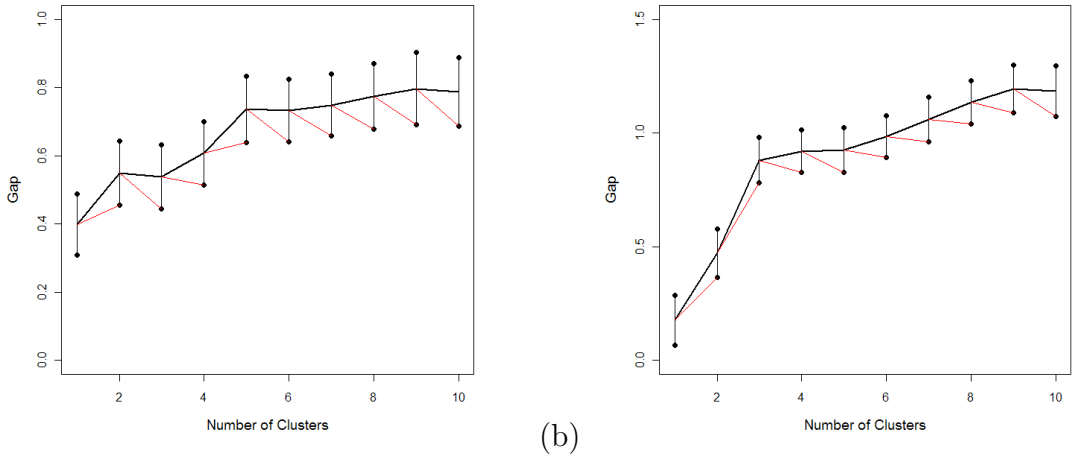
Figure 4.6(a) displays a plot of the within-cluster dispersion against the number of clusters for alkalinity and Figure 4.6(b) displays the gap curve with bars representing one standard error. It is evident from Figure 4.6(b) that the gap statistic method suggests three groups is the number that is statistically optimal for alkalinity. Repeating the same procedure for phosphorus indicated that 2

groups would be optimal (Figure 4.7 a) and for chlorophyll 3 groups was chosen as being the most suitable number (Figure 4.7 b). The results of the gap statistic approach for selecting the appropriate number of clusters for the FCM model agree both with those determined using BIC and with the number of clusters identified as suitable within the hierarchical clustering of the same data. While the number of groups selected for chlorophyll using BIC was 5, 3 groups was the next nearest in terms of BIC value. This may indicate that for chlorophyll there are 3 more separated groups or 5 groups which are not as distinct.



(a) (b)

FIGURE 4.6: L-curves plot (a) and Gap Statistic plot (b) for alkalinity



(a) (b)

FIGURE 4.7: Gap Statistic plots for phosphorus (a) and chlorophyll (b)

In terms of fitting the FCM, for the lakes data, the gap statistic provides a reliable way of selecting the number of clusters, however unlike BIC, it cannot be used to simultaneously answer the question of how to choose h, the value which

determines how the cluster means are parameterized. In the univariate case it is fairly clear from the BIC values calculated, that $h = 1$ seems to a sensible choice for all three single determinand models. For the later multivariate models however, h was not selected using BIC as it proved to be overly computationally intensive. An alternative method to choose h is by fitting models with the number of clusters identified using the gap statistic, and then look at a plot of the estimated curves projected onto h dimensional space. The primary purpose of the h parameterisation within the FCM is to allow low dimensional projections of the curves and it can be seen from the plots of the projected curves whether h is inappropriate. James and Sugar (2003) suggest that if the projected curves appear to lie in a lower dimensional space then the model should be re-fitted with h adjusted accordingly. For example, if $h = 2$ is used and the projected curves (plotted onto 2 dimensional space) lie approximately in a straight line then this implies that $h = 1$ would be more appropriate choice. The choice of which h is also determined by the restriction that $h \leq \min(P, G-1)$, meaning that if $G = 2$ is selected as best, as it is for the phosphorus data, then the only option is h=1.

Choosing G as determined by the gap statistic, $h = 1$ as identified using BIC and $P = 15$, an FCM was fitted to the log transformed alkalinity, phosphorus and chlorophyll data. All models were fitted using the EM algorithm and were said to converge when there was less than 0.1% change in the estimated error variance σ^2. Using this tolerance all of the univariate models converged quickly with 6 iterations. The stopping criterion used here is a lack of progress criterion rather than a convergence criterion such as Aitken's acceleration procedure. The results of these models are summarised in Table 4.3 which presents for each determinand a group allocation for each site and the estimated membership probability. For all three determinands the estimated cluster means highlight that the division between the groups is based solely on mean level. For this reason, for alkalinity and chlorophyll the three groups have been labelled high (H), intermediate (I) or low (L), and for phosphorus the labels high(H) and low(L) have been used. These labels are only intended to give a broad idea of the relative position of the cluster mean trajectories. Figures 4.8, 4.10 and 4.12 show four plots which present the estimated cluster structure for each determinand. The four figures within each set are:

(a) the observed data for each lake;

(b) the predicted trajectories for each lake;

(c) the value of each curve when projected onto 1 dimensional space against the area of the lake in km^2. The projected cluster centres are also shown as vertical lines;

(d) the predicted curves are shown again (dashed lines) with the estimated cluster means superimposed (solid lines).

In all of these plots different colours represent the different groups. It should be noted that there is no particular reason why the projected curves have been plotted against area of lake in (c), and while plotting the points in this way is primarily to scatter them enough that their distance from the group mean can be more clearly seen, it could also indicate if there was any relationship between the clusters and this determinand. Other determinands could be used here if they were specifically of interest.

Discussion of Univariate FCMs

For alkalinity the estimated cluster means appear to be fairly well separated. The low group mean displays the clearest evidence of a seasonal pattern of the three groups although this is not particularly strong and is heavily influenced by a few lakes where the seasonal signal is large. The groups vary in size, with the highest group (shown in red) consisting of only three lakes, one of which has a markedly higher mean alkalinity level than the other two. This is lake 20, Loch Harray, which has already been identified in the hierarchical cluster analysis as being particularly distinct. Both of the lakes in SEPA group 5 (lakes 19 and 20), are in the high group. This is consistent with the earlier functional analysis since group 5 was the one which was shown to be significantly different to all other SEPA groups when pairwise functional t-tests were implemented. The results of the univariate models are summarised in Table 4.3. The predicted cluster memberships are shown along with the corresponding confidence in classification. A cross-product classification has also been included.

The within-group variability seems fairly small in the group of 10 Low concentration lakes (shown in green) and in the group of eight Intermediate lakes (shown in black), with the exception of one Intermediate site which displays a much stronger seasonal component than all other group members. The projected

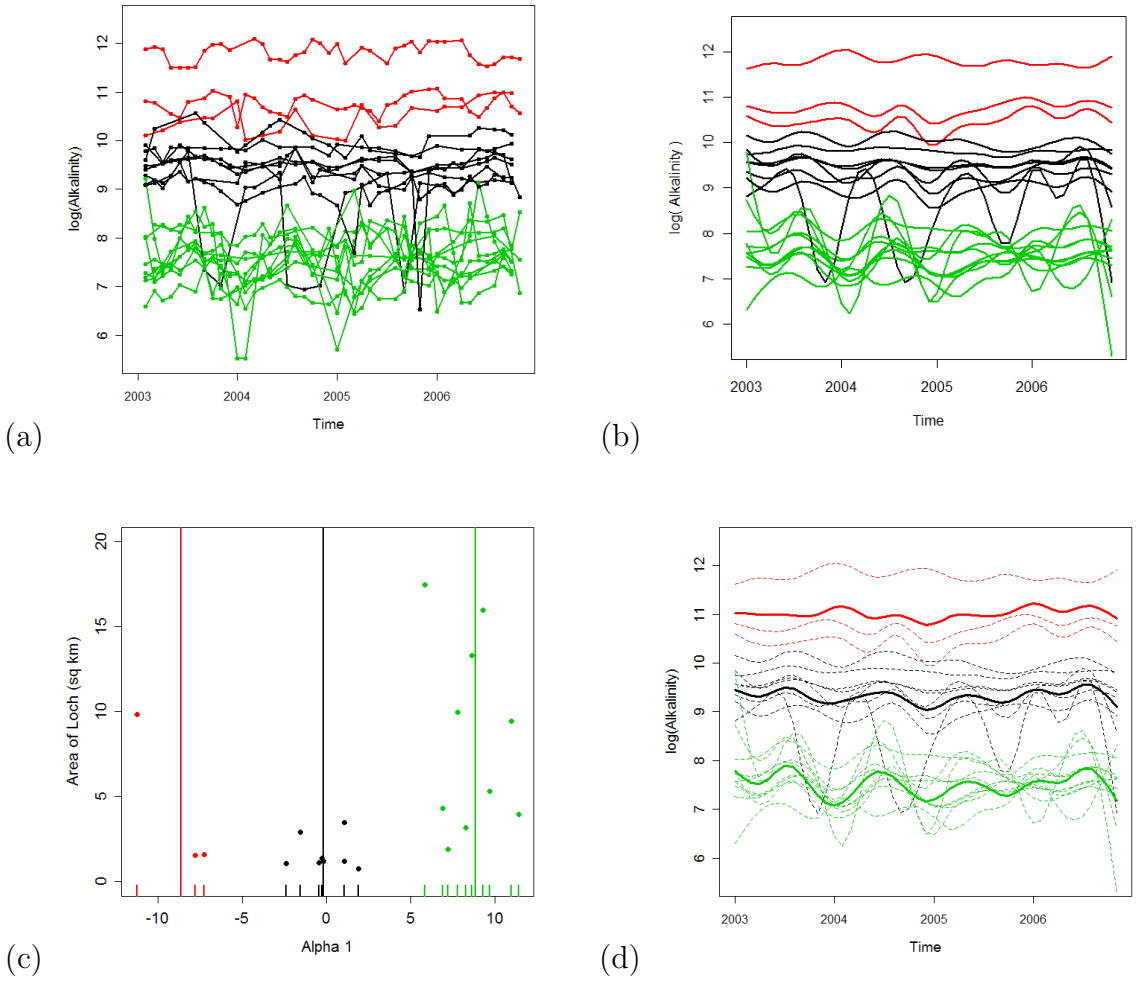


FIGURE 4.8: Summary of fitted FCM for alkalinity; (a) observed data, (b) predicted curves, (c) linear discriminant plot, (d) cluster means

curves and cluster means shown in Figure 4.8(c) provide further evidence of separation between the groups as the values are clustered in three sets along the x-axis. There is little evidence of a connection between the size of the lakes (in km^2) and the groups based on alkalinity although all Intermediate lakes tend to be relatively small, while the Low alkalinity lakes are much more varied in terms of size. The probability of cluster membership is 1 (Table 4.3) for all lakes for alkalinity. In view of the overlapping estimates for each of the curves representing different lakes this level of certainty in the partition may initially seem questionable, however, looking at Figure 4.8(c) it can be seen that each of the projected curve values are definitely closer to one group centre line than any other. There are no points on this plot which are equidistant from two of the projected group centres.

Figure 4.9 shows a map of Scotland with the new group structure for alkalinity.

The different groups are coloured as before, and the site numbers again correspond to those in Table 3.1. From this map there does appear to be a spatial pattern in the group structure as determined by the FCM in terms of alkalinity. The lakes in the high group are all along the East coast and those in the Low group tend to be located in the North West. The Intermediate groups are the most spread out geographically although they tend to be South of the other lakes with the exception of site 22, Loch of Cliff which is the furthest north of all the lakes. Although there are only three FCM groups for alkalinity, when comparing these new groups to the current SEPA ones it can be seen there is a reasonable agreement between the two. Both lakes in SEPA group 6, and 3 out of 4 of the SEPA group 1 lakes form the Intermediate group while all of the SEPA group 4 lakes are in the Low group. In addition, as already noted, SEPA group 5 are two of the three lakes which make up the High group in the new FCM structure. SEPA groups 2, 3 and 7 are split between the new groups. There was some indication in the initial analysis of the functional data that fewer groups, comprised from concatenation of the existing groups may be sufficient in capturing the variability between the lakes. Analysis of the results of the FCM for alkalinity confirm this to be the case.

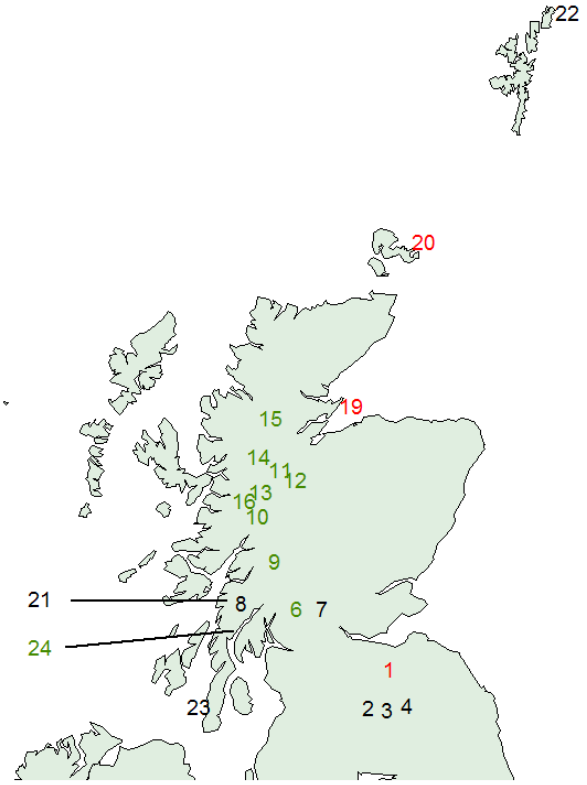


FIGURE 4.9: Map of Scotland showing FCM group structure for alkalinity

For phosphorus, the split between the two groups is again based primarily on mean level, with one smaller group consisting of 5 High concentration lakes and a larger Low group formed from the remaining 16 lakes. From Figure 4.10(c) there appears to be a reasonable amount of variability in the Low concentration lakes as the projected site values are quite spread out around the projected cluster mean. The map of the group structure for phosphorus shown in Figure 4.11 highlights that there is again some evidence of a spatial pattern in the distribution of the groups since all of the High lakes seem to be close to the coast. The three lakes which were in the High group for alkalinity are also in the High group for phosphorus. In terms of comparison between the FCM groups for phosphorus and the existing SEPA groups it can be seen that, as with alkalinity, there are several groups which have been amalgamated to form a larger group in the new group structure. All lakes from SEPA groups 2, 3 and 4 are contained within the Low concentration group. This is consistent with the earlier impressions gained from the functional data that there is a huge degree of overlap in the existing SEPA groups for this determinand and hence fewer groups can accurately represent the similarities in these lakes.

For chlorophyll, the cluster means for the Low and Intermediate concentration groups display a strong seasonal signal. Although the cluster means for the Low and Intermediate groups appear to be fairly close to one another in Figure 4.12(d), it can be seen from the plot of projected curve values in Figure 4.12(c) that there is clear separation between the FCM groups for chlorophyll, with all projected site values being close to their corresponding projected group centre. Figure 4.13 displays a similar geographical pattern to alkalinity in terms of the High concentration group, with all lakes in the High chlorophyll group also being found along the east coast. Of the 17 lakes in the Low and Intermediate groups the geographical pattern is not quite as clear although the Low concentration groups tend to be located fairly close to one another in the West coast and in central Scotland. The Intermediate lakes are more spread out in terms of location although all but one of these lakes falls into one of two groups; one in the North West and one in the South East.

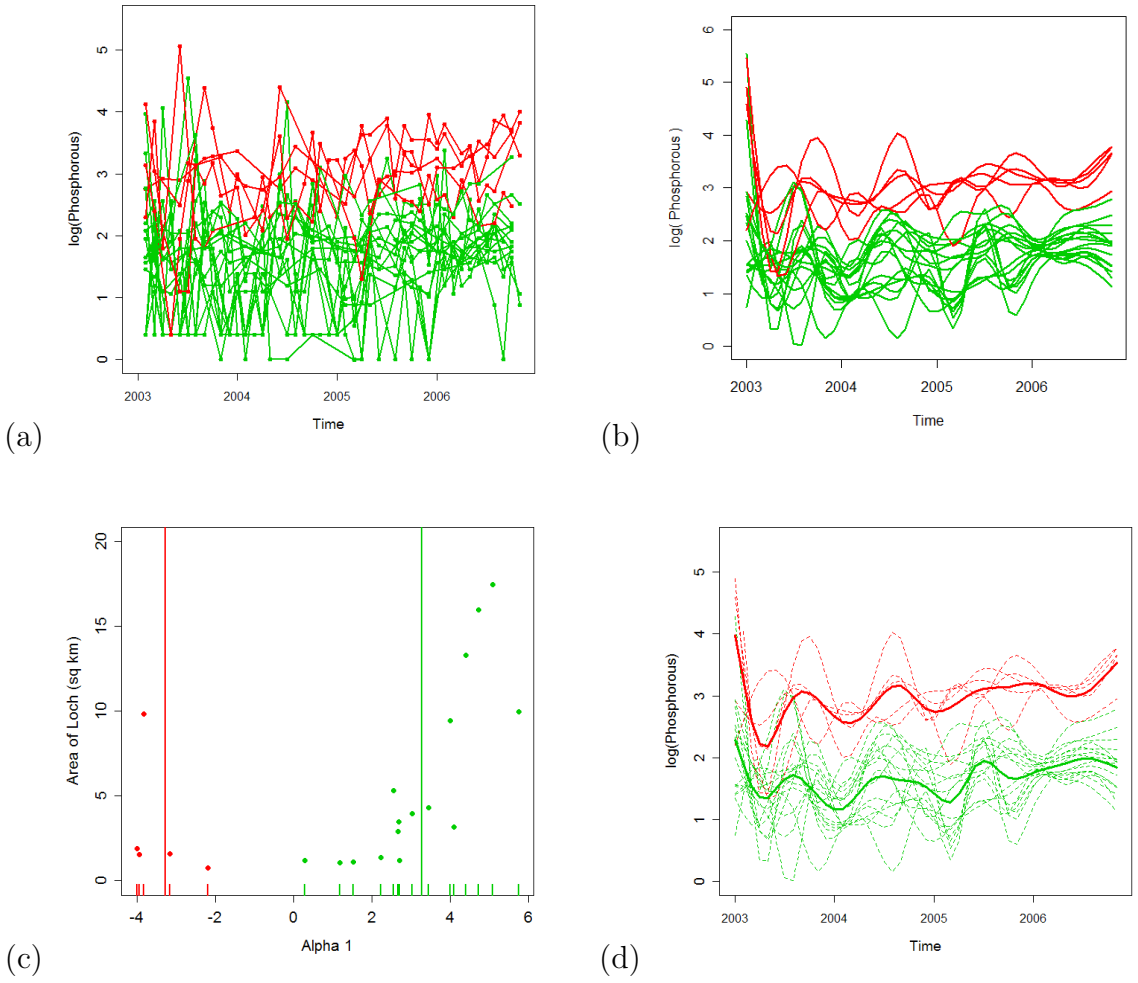


FIGURE 4.10: Summary of fitted FCM for phosphorus; (a) observed data, (b) predicted curves, (c) linear discriminant plot, (d) cluster means

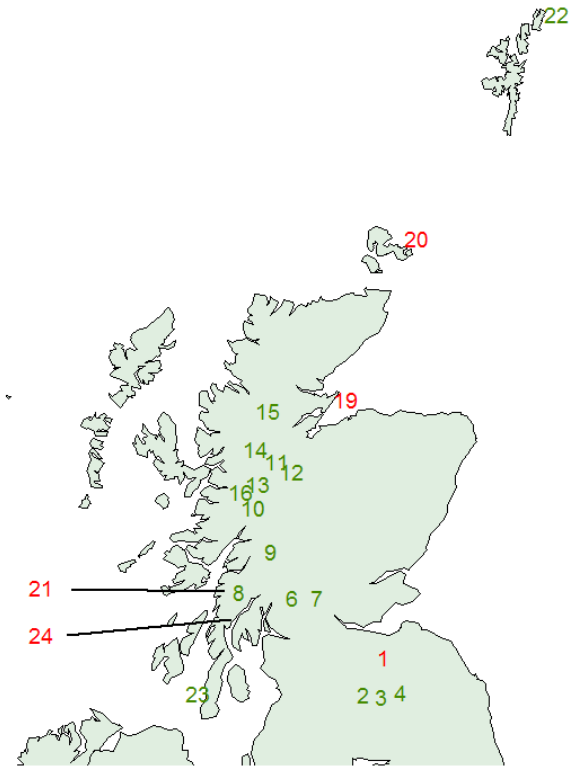


FIGURE 4.11: Map of Scotland showing FCM group structure for phosphorus

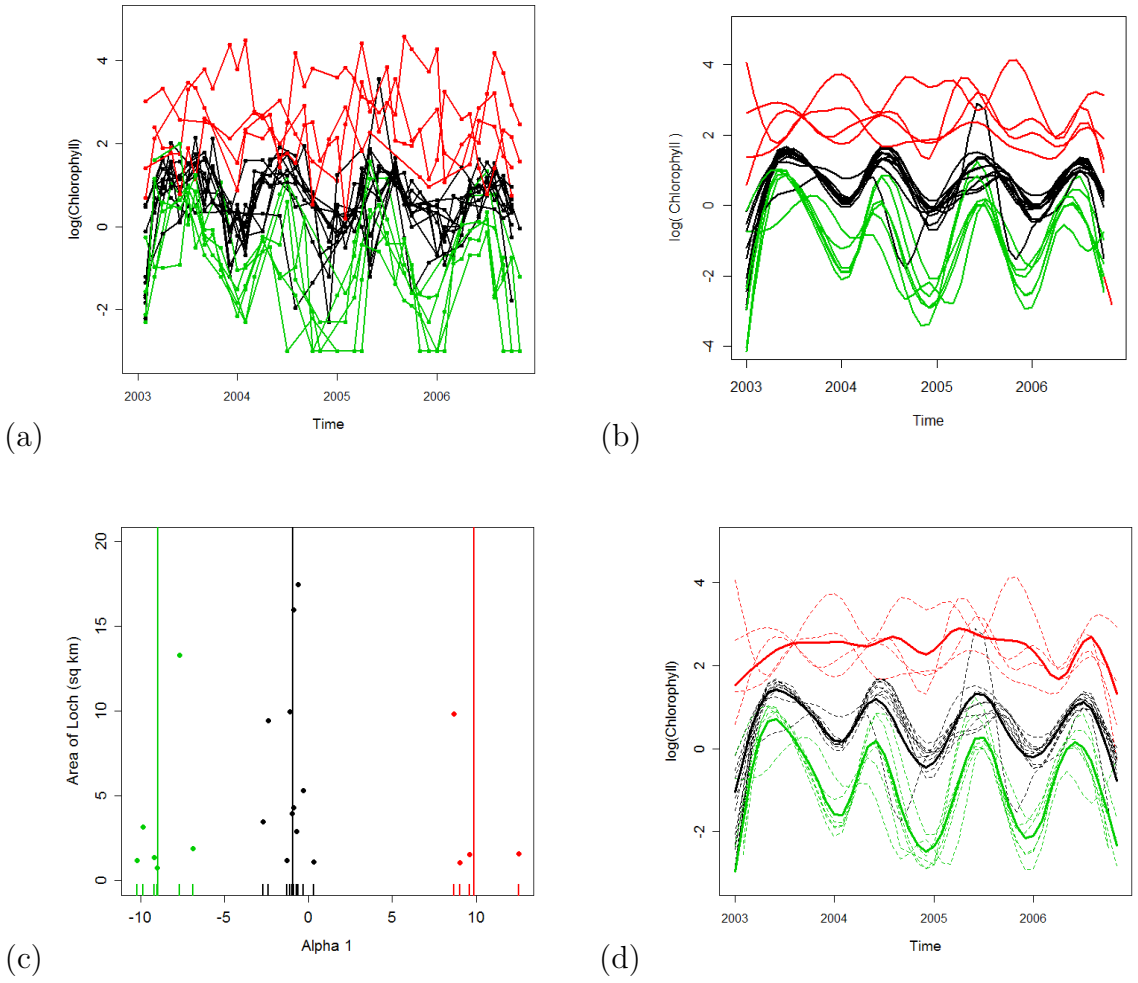


FIGURE 4.12: Summary of fitted FCM for chlorophyll; (a) observed data, (b) predicted curves, (c) linear discriminant plot, (d) cluster means

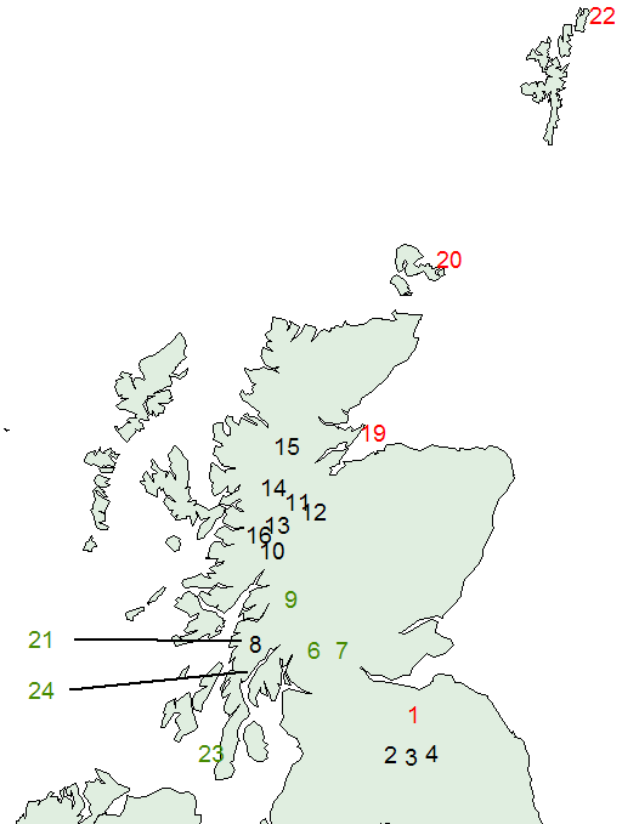


FIGURE 4.13: Map of Scotland showing FCM group structure for chlorophyll

Lake	**Name**	**Alk (Pr)** G=3	**Phos (Pr)** G=2	**Chl (Pr)** G=3	**SEPA group**	**cross prod**
1	Gladhouse Reservoir	H (1)	H(1)	H(1)	1	A
2	Talla Reservoir	I (1)	L(1)	I(1)	1	B
3	Fruid Reservoir	I (1)	L(1)	I(1)	1	B
4	St Marys Loch	I (1)	L(1)	I(1)	1	B
6	Loch Katrine	L (1)	L(1)	L(1)	2	C
7	Glen Finglas Reservoir	I (1)	L(1)	L(1)	2	D
8	Loch Avich	I (1)	L(1)	I(1)	3	B
9	Loch Ba	L (1)	L(1)	L(1)	3	C
10	Loch Arkaig	L (1)	L(1)	I(1)	3	E
11	Loch Beinn a Mheadhoin	L (1)	L(1)	I(1)	4	E
12	Loch Mhor	L (1)	L(1)	I(1)	4	E
13	Loch Mullardoch	L (1)	L(1)	I(1)	4	E
14	Loch Monar	L (1)	L(1)	I(1)	4	E
15	Loch Glascarnoch	L (1)	L(1)	I(1)	4	E
16	Loch Quoich	L (1)	L(1)	I(1)	4	E
19	Loch Eye	H (1)	H(1)	L(1)	5	F
20	Harray Loch	H (1)	H(1)	H(1)	5	A
21	Loch Tralaig	I (1)	H(1)	H(1)	6	G
22	Loch of Cliff	I (1)	L(0.96)	L(1)	6	D
23	Lussa Loch	I (1)	L(0.84)	H(1)	7	H
24	Loch Glashan	L (1)	H(1)	L(1)	7	I

TABLE 4.3: Table of FCM groups for univariate models.

G represents the statistically optimal number of groups.

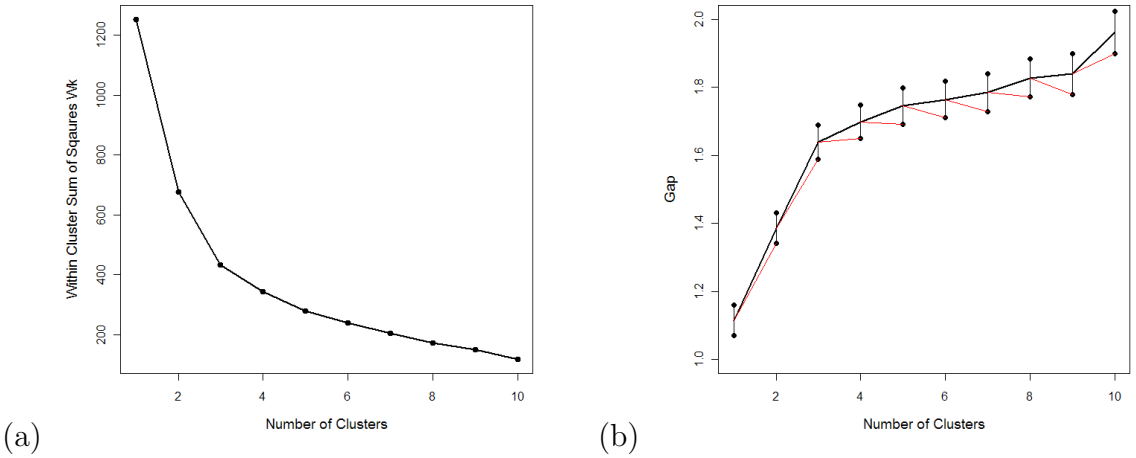
4.5.1 Multivariate Model

Ideally, as the WFD classification of lakes is based on multiple determinands of interest, any grouping structure will be formed using information from a combination of these determinands. Following on from the univariate models the next step was to combine the information from the alkalinity, phosphorus and chlorophyll data available and fit the FCM to these data to obtain a group structure. In all of the multivariate models presented, the data have been scaled to normalise the data before any clustering model is applied. It is necessary to suitably standardize the variates in order to avoid any one of the variates dominating the differences between the sites. For each determinand, the functional mean of all lakes has been removed from each individual lake and the observations have been divided by the standard deviation. To fit the FCM to multivariate data the first step was to fit

a different curve to the observed data for each determinand at each site. For a single site the basis coefficients for the curves corresponding to the multiple determinands are then concatenated, and k-means is applied to these combined sets of basis coefficients. The spline functions were fitted to each determinand separately and so the same number of splines (15) and ridge parameter (0.01) were used for consistency with the univariate model.

In this application BIC was found to be unstable when calculated for the multivariate models in situations where there were a relatively large number of potential groups (6 or more). As the gap statistic proved to be a reliable choice when determining the statistically optimal number of clusters for the univariate data, this was again used to select the optimal number of clusters in the multivariate case. For all univariate cases BIC and the gap statistic selected the same number of clusters as most appropriate. The same range of possible numbers of clusters, $G = 1, .., 10$, was considered and again 500 sets of reference data were generated for each number of clusters. To compute the within-cluster sums of squares for the reference data, a separate set of reference data was generated for each of the determinands and this was treated in the same way as the observed data, whereby estimated basis coefficients were combined, and a k-means approach was subsequently applied. The L-curve for the multivariate data is shown in Figure 4.14(a) and the gap statistic is shown in Figure 4.14(b). The L-curve itself gives no clear indication of what number of groups is appropriate as there is no single value after which the curve starts to flatten. The gap statistic plot indicates that four groups is most appropriate for the multivariate data, although the gap corresponding to three groups is only marginally smaller than the value required for it to be the optimal value. For this reason, the gap statistic was calculated several times for the multivariate data and 4 groups was consistently selected as most suitable, although each time, 3 groups was a close second.

To ensure that the number of clusters identified as optimal using the gap statistic was not sensitive to the number of basis functions used in the estimation of the curves the gap statistic was run a number of times; each time for a set of curves fitted using a different number of basis functions. For the multivariate FCM when the number of basis functions was between 10 and 18 the number of clusters which was identified as statistically optimal was consistently chosen to be 4. Figure 4.15 shows a set of different L-curves corresponding to clustering of curves which have been estimated with different numbers of basis functions. As



(a) (b)

FIGURE 4.14: L-curve and gap statistic plots for multivariate data

can be seen, there is only a minimal shift in the curves in terms of the within cluster sum of squares and the shape of each L-curve is almost identical. This implies that the underlying group structure of the curves is prominent, despite small changes to the variability of the curves expressed through differing numbers of basis functions. This is unsurprising as the formation of the groups is based primarily on mean level, and so it is likely this will dominate the number of clusters selected as optimal, rather than relatively minor fluctuations within the curves that may, or may not, be picked out with different numbers of basis functions. A similar procedure was also carried out to ensure the number of clusters selected as optimal for each of the univariate clusterings was not sensitive to the number of basis functions used to estimate the curves.

Following the selection of the number of groups, the model was fitted using the EM algorithm and the convergence criterion was set as there being a less than a 0.5% change in all three error variance estimates (corresponding to the three different determinands). Although this tolerance was higher than that used for the univariate models it seemed appropriate given that the three parameter estimates were required to converge simultaneously and worked well given the data. With the number of groups being four, the possible values of h were $1, 2$ and 3. A model was first fitted which had $h = 3$ however the projected values of the curves appeared to lie on a plane in 3-dimensional space and so the model was re-fitted with $h = 2$. With this second model, there was no evidence the projected curves lay in one-dimensional space rather than two dimensional space, and there was also much clearer separation between the groups. The final multivariate model with

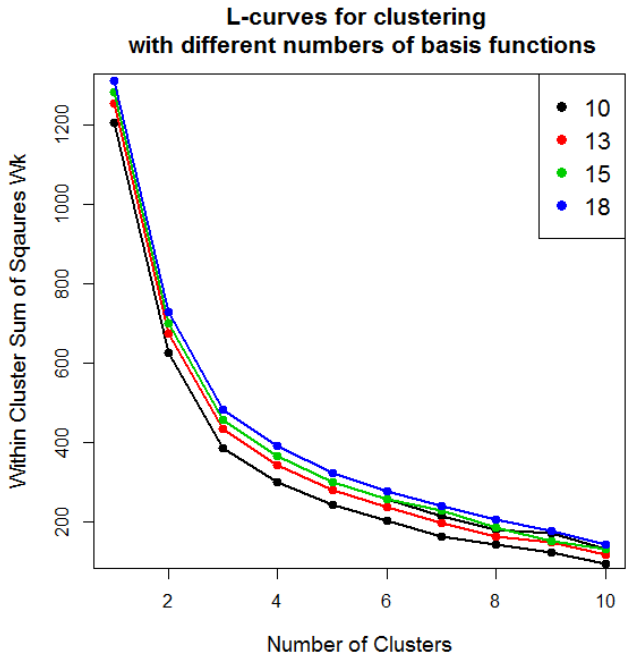


FIGURE 4.15: L curves for multivariate FCM where curves have been estimated using different numbers of basis functions

$P = 15$ for each determinand, $G = 4$ and $h = 2$ converged within 10 iterations when a tolerance of 0.05% difference was used.

The results of the multivariate FCM are summarised in Table 4.4 which shows the group structure of the multivariate model with probabilities of cluster membership, alongside the group structure corresponding to the existing SEPA groups. The cross-product classification according to the univariate FCM models has also been included in Table 4.4 in the final column. In addition, Figure 4.16 shows the four predicted group mean curves for each determinand. As the data was scaled and centered before the model was fitted these predicted curves are not directly comparable to the univariate FCM predicted mean curves for each determinand. It is clear while the split in the groups is again primarily based on the mean level of the determinands, there is some overlap in the group means, which indicates that other temporal features of the data have played a role in the formation of the groups. For example, the blue and black curves for alkalinity are very close in terms of mean level, however, while the blue curve is flat, the black curve exhibits a strong seasonal signal. The four groups have been labelled A, B, C and D and can be broadly summarised as follows;

Group A (Red) 3 lakes which appear to have high concentrations for all three determinands. The lakes in this group have a strong seasonal signal in terms of chlorophyll.

Group B (Green) 8 lakes with low alkalinity and phosphorus concentrations and intermediate levels of chlorophyll.

Group C (Black) 5 lakes which have intermediate concentrations for alkalinity and phosphorus and low levels of chlorophyll. This group, in general, is formed from lakes which have strong seasonal components.

Group D (Blue) 5 lakes which have intermediate levels for all determinands. The group mean for this group is very close to the mean value for all 3 determinands and seems to be very flat. While the flat curve is not representative of all lakes in this group, often the lakes within this group have only a weak seasonal pattern.

More groups are required for the multivariate model than any of the univariate models. This implies that the group means are being pulled in different directions by the different determinands. Groups are not always consistent in terms of the characteristics for each determinand, for example, there is no group which consists of lakes that have low concentrations of all three determinands.

While there are nine distinct cross-products from the univariate classifications, rather than the four groups identified by the multivariate FCM it can be seen that there is overlap between the two sets. Group E from the multivariate FCM corresponds exactly with the univariate FCM cross product classifications labelled by group B. As well as the relationship between the univariate and the multivariate FCM groups, it is also of interest to consider the relationship between the multivariate FCM groups and the original SEPA groups. This relationship is shown in the cross classification table provided in Table 4.5. As can be seen there is again a considerable degree of overlap between the original SEPA groups and those determined by the model based functional clustering model approach. For example, 75% of the SEPA group 1 lakes are contained in the multivariate FCM group labelled D, while all of the SEPA group 5, 4 and 7 lakes are contained within the multivariate FCM groups A, B and C respectively. Although the multivariate FCM approach suggested a smaller number of groups as statistically optimal, it is clear that a great deal of the original SEPA group structure is preserved within

Lake	**Name**	**Multi Grp (Pr)** $G = 4, P = 15(\times 3)$	**SEPA group**	**cross product**
1	Gladhouse Reservoir	A (1)	1	A
2	Talla Reservoir	D (1)	1	B
3	Fruid Reservoir	D (1)	1	B
4	St Marys Loch	D (1)	1	B
6	Loch Katrine	B (0.99)	2	C
7	Glen Finglas Reservoir	C (1)	2	D
8	Loch Avich	D (1)	3	B
9	Loch Ba	C (1)	3	C
10	Loch Arkaig	B (1)	3	E
11	Loch Beinn a Mheadhoin	B (1)	4	E
12	Loch Mhor	B (0.97)	4	E
13	Loch Mullardoch	B (1)	4	E
14	Loch Monar	B (1)	4	E
15	Loch Glascarnoch	B (1)	4	E
16	Loch Quoich	B (1)	4	E
19	Loch Eye	A (1)	5	F
20	Harray Loch	A (1)	5	A
21	Loch Tralaig	C (1)	6	G
22	Loch of Cliff	D (1)	6	D
23	Lussa Loch	C (1)	7	H
24	Loch Glashan	C (1)	7	I

TABLE 4.4: Table of FCM groups for multivariate models.

G represents the statistically optimal number of groups and P represents the number of spline coefficients

the definition of the multivariate groups. The estimated multivariate FCM clusterings, both univariate and multivariate are also consistent with the results of the functional F-tests carried out on the original SEPA groups shown in Table 3.2. The functional F-tests highlighted in particular that SEPA group 5 was distinct from the rest of the SEPA groups in terms of alkalinity and phosphorous and this is also reflected in the estimated FCM clusterings. An example of the predicted curves for each determinand at a single site (Site 1, Gladhouse Reservoir) are shown in Figure 4.17. The observed data are shown on these plots and the dashed lines represent 95% confidence bands. As could be expected, given these curves are marginal predictions from a multivariate model, the predicted curves for each determinand are not as good a fit to the observed data as the curves estimated using each of the univariate models. Despite this, Figure 4.17 shows the curves do capture the main features of the observed data and the amount of agreement

FCM SEPA	**A**	**B**	**C**	**D**	**Total**
1	1	0	0	3	4
2	0	1	1	0	2
3	0	1	1	1	3
4	0	6	0	0	6
5	2	0	0	0	2
6	0	0	1	1	2
7	0	0	2	0	2
Total	3	8	5	5	21

TABLE 4.5: Cross Classification Table of multivariate FCM and SEPA groups

between the predictions and the observed data shown in this particular example is typical of all other lakes. It is also worth noting that these curves are fitted to data which have been scaled and centered and so any minor fluctuations in the pattern over time are not of any real interest.

Although the multivariate predicted curves may not fit the observed data quite as well as the univariate curves, there is, in general, agreement between the groupings of lakes according to the univariate models and the multivariate model. For example, lakes 1 and 20, which are classed as being in the high group for all three of the univariate models, are in group A in the multivariate group structure. From Table 4.4 it can also be seen that some of the current SEPA group structure has been preserved within the multivariate FCM group structure. As with the univariate models, all of the SEPA group 4 lakes remain grouped together when the multivariate functional clustering model is used to determine the clusters. For the multivariate FCM both SEPA group 5 lakes continue to be grouped together under the new structure, as do the SEPA group 7 lakes.

With the multivariate model it is especially difficult to see how the groups can be separated considering only the predicted curves. The projected curves are again a useful tool in visual identification of the clusters and Figure 4.18 displays the combined information from 3 determinands projected onto two dimensional space. The orange points represent the projected cluster centres for the four groups. There is evidence from this plot that the four groups are well separated and no site is on the border of being in two groups which not only indicates that four groups is suitable for the given data, but also explains why the cluster membership probabilities for each of the lakes are so high. Without the plot of the

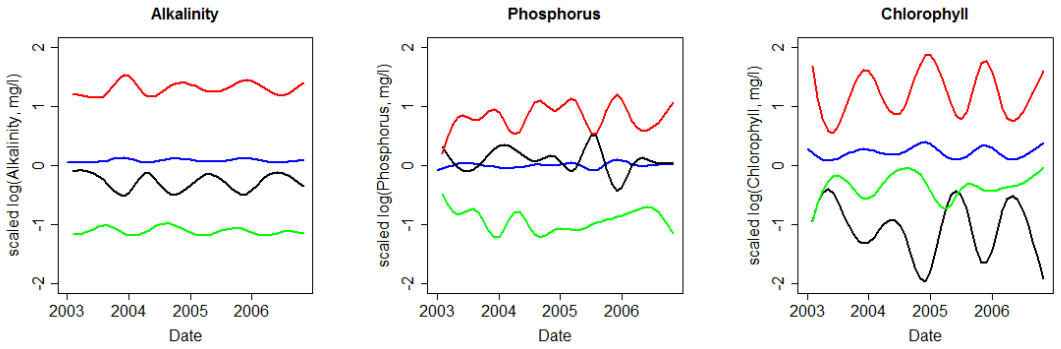


FIGURE 4.16: Multivariate model predicted group mean functions for alkalinity, phosphorus and chlorophyll

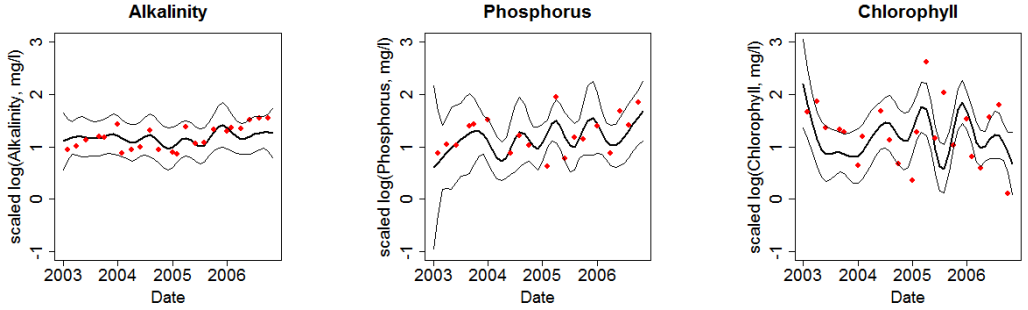


FIGURE 4.17: Predicted marginal functions of alkalinity, phosphorus and chlorophyll for Site 1

projected curves it can be difficult to see how we can predict with such certainty which site is in which group.

Figure 4.19 shows a map of the multivariate FCM group structure. As with the univariate models there is a group of high concentration lakes (group A lakes) which lie along the east coast. The group B lakes are all located in the north west of the country while the other two groups are less separable in terms of geographical location. While new group structures have been determined using the FCM, the question of how to choose the representative site within each group has yet to be addressed. At present, the representative site is often determined by logistics. Initially it was thought that the representative site within each group could be chosen by selecting the site which had the highest membership probability according to the FCM. However, as nearly all lakes have a cluster membership probability of one, it was decided that another sensible approach to choosing the representative site would be to chose the site whose projected value was closest to the appropriate projected cluster centre. The representative lakes which were

selected using this method are the lakes which are highlighted in orange on Figure 4.19.

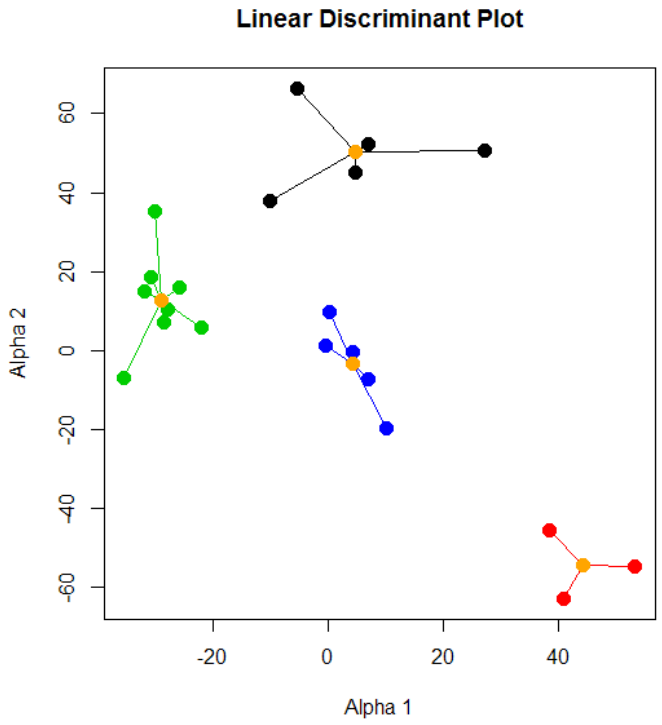


FIGURE 4.18: Projected curves and cluster means for multivariate FCM

4.6 Summary

In summary, functional clustering provides a statistical approach to defining groups of lakes based on the observed data. Hierarchical functional clustering of each individual determinand provided a good first step in terms of allowing us to visually assess if there was any clear underlying group structure. The large amount of overlap in the existing groups observed in the initial analysis of the data indicated that a clear grouping of the lakes would be unlikely and the hierarchical clustering reinforced this. Investigation of a statistically optimal number of groups for hierarchical functional clustering of the lakes data using the gap statistic indicated that fewer groups than the number currently used by SEPA would be sufficient in capturing the variability amongst the lakes.

Following from the hierarchical methods, model based clustering had several advantages over the non-probabilistic approach, including enabling standard model

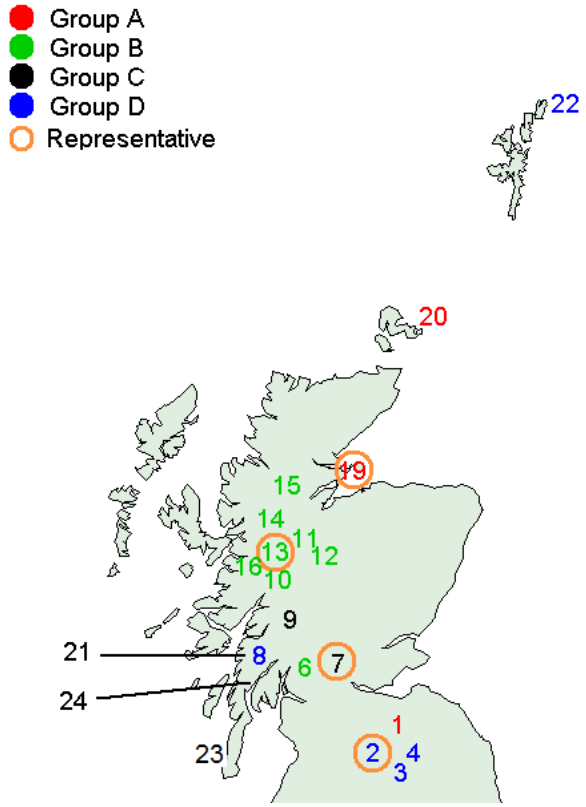


FIGURE 4.19: Map of Scotland showing multivariate FCM group structure

comparison techniques to be used to select the optimal number of groups and enabling us to calculate a confidence in classification of each site. Furthermore, the functional clustering model proposed by James and Sugar (2003) had the additional key advantage of being able to deal with irregular and sparsely sampled data, which was a problem in the Scottish lakes data. Treating the curves as a random effect lessens the importance of having a regularly spaced, complete dataset, which is a rarity in environmental settings due to situations such as adverse weather conditions stopping samples being collected or failure of equipment used for analysis of samples.

In order to compute initial estimates of the spline coefficients while taking into account the fact there may be an incomplete dataset at each site, ridge regression was used rather than unconstrained least squares which would overfit the data. This ridge regression avoided interpolation of the data and meant that missing data no longer had to be imputed by initially fitting interpolating splines. The curves fitted to the observed data using ridge regression were only slightly different than the curves fitted in the exploratory functional data analysis, which were fitted

to a regular dataset using penalised regression splines. It was clear that the ridge regression fits provided a good first step in fitting the FCM, particularly in the univariate case, as the final predicted curves for each site obtained using the fitted model were a good fit for the observed data. Even for the multivariate model, after using these estimates for the basis coefficients as a starting point for the EM algorithm, the marginal projections captured most of the main features of the underlying data.

Although being able to employ standard model comparison techniques to identify the best number of clusters was one of the attractive features of the model based clustering approach, problems were encountered when trying to use BIC with the multivariate models. Despite this, the gap statistic provided a suitable alternative approach which produced consistent results that were sensible given the earlier exploratory analysis of the current groups. It was reassuring that for the univariate models, where BIC could be calculated, there was agreement between the number of clusters identified as being best using both methods. The number of clusters determined using the functional hierarchical approach also agreed with the number of clusters identified for each of the univariate determinand FCMs.

All of the clustering methods, hierarchical and both univariate and multivariate models, indicated that the statistically optimal number of groups is less than the number of groups currently used by SEPA. The existing SEPA groups are in many cases combined to form larger groups within the new FCM based group structures. As already discussed, this is not surprising given the overlap in the current groups and the small number of lakes. There are, however, clearly some differences between groups of lakes, especially for alkalinity. Even though these differences are primarily based on mean level, which could be expected as there is no evidence of any trend in the data, the multivariate model did indicate a split in the groups in terms of the strength of seasonal signals at the lakes.

From a statistical viewpoint, the results of the functional clustering models for these data may be considered as slightly disappointing because of the fact that the mean level is essentially the sole driving force behind separation of the groups. This is due to the underlying data, and the method itself has the potential to determine a more interesting group structure which captures more than mean levels, if there was more happening in terms of temporal dynamics at the lakes. It is worth noting there are only 21 lake which have to be clustered, and these are based on a time series of just under 4 years of monthly observations. On the

basis of the results of the simulation study presented in Chapter 2, the available data considered for the Scottish lakes does not cover a long enough time period for even a moderate long term trend to be observed. Regardless of the formation of the groups on mean level, the functional approach to clustering continues to be worthwhile. For example, consider the situation where there is a group of lakes which display a negative trend over time and a group where the level remains constant over time. If clustering was based only on a single measurement, for example, an annual mean, the lakes which are potentially rapidly deteriorating may be grouped with the lakes which display no change.

The practical implication of defining fewer groups is that fewer representative lakes can subsequently be monitored, while still ensuring that differences between the lakes are being taken into account. As discussed previously, any possible reduction in monitoring has become especially important in view of financial and time constraints which are currently being imposed on environmental regulator agencies such as SEPA. While the WFD is an extremely complex piece of legislation, in which the classification of lakes encompasses a huge range of different determinands, both chemical and biological, grouping the lakes using a functional clustering approach provides a solid basis for the group structure which is based on observed data from some of the determinands of interest. Although the application of the functional clustering approach has been demonstrated in a particular setting there is potential for these methods to be used in conjunction with a wide variety of datasets. Slight modifications of the methods may be required for specific contexts and this is considered in the next chapter for data from a river network.

Chapter 5

Incorporating Spatial Correlation

The investigation into functional clustering of water quality data has thus far only considered groups of locations which have different geographical locations, are unconnected and are therefore assumed to be spatially independent. Situations where this assumption is not appropriate, and where it may be of interest to account for spatial correlation between locations have not yet been considered. Examples where spatial correlation is present in environmental data are abundant in the literature such as in air quality studies (Guttorp et al., 1994) and in the analysis of water temperature (Akita et al., 2007). While this spatial correlation is often incorporated into models used for prediction (Bowman et al., 2009, Shaddick and Wakefield, 2002), there are far fewer examples which discuss the inclusion of spatial correlation within clustering methods, and only a couple which examine the presence of spatial correlation for geographically referenced functional data.

This chapter will examine clustering of spatially correlated data in order to obtain groups of monitoring stations that are not only similar in terms of mean levels and temporal patterns of the determinand of interest, but which are also spatially homogenous. Hierarchical clustering will be applied to a set of nitrate data from monitoring stations along the River Tweed in the South of Scotland. River network data introduces a set of new challenges when considering functional clustering which go beyond the inclusion of spatial correlation. As well as including correlation based on standard Euclidean distance between the stations when forming clusters, the effects of stream distance based correlation will also be considered, along with flow-connectedness amongst the stations.

5.1 Spatial Functional Data Analysis

As mentioned, there is a wide literature available on modelling spatially correlated variables measured at different locations within a geographical region. Often the aim is to use these models to predict the determinands of interest at unobserved locations using kriging, which is a method of spatial prediction that interpolates between previously observed locations. There are also several examples of statistical analysis of spatially correlated data in multivariate settings where there are several response variables, such as Ver Hoef and Cressie (1993) and Pebesma (2004). In the multivariate context the determinands of interest are considered simultaneously and the spatial covariance has to be estimated for each.

Further to this, multivariate spatial statistical tools are also starting to be generalized for use with functional data. Delicado et al. (2010) provides a summary of recent contributions to methods of interpolation for the three classic types of spatial data structures; geostatistical data, point patterns and areal data. The different methods discussed include proposed approaches by Goulard and Voltz (1993), Nerini et al. (2010) and Giraldo et al. (2011) which all aim to offer a solution to the problem of predicting curves at unsampled locations. In addition to functional kriging, there are a number of other recent examples which consider other methods of spatial functional data analysis. Sun and Genton (2011*a*) propose a spatial correlation adjustment of the functional boxplots proposed in Sun and Genton (2011*b*). These boxplots can be used as an exploratory tool for visualizing spatio-temporal functional data and for outlier detection. Yamanishi and Tanaka (2003) develop a regression model for spatial functional data in which both response and explanatory variables are curves, and where the relation amongst the variables can change over space. This model combines two existing methods, geographically weighted regression (Brunsdon et al., 1998) and functional multiple regression (Ramsay and Silverman, 1997).

There are a number of examples in which spatial correlation is included within clustering techniques, particularly in the image analysis and remote sensing context. For example, Soares et al. (1996) considers a clustering approach by adapting the EM algorithm to include a spatial constraint based on neighbourhood information. The number of examples in the literature where spatial covariance has been incorporated into clustering of curves is, however, fairly limited. Two examples of spatial functional clustering are provided in Romano et al. (2010) and Secchi

et al. (2011), both of whom use iterative algorithms to partition geographically referenced data. In addition, Giraldo et al. (2010) extends existing ideas used for clustering to include spatial correlation between curves. Hierarchical clustering methods are adapted via weighting the dissimilarity matrix by a measure of spatial functional covariance. The methods in this paper are an extension of methods previously considered in the investigation of the lakes data in Chapter 4 and hence will be explored in more detail later in this chapter.

As with the majority of spatial statistical analysis, the initial aim here is to estimate the spatial correlation between stations however, rather than use this to predict concentrations of the determinand of interest at unobserved locations, it is of interest to build this information into clustering methods in order to identify groups of stations which display similar spatio-temporal characteristics.

5.2 Estimating Geostatistical Covariance

5.2.1 Covariance and Semi-variance

Calculating the covariance between observations collected at pairs of locations is carried out in order to estimate the association between them, however the covariance is unobtainable if there is only one observation collected at each location and the mean cannot be calculated. Although it is often the case that there is only one sample collected at each location, the assumption of second order stationarity overcomes this problem. If the underlying process is stationary then the distribution of the variable of interest has attributes, such as the mean, which are constant across space. The variance of the underlying process is also assumed to be finite and constant. If there are two different locations x_i and x_j which are separated by a lag of $h = x_i - x_j$ then let $Z(x_i)$ and $Z(x_j)$ represent observed values of a variable of interest Z at these locations. The covariance can then be defined as

$$Cov(x_i, x_j) = E[\{Z(x_i) - \mu\}\{Z(x_j) - \mu\}] \tag{5.1}$$

where μ is the mean of Z (assumed be constant). Since $x_j = x_i + h$, the above equation can be rewritten as

$$\begin{aligned} Cov(x_i, x_i + h) &= E[\{Z(x_i) - \mu\}\{Z(x_i + h) - \mu\}] \\ &= E[\{Z(x_i)\}\{Z(x_i + h) - \mu^2\}] \\ &= Cov(h) \end{aligned}$$

Hence the covariance between points depends only on the distance or the 'lag' between them. The dependence between values of Z separated by different lags is therefore known as the autocovariance function and is related to the autocorrelation function which has previously been discussed when assessing the temporal correlation. The relationship between the autocorrelation function and the autocovariance function is discussed in Webster and Oliver (2007) who provide a comprehensive introduction to spatial statistics and from whom the above notation has been taken. To ensure stationarity it is often necessary to estimate a spatial trend surface. This spatial trend is subsequently removed from the data and the residuals are modelled to determine an estimate of the underlying spatial correlation.

In addition to the covariance function, semi-variances, which are half the variance at a particular lag, are also used widely within geostatistics. The reason for this is that points are considered in pairs and so the semi-variance is equivalent to the variance per point at a given lag. Using the same notation as above, the variance of points separated by lag h can be defined as

$$\begin{aligned} Var[Z(x_i) - Z(x_i + h)] &= E[\{Z(x_i) - Z(x_i + h)\}^2] \\ &= 2\gamma(h). \end{aligned}$$

and hence the semi-variance is defined as $\gamma(h)$. Plots of empirical covariances and semi-variances against different lags, respectively referred to as covariograms and variograms, are used to summarise spatial relationships. Since semi-variances do not require the mean to be calculated, the variogram is far more popular than the covariogram in geostatistics. However, for reasons discussed later in this chapter, it is necessary for stream-distance correlation to be modelled using the covariogram. As a result of this, both Euclidean and stream based spatial correlation will be modelled using the covariogram, in order to ensure that any results obtained are comparable. Writing $Cov(h)$ to represent the covariogram at lag h, and $\gamma(h)$

to represent the variogram at lag h, then the relationship between the two is straightforward and can be written as

$$\gamma(h) = Cov(0) - Cov(h) \tag{5.2}$$

As $Cov(0)$ is a constant, it can be seen from the relationship in Equation 5.2 that the variogram and covariogram are mirror images of one another. There are a range of functions which are frequently employed to model spatial covariance including the Gaussian, exponential, linear with sill, and spherical models. The exponential model and the wider class of models that this is a part of, the Matérn class, will be discussed later in this section. All of these models have the ability to represent several key features found in the covariogram. For example, since distances are always positive, any covariance function should be positive definite. In addition, the function should monotonically decrease from $Cov(0)$ as the lag increases (locations which are far apart are less similar than those close together) and have a constant minimum which may either be finite or can be approached asymptotically. Conversely, as the variogram is a mirror image of a covariogram, variograms must be negative semi-definite and monotonically increase, again possibly asymptotically, to a constant maximum. In the literature, the signal variance of the process is commonly referred to as the sill variance. The sill is therefore the maximum variance reached in a variogram or covariogram. For monotonic variograms, if the sill is reached at a finite distance, then this distance is known as the range. The range is the distance beyond which locations will be spatially independent. Hence, responses at locations separated by distances greater than the range are spatially uncorrelated. For variograms which reach their sill asymptotically, the effective, or practical range can be identified. One definition of the effective 'range' is provided in Cressie (1993) who proposes that the effective range is the distance at which the variogram reaches 95% of its sill. In this chapter, the term range shall refer to either the effective range or the exact range. As well as the range and sill, often variograms will have a discontinuity at the origin which is referred to as the 'nugget'. This parameter represents measurement error, or the spatial variability on a smaller scale than the distance between the two closest points in the sampling region (Diggle and Ribeiro, 2007). An example of a variogram function to illustrate the different components is shown in Figure 5.1.

In order to construct an empirical covariogram, a covariogram cloud is first constructed by calculating the covariance between all possible pairs of observations

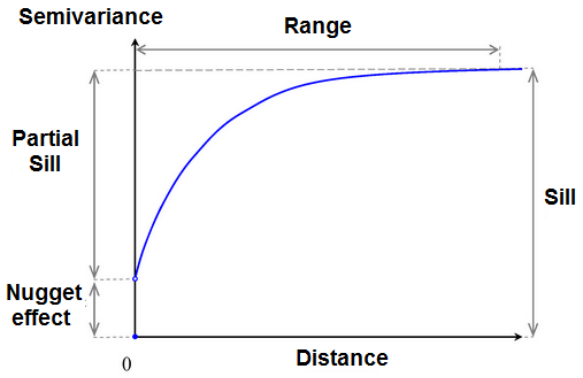


FIGURE 5.1: Example variogram function

from different locations separated by lag h. These estimated covariances are then plotted against the corresponding lags and the plot is binned by averaging at regular intervals. A variogram is obtained in the same way, but by using semi-variances in place of covariances. After an empirical covariogram plot has been obtained from the observed data, one of a set of valid covariogram models can subsequently be fitted to estimate the underlying covariance structure.

A broad class of covariance models is the Matérn family functions. For two observations separated by a distance of h units the Matérn covariance function is given as

$$Cov(h) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(2\sqrt{\nu}\frac{h}{\theta}\right)^{\nu} K_{\nu}\left(2\sqrt{\nu}\frac{h}{\theta}\right) \tag{5.3}$$

where Γ is the gamma function, K_{ν} is a modified Bessel function of the second kind and θ and ν are non-negative covariance parameters corresponding to the range and smoothness of the function respectively. The chosen value of ν influences the relationship between the range and the sill and in general, the smoothness of the function increases as ν increases. In Equation 5.3, as the smoothness parameter ν tends to infinity, the Gaussian covariance function is approached and when $\nu = 0.5$, the Matérn function is equivalent to the exponential covariance function. The exponential model can be written as

$$Cov_{exp}(h|\theta) = \begin{cases} \theta_0 + \theta_1 & \text{if } h = 0 \\ \theta_1 \exp(\frac{h}{\theta_2}) & \text{otherwise} \end{cases} \tag{5.4}$$

where θ_0, θ_1 and θ_2 respectively correspond to the nugget, partial sill and range parameters.

Webster and Oliver (2007) state that choosing variogram and covariance models and fitting them to data remains one of the most controversial topics in geostatistics. Some people prefer fitting models using subjective judgement while others use mathematical criteria such as ordinary least squares, weighted least squares or AIC to select the best fit. Cressie (1985) suggests a method of weighted least squares to fit variogram models, where the weights used are proportional to the number of observations within each 'bin'. The result of this is that more weight is given to the shorter lags and less weight is given to the larger lags as these are the points on the variogram which have been estimated using a small number of paired differences. While the modelling of the spatial data within this chapter has been carried out using covariances, the equivalence between variogram and covariogram in Equation 5.2, means that these weights can also be used with the covariances.

As the Matérn class of covariance functions is thought to be a very flexible and general set of functions which encompasses several of the covariance functions commonly used to estimate the spatial covariance of environmental data, the Matérn function has been used as the model of choice within this thesis. Weighted least squares, with the weights as defined in Cressie (1985), has been used as the method of selecting the optimal covariance parameter values.

5.2.2 Spatial Functional Covariance

One of our main areas of interest for the investigation of environmental data are functional data. Two methods of estimating spatial correlation for functional data are discussed in Giraldo et al. (2010), namely, the trace variogram and the multivariate variogram. Using Giraldo et al. (2010) as the main reference for this section, each of these methods will now be discussed.

The Trace Variogram

The idea of generalizing the variogram to be used with spatially correlated functional data are discussed in Giraldo et al. (2011) who suggest the use of the trace variogram. Let $g_1(t), ..., g_N(t)$ defined for $t \in [a, b] \subset \mathbf{R}$ be a set of curves which are realizations of a stationary, isotropic functional random process collected from N stations with corresponding location co-ordinates denoted by $x_1, \ldots, x_N$. Then

writing the distance between two locations (i, j) as h, the trace variogram can be defined as,

$$\gamma^*(h) = \frac{1}{2}\mathbb{E}\left[\int_{[a,b]} (g_i(t) - g_j(t))^2 dt\right]. \tag{5.5}$$

If the curves are estimated using splines then each curve can be expressed (as shown in Equation 1.14) as the product of a set of spline coefficients ($\mathbf{c}_i$) and a matrix of basis functions, $\Phi(t)$.

$$g_i(t) = \mathbf{c}_i^T \Phi(t),$$

where $i = 1, ..., N$. The integral in Equation 5.5 is then equivalent to the square of the functional distance, d_{ij}, as defined in Equation 4.1 since

$$\begin{aligned}\int_{[a,b]} (g_i(t) - g_j(t))^2 dt &= \int_{[a,b]} (\mathbf{c}_i - \mathbf{c}_j)^T \Phi(t)\Phi(t)^T (\mathbf{c}_i - \mathbf{c}_j) dt \\ &= (\mathbf{c}_i - \mathbf{c}_j)^T W (\mathbf{c}_i - \mathbf{c}_j)\end{aligned}$$

where $W = \int_{[a,b]} \Phi(t)\Phi(t)^T dt$. As with standard variograms, to obtain the empirical trace variogram, the trace variogram cloud can be computed by calculating the differences between all pairs of curves and plotting these differences against the corresponding distance between the locations. The points on this plot can then be 'binned' and averaged at a series of regular intervals. The estimated trace variogram can therefore be written as

$$\hat{\gamma^*}(h) = \frac{1}{2|N(h)|} \sum_{i,j \epsilon N(h)} (\mathbf{c}_i - \mathbf{c}_j)^T W (\mathbf{c}_i - \mathbf{c}_j) \tag{5.6}$$

where $|N(h)|$ is the number of curves separated by a distance of h units. After obtaining the empirical trace variogram from observed data, any standard variogram model can be fitted as if it were a standard univariate variogram. Giraldo et al. (2010) note that the fitted parametric trace variogram is always a valid variogram because its properties are those of a parametric variogram fitted from a univariate geostatistical dataset.

The Multivariate Variogram

In addition to the trace variogram, the multivariate variogram can also be used to describe spatial covariance for functional data. The multivariate covariogram, $\Gamma(h)$, was formalized in Bourgault and Marcotte (1991) to be used for spatially correlated multivariate data where there are m variables collected at a series of N locations with coordinates $x_1, \ldots, x_N$. Using notation from Giraldo et al. (2010), the multivariate variogram, $\Gamma(h)$ is defined as follows. For an m multivariate spatial process $\{\mathbf{Z}(x) = Z_1(x), Z_2(x), ..., Z_m(x)\ x \in D \subset R^d\}$ then $\Gamma(h)$ can be written as

$$\Gamma(h) = \frac{1}{2}\mathbb{E}[(Z(x) - Z(x+h))^T \mathcal{M}(Z(x) - Z(x+h))] \tag{5.7}$$

where $\mathcal{M}$ is a symmetric positive definite matrix used as a matrix in the calculation of dissimilarities, such as Euclidean distance where $\mathcal{M} = I$. In this case, the multivariate variogram is the sum of all m single variograms corresponding to each variable.

$$\Gamma(h) = \frac{1}{2}\mathbb{E}[(Z_l(x) - Z_l(x+h))^2] \tag{5.8}$$

$$= \sum_{l=1}^{m} \gamma_{ll}(h) \tag{5.9}$$

where $\gamma_{ll}(h)$ is the variogram for the l^{th} variable. Alternatively, if the Mahalanobis distance is used and $\mathcal{M}$ is the inverse of the variance-covariance matrix between the m variables, then $\Gamma(h)$ becomes a weighted sum of single and cross variable variograms.

It has already been shown (in Chapter 4) that the distance between any two curves can be calculated from the distance between the coefficients of the basis functions which define those curves. With this is mind, the multivariate variogram can also be used to estimate the spatial covariance of functional data by treating the set of basis coefficients as a multivariate random variable. As before, if there are a set of N curves, each expressed using P basis functions, then each set of coefficients, $\mathbf{c}_i = (c_{i1}, \ldots, c_{iP})$, is a vector of length P. Then for $p = 1, \ldots, P$, the set of p^{th} basis coefficients, can be written as $M_p = (c_{1p}, \ldots, c_{Np})$. Giraldo et al. (2010) state that these coefficients form a realization of a multivariate random variable, $\mathbf{M}(x) = (M_1(x), \ldots, M_P(x)) : x \in D \subset \mathbf{R}^d$. Here x is used to denote

the geographical co-ordinates of each of the curves. Subsequently, replacing the multivariate random field $\mathbf{Z}(\mathbf{x})$ with the coefficients $\mathbf{M}(\mathbf{x})$ in Equation 5.9 provides a variogram that can be used with functional data. In practise, it is necessary to simultaneously compute a $P \times P$ matrix of variograms and cross-variograms between the basis coefficients,

$$\Upsilon(h) = \begin{pmatrix} \gamma_{11}(h) & \gamma_{12}(h) & \dots & \gamma_{1P}(h) \\ \gamma_{21}(h) & \gamma_{22}(h) & \dots & \gamma_{2P}(h) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{P1}(h) & \gamma_{P2}(h) & \dots & \gamma_{PP}(h) \end{pmatrix}$$

where

$$\gamma_{ll}(h) = \frac{1}{2}\mathbb{E}(M_l(x_i) - M_l(x_j))^2$$
$$\gamma_{lq}(h) = \frac{1}{2}\mathbb{E}(M_l(x_i) - M_q(x_j))^2$$

Above, $l, q = 1, ..., P$ and $h = |x_i - x_j|$ is the distance between curves i and j, with coordinates x_i and x_j respectively. This matrix can be estimated using the Linear Model of Coregionalization (LMC) which is discussed in Goulard and Voltz (1992).

5.3 Covariance between locations on River Networks

The covariogram models discussed thus far are all intended to be used with standard Euclidean distances and while Euclidean distance is probably the most frequently used distance measure in spatial statistics, there are some contexts where it may not be the most appropriate metric for describing the spatial dependence. Measuring the distance between two stations on a river network is one such situation. The stream distance, which can be described as the "shortest distance between two locations, where distance is only computed along the stream network" (Ver Hoef et al., 2006) can be used in problems involving stations spaced along a river network. However, when fitting a covariogram model to an experimental covariogram based on stream distances there are some additional issues that need to be addressed. Ver Hoef et al. (2006) highlight that spherical and

linear models used in combination with stream distances, as opposed to Euclidean distances, can result in a covariance matrix which is not positive-definite and therefore invalid. Ver Hoef and Peterson (2010) discuss two classes of stream distance based covariance model; the 'tail-up' model (in reference to the tail of the moving average process moving upstream) and the 'tail-down' model (in reference to the tail of the moving average process moving downstream). A mixture of the models can also be used, this is called a variance component model and is discussed in Ver Hoef and Peterson (2010) and Cressie and O'Donnell (2010).

The stream-distance based covariance models are often used in conjunction with flow data, which is a measure of the volume of water passing a point per unit time. If stream segments flow into one another, they are called flow-connected. A matrix of flow-connectedness, F, can be defined to summarise the flow-connectedness between stations. If there are n stream segments, then the flow-connectedness matrix is an $n \times n$ matrix with the $(ij)^{th}$ entry

$$F_{ij} = \begin{cases} 1 & \text{if } i \text{ and j are flow-connected} \\ 0 & \text{otherwise} \end{cases} \tag{5.10}$$

One of the features of tail-down models (Ver Hoef and Peterson, 2010) is that although stream distance is used, correlation is permitted between locations which are not flow-connected. While this initially may seem unrealistic, enabling observations at locations which are not flow-connected to be spatially related to one another is a property of practical use in situations where the variable of interest is an organism which can swim against the flow, or is a chemical determinand which is associated with such organisms. An example of such a scenario would be studies which look at fish populations. Within this thesis interest lies in modelling the spatial relationships of nitrates data and hence the tail-up model appears to be the most suitable and will be discussed further.

5.3.1 Tail-up Model

The tail-up model was introduced in Ver Hoef et al. (2006) and Cressie et al. (2006) and unlike the tail-down model, it assigns a covariance of zero to stream segments which are not flow-connected. The consequence of this is that observations collected at locations which do not flow into one another are assumed to be

uncorrelated. To define this model, let $Z(x_s)$ and $Z(x_t)$ be the values of a random variable at locations x_s and x_t which are located on stream segments s and t respectively, and let h_{str} be the stream distance between them. Let $\mathbf{k} \in B_{x_s,x_t}$ be the set of all stream segments on the river network that are between segment s and segment t. Following this it is shown in Ver Hoef and Peterson (2010) that a class of tail-up models suitable for use with stream distance can be written as

$$Cov(h_{str}|\theta) = \begin{cases} 0 & \text{if } s \text{ and } t \text{ are not flow-connected} \\ \Pi_{\mathbf{k}\in B_{x_s,x_t}} \sqrt{\omega_{\mathbf{k}}} Cov_u(h_{str}) & \text{otherwise} \end{cases} \tag{5.11}$$

where $Cov(h_{str})$ is the standard Euclidean distance based model formulation of a chosen covariance function, such as the Matérn function previously discussed in Equation 5.3 and $\omega_{\mathbf{k}}$ refers to a set of weights. From Equation 5.11 it can be seen that the weighting between any two flow-connected points on the river, say s and t, is obtained by taking the product of the square root of the k weights over the set $B_(x_s, x_t)$, which corresponds to all stream segments that lie between s and t.

It is suggested by Ver Hoef et al. (2006) that the best way in which to define the weights, $\omega_{\mathbf{k}}$, is using flow volume. Wherever there is a point of confluence in the river network, and two feeder streams join to form one larger stream, then the weights ω can be computed provided there are flow data available. O'Donnell (2012) states that the weight at each of the feeder streams is the 'proportion of the overall contribution that the streams make to the overall volume after the join'. Hence, if there are two stream segments, labelled s and t say, and these have volumes vol_s and vol_t respectively, then the weight for stream segment s can be defined as

$$\omega_{s,t} = \frac{\text{vol}_s}{\text{vol}_s + \text{vol}_t}.$$

Including these weights in this way effectively means that covariances between two stations which are flow-connected will be relatively small if one of the stations is on a minor (low flow volume) stream segment, even if the two stream segments are close together in terms of the stream distance.

If the flow data cannot be obtained or simulated for the entire river network, Ver Hoef et al. (2006) suggest that a proxy variable such as stream order can be used as a suitable alternative. While Ver Hoef and Peterson (2010) indicate that

the use of stream order data are less computationally intensive than using flow volume data, O'Donnell (2012) suggests that the use of flow data are far more descriptive of the river network than stream order, and highlights the potential of using flow data calculated at a series of different time points so that weightings in the covariance model could change over time depending on flow-volumes.

5.3.2 Estimating Covariance with the tail-up model

As mentioned earlier, it is necessary for the stream distance spatial covariance to be modelled via a covariogram. The reason for this is the more complex structure of the tail-up model shown in Equation 5.11. The exponential covariance function is used to illustrate why a variogram is inappropriate for stream distance based modelling in O'Donnell (2012). Using the relationship between the variogram and covariogram shown in Equation 5.2 and the tail-up model shown in Equation 5.11, where $Cov(h_{str})$ is defined as in Equation 5.4, then the stream distance based exponential variogram can be written as,

$$\begin{aligned} \gamma(h_{str}) &= Cov(0) - Cov(h_{str}) \\ &= \theta_0 + \theta_1 - (\omega_{s,t}\theta_1 \exp(-\frac{h_{str}}{\theta_2})) \\ &= \theta_0 + \theta_1(1 - \omega_{s,t}\exp(-\frac{h_{str}}{\theta_2})) \end{aligned} \tag{5.12}$$

Here, as before θ_0, θ_1 and θ_2 respectively are the nugget, partial sill and range parameters. From this it can be seen that a key issue is that is not possible to separate the weighting structure, $\omega_{s,t}$, from the observed semi-variances in the data in order to estimate the variogram model parameters. O'Donnell (2012) states that the consequence of this is that "the estimation of θ_0, θ_1 and θ_2 will need to factor in the impact of the weights, as they are likely to affect θ_1 and θ_2 if the variogram formulation is to be used". It is also noted that failure to account for this is likely to result in poor descriptions of the underlying correlation structure. Alternatively, using the covariance function overcomes this problem as the parameters and the weight structure can be separated,

$$Cov(h_{str}) = \omega_{s,t}\theta_1 \exp(-\frac{h_{str}}{\theta_2}) \tag{5.13}$$

In order to fit the tail-up model, the first step is to compute the observed pairwise covariances (see Equation 5.1) for all pairs of stations which are flow-connected.

Subsequently, these covariances can be plotted against lags (measured in terms of stream distance) and binned at regular intervals to obtain an empirical stream-distance based covariogram. A standard covariogram model can then be fitted and evaluated to obtain a covariance matrix, denoted V. Subsequently, as discussed in Ver Hoef et al. (2006), to obtain a valid stream based covariance matrix the Hadamard product of V and a weight matrix Θ is computed. The matrix Θ contains zeros whenever stations are not flow-connected, and when stations are flow-connected, Θ contains the square root of the percentage of flow volume weight, $\omega_{s,t}$. The construction of the matrix Θ ensures both that the flow weights are incorporated in the estimate of the covariance structure, and that stations which are not flow-connected are not spatially correlated.

It is the aim of this work to extend the application of the tail-up model to functional data and then to incorporate the estimate of stream based covariance within functional clustering approaches. Methods of including spatial correlation into different clustering methods for both standard and functional data will now be discussed.

5.4 Including Spatial Covariance within Clustering Methods

If the spatial covariance in a process has been estimated using a variogram or covariogram, this can subsequently be incorporated within hierarchical clustering in order to group the data into contiguous zones where the attributes of one, or more, variables are similar. Both Oliver and Webster (1989) and Bourgault et al. (1992) propose the idea of weighting the distance matrix which represents the dissimilarities between samples by using the variogram and the multivariate variogram, respectively. Giraldo et al. (2010) not only extend the ideas of estimating spatial covariance to the functional data setting but furthermore develop the idea of using these functional spatial covariance matrices as weight matrices when clustering hierarchical data.

For standard univariate data, where there is a set of locations with a value of a single observed variable collected at each location, Oliver and Webster (1989) suggest weighting the original distance matrix, d_{ij}, using the variogram calculated

for the distance between the stations as follows,

$$d_{ij}^{w} = d_{ij}\gamma(h). \tag{5.14}$$

Above, $\gamma(h)$ is the corresponding value of the variogram calculated at the distance between stations i and j. In the multivariate case, where there is more than one variable collected at each location, Oliver and Webster (1989) suggest using principal component analysis (PCA) on the set of variables, and then calculating a variogram which corresponds to the value of the first principal component at each location. The distance matrix can then be weighted by this principal component based variogram in the same way as in the univariate case. Alternatively, for multivariate data collected at a series of different locations, Bourgault et al. (1992) propose weighting the distance matrix by the Multivariate variogram initially discussed in Bourgault and Marcotte (1991).

As already stated, both of the approaches outlined above were initially intended for use with standard non-functional data, however they can be generalised to be used in conjunction with functional data. Incorporating spatial covariance into hierarchical functional clustering can be achieved by weighting the functional distance matrix defined in Equation 4.1 using either the trace variogram (Equation 5.6) or the functional multivariate variogram (Equation 5.7). Both methods of hierarchical clustering for spatially correlated functional data are applied to a climatology dataset in Giraldo et al. (2010), and while it is noted there are slight differences between the clusters obtained under each approach, neither is selected as being better than the other in terms of identifying spatially homogenous groups. As both the methods discussed here for estimating spatial association between functional data are based on the variogram, neither are suitable for use with stream distance. The hierarchical clustering approach was thought to be a suitable method to be used in this context rather than the functional model based clustering approach explored in Chapter 4. The key reason for this is that the FCM clusters sets of basis coefficients, rather than stations, and hence the spatial covariance matrix incorporated within the FCM would need to be defined in terms of covariance between individual spline coefficients. The use of river network data and stream distance would add further complications with the unique difficulties of also requiring flow-connectedness and flow weights to be defined in terms of the spline coefficients.

5.4.1 Clustering stations on a River Network

Following on from the approaches discussed in Giraldo et al. (2010) for hierarchical clustering of spatially correlated functional data it is the aim here to extend these ideas further so that they can be applied to data which have been collected at monitoring stations which lie on a river network. An obvious extension of the work already proposed by Oliver and Webster (1989) is to weight the functional dissimilarity matrix using stream distance covariance rather than Euclidean distance based covariance. However, the complex structure of river network data introduces several additional features which need to be taken into consideration. A stream distance based covariance model such as the tail-up model is required in order to ensure that the covariance structure estimated is valid. As already discussed, it is necessary to estimate the stream distance based spatial relationships using covariances as opposed to the semi-variance to ensure that the tail-up model parameters are identifiable.

To estimate the stream distance covariance for functional data, we need to define a metric for measuring the covariance between two curves. For standard, non-functional data, Cressie (1993) states that the covariance between stations s_i and r_i at one particular point in time is given by

$$Cov(s_i, r_i) = \sum_{i=1}^{N} \frac{(Z(s_i) - \bar{Z}(s_i))(Z(r_i) - \bar{Z}(r_i))}{N}, \tag{5.15}$$

where $Z(s_i)$ and $Z(r_i)$ are the values of the variable at stations s and r at time point i. Keeping in mind both the above equation and the definition of the trace variogram in Equation 5.6, which uses the area between two curves to represent the difference between them, one potential measure for estimating functional covariance has been developed. Using the same notation as before, let $g_1(t), \ldots, g_N(t)$ defined for $t \in [a, b] \subset \mathbf{R}$ be a set of curves which are realizations of a stationary, isotropic functional random process collected from N stations with corresponding location co-ordinates denoted by $x_1, \ldots, x_N$. Also as before, if each curve is expressed using P basis functions $g_i(t) = \mathbf{c}_i^T \Phi(t), i = 1, \ldots, N$, then each set of coefficients, $\mathbf{c}_i = (c_{i1}, c_{i2}, \ldots, c_{iP})$, is a vector of length P. To calculate the covariance let us first define the mean curve, denoted $\bar{g}(t)$. Following Equation 3.1, $\bar{g}(t)$ can be defined by the mean of the basis coefficients representing the set on all N

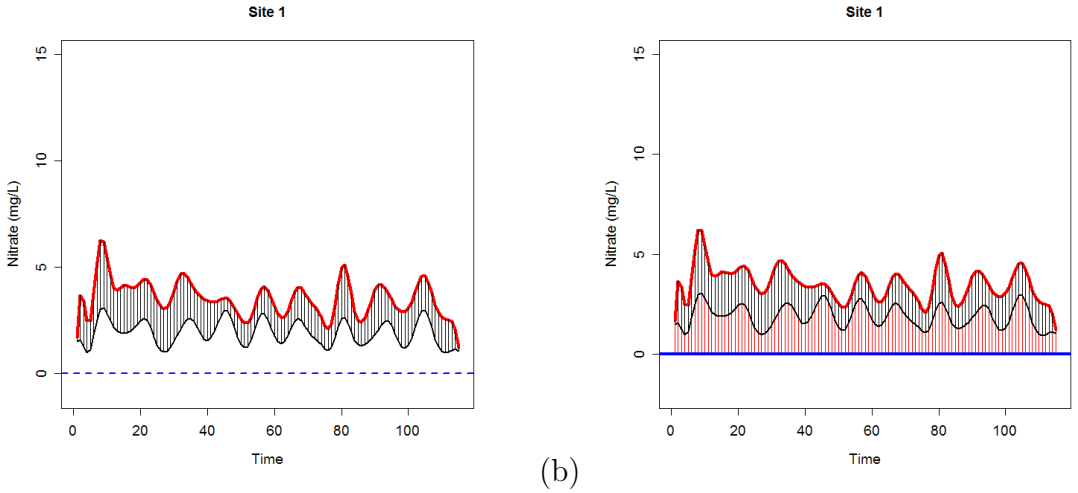
curves at time point t,

$$\begin{aligned} \bar{g}(t) &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{c}_i\Phi(t) \\ &= \bar{\mathbf{c}}^T\Phi(t) \text{ where} \\ \bar{\mathbf{c}} &= \left(\frac{1}{N}\sum_{i=1}^{N}c_{i1}, \frac{1}{N}\sum_{i=1}^{N}c_{i2}, \ldots, \frac{1}{N}\sum_{i=1}^{N}c_{iP}\right) \end{aligned}$$

For any two stations, a naive approach to computing the functional covariance could be to multiply the differences between the curve representing each station and the curve representing the overall mean, in other words

$$Cov(g_i(t), g_j(t)) = \int (g_i(t) - \bar{g}(t))^2 dt \int (g_j(t) - \bar{g}(t))^2 dt \tag{5.16}$$

There are however two problems with this approach. The first is that the direction of the difference is not reflected by the above equation as it is essentially the product of two areas, the area between the curve representing each station and the overall mean curve. While covariances can be positive or negative, the area



(a) (b)

FIGURE 5.2: Difference between station 1 Nitrate and Mean Nitrate (original and de-trended)

between the station curve and the mean curve will always be a positive value and so using this approach there is no way to determine if the station in question has values which are below or above the overall average. For example, on Figure 5.2 (a) the nitrate concentration at station 1 is shown by the solid black line, the overall mean curve is shown by the solid red line and the difference between the two curves is represented by the black area between the two curves. Calculating

the difference by computing the area between the curves gives no indication that the station 1 concentration is less than the overall mean concentration. The second problem is that multiplication of two areas is likely to result in large values which are not suitable in the computation of covariance and it is thought that covariance will, at some point, tend to zero as the distance between stations increases. To overcome both of the above problems a 'reference line' can be defined. The area between the mean curve and this reference line can be used to both reflect the direction of the difference between a given station and the overall mean, and can be used to standardize the areas so that the measures of covariance are on a suitable scale. Since curves should not fall below this reference line it should be set as a horizontal line which is below the minimum value of the set of curves $\mathbf{g}_i(t)$ where $i = 1, \ldots, N$.

Writing the reference line as $g_{ref}(t)$ and the corresponding set of basis coefficients which define this line as c_{ref}, then the area between the reference line and mean curve can be written as

$$\begin{aligned} \int (\bar{g}(t) - g_{ref}(t))^2 dt &= (\bar{c} - c_{ref})^T W (\bar{c} - c_{ref}) \\ &= \bar{M} \end{aligned}$$

Similarly, the area between the curve representing station i and the reference line can be written as

$$\begin{aligned} \int (g_i(t) - X_{ref}(t))^2 dt &= (\mathbf{c}_i - c_{ref})^T W (\mathbf{c}_i - c_{ref}) \\ &= M_i \end{aligned} \tag{5.17}$$

Then the difference between these two measures, $M_i - \bar{M}$, will not only reflect the magnitude of the difference between curve $g_i(t)$ and the mean curve $\bar{g}(t)$ but will also give an indication if the station has higher or lower than average values. Additionally this difference can also be standardized by dividing by the difference between the mean curve and the reference line, $\bar{M}$, in order to ensure that the scale is appropriate. For example, on Figure 5.2 (b), the blue horizontal line represents the reference line $g_{ref}(t)$, the black area represents M_1 the difference between the station 1 curve (shown by the black line) and the mean curve (shown by the red line). The reference area M_{ref} is the area between the mean curve and the reference line.

Following from this, the functional covariance between stations i and j can be defined as

$$Cov(g_i(t), g_j(t)) = \frac{(M_i - \bar{M})(M_j - \bar{M})}{\bar{M}^2} \tag{5.18}$$

where $i \neq j$. This results in a single value which summarizes the covariance between the functions at the two stations over the time period of interest. These point summaries of the covariance between pairs of curves can then be used to create an adjusted covariogram cloud as described in Section 5.3.2. In accordance with the definition of the tail-up model described in Equation 5.11 a standard covariogram model, such as the Matérn function, can next be fitted to this empirical covariogram (Equation 5.10). Evaluating this model at the relevant stream distances will result in a stream distance based functional covariance matrix, V. In order to obtain a valid stream distance based covariance matrix, $Cov^*(h_{str})$, the element-wise product of V and the weight matrix Θ, discussed in Section 5.3.1, can be computed.

To find spatially homogenous clusters of stations on a river network, $Cov^*(h_{str})$ can be used as a weight matrix in a similar way to the trace variogram shown in Equation 5.14.

$$d^c_{i,j} = d_{i,j} Cov^*(h_{str}) \tag{5.19}$$

Here, as before, $d_{i,j}$ is the functional distance matrix.

The key difficulty in clustering stations on a river network is due to the necessity to model the covariance between the stations rather than the semi-variance. Defining the covariance between curves in the manner outlined in Equation 5.17 enables a summary value of the spatial dependence between stations to be obtained. Consequently, these values can be used to fit the tail-up covariance model which can be used to estimate a valid stream distance based covariance structure that takes into account stream distance, flow-weights and flow-connectedness between stations. These are features of the data which are not considered when using the Euclidean distance based trace-variogram to cluster river network data.

5.5 The River Tweed

In this section the methods for estimating spatial covariance already discussed will be applied to a set of functional data from monitoring stations located along a river network. Spatial covariance based on both Euclidean and stream distance, the latter also including flow-connectedness, will then be incorporated within hierarchical clustering methods in order to identify groups of stations which are spatially similar.

The data used in this section come from a network of locations along the River Tweed which is located in the Scottish borders. Agriculture is a significant industry within the area and the water environment in the wider River Tweed catchment is an important economic, social and environmental asset (SEPA, 2009). The river is primarily surrounded by arable land, passing through only a small number of built-up areas. The location of the Tweed within Scotland is shown on Figure 5.3 while the river network itself is shown on Figure 5.4 where each of the points shown on the network is a monitoring station location. Data have been provided by SEPA on different chemical and biological determinands at 83 unique monitoring stations, covering dates between January 1986 and October 2006. As with the lake data, there are not data available at all stations over the entire time period and so a subset of data have been selected such that there are a set of stations which each have a reasonable number of observations that cover a common time period. Although the exact sample dates and frequency of sampling do not need to match across stations, it is important to ensure there is both a reasonable quantity of data at each station, and that the start and end dates of the time series are similar. The exact criteria for selecting the stations used in our analysis will be discussed later.

While the width of the river streams in Figure 5.4 are not shown to scale (since no data are available for this) the thickness of the lines representing each of the streams is proportional to the estimated average flow volume. This means that thicker lines on Figure 5.4 are likely to correspond to wider stretches on the river. It is clear that the Tweed network has a complex structure with many tributaries flowing into the main stream. The main river is the heavy line which runs from the far South-West to the North-East and in total there are 298 stream segments. Flow data provided by SEPA were estimated for the Tweed using a computer package called Low Flows 2000 (Goodwin et al., 2004). As the flow data

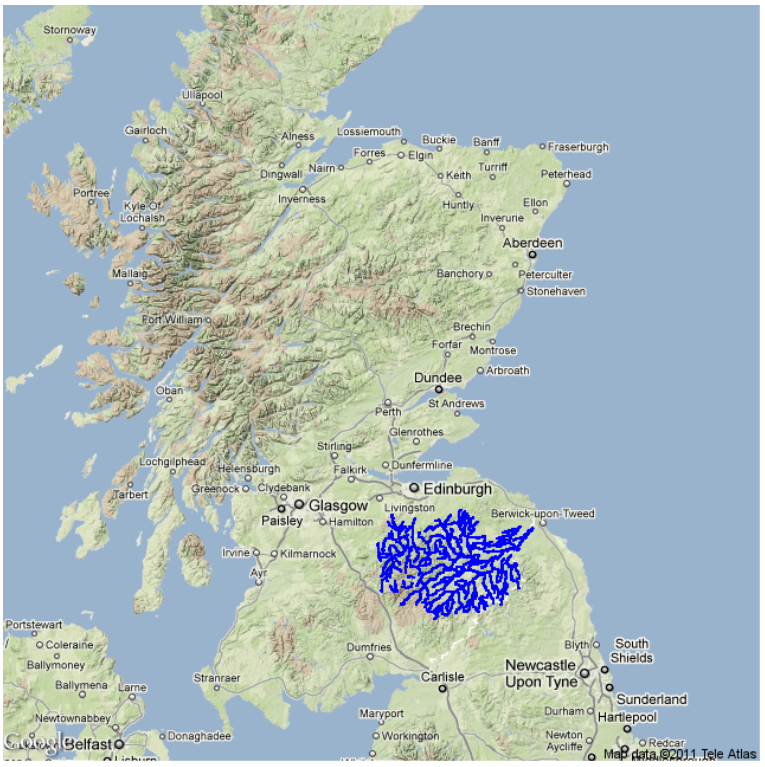


FIGURE 5.3: Map of Scotland showing location of River Tweed

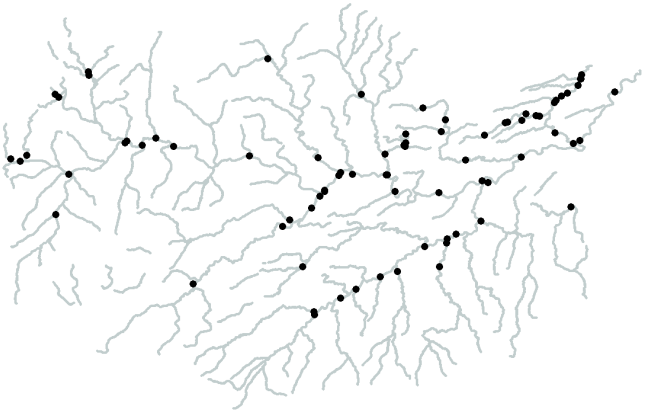


FIGURE 5.4: River Tweed Network showing location of monitoring stations

are estimated rather than observed they remain static throughout the time period considered and hence do not incorporate irregular weather events which may result in changes in flow.

Although there are a range of different determinands available, only nitrate

will be considered in this investigation. O'Donnell (2012) states that nitrate and phosphorus are regarded as being amongst the most interesting of the variables collected on the Tweed as they reflect the changing status of the network. However, as the phosphorus data collected are heavily affected by limit of detection issues (also discussed in O'Donnell, 2012) they will not be considered here.

5.5.1 Nitrate Data Exploration

Figure 5.5 displays a plot of the sample dates for nitrate at each of the 83 stations. As already discussed there are inconsistencies both in the time period the data covers and in the sampling dates. The large majority of the stations have samples which were collected at an approximately monthly frequency, however, if there was more than one observation collected in a single month at any station, the average of the observations within that month was computed. Very few of the stations had multiple observations each month. The black dashed line on this plot indicates the start date of the subset of data used here, while the stations shown in red indicate those which were removed as they were deemed to have an insufficient quantity of data to be compared to other stations. The criteria for removing stations were that stations were removed if the first date of sampling was after June 1997, or if there were less than 36 observations (approximately 3 years worth of data) between 1997 and 2006. This means there are approximately monthly data for a set of 77 stations covering a 9 and a half year time period from Spring 1997 to Winter 2006. In all subsequent analysis of the River Tweed data, this subset of data is used.

5.5.2 Log transformed or Raw data?

Exploratory analysis of the Tweed nitrate data highlighted that there were substantial differences between the stations in terms of the amplitude of the seasonal patterns observed, and in terms of the mean levels. There also appeared to be a change in the seasonal component over time at a few of the stations. Examples of the pattern in nitrate levels over time at two of the stations (Station 1 and 15) are shown in Figure 5.6(a). It can be seen that there are large differences between these two stations; while station 1 shows a low mean nitrate level and

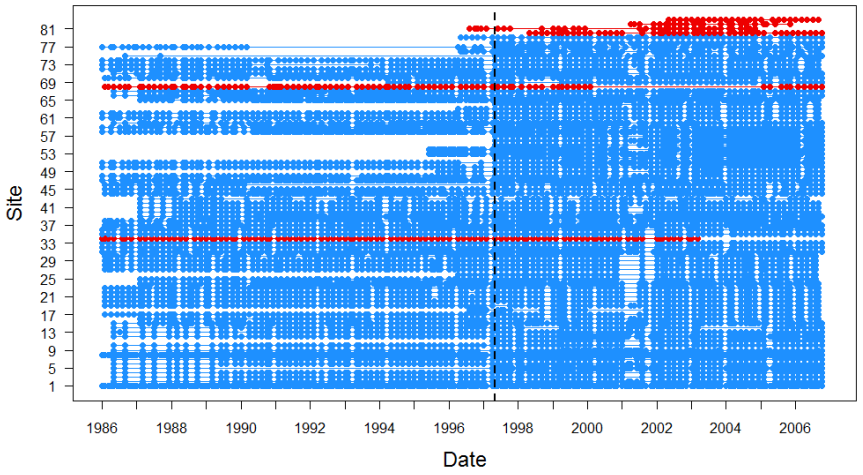


FIGURE 5.5: Nitrate sample dates for River Tweed data with vertical line showing start of time period considered

The red points correspond to stations which have not been included in further analysis

a fairly constant seasonal signal, station 15 exhibits a higher mean level and far more variability with evidence of a change in the variability over time.

A natural next step was to investigate applying a transformation to the data. A natural log transform is commonly used to stabilize the variability in data such as that observed at station 15. The log transformed nitrate data for stations 1 and 15 are shown in Figure 5.6(b). Here it can be seen that while there continues to be a difference between the mean levels at the stations, albeit less distinct, there is less disparity between the strength of the seasonal patterns at the stations. In addition, there remains evidence of non-constant variance at station 15 and the log transform has done little to overcome this potential issue in the data.

The variability in the nitrate data at station 15 was explored further to assess whether the apparent change in variability was actually a change in the seasonal pattern throughout time. In order to check this, an additive model, with additive smooth trend and seasonal terms (Equation 1.7), and a bivariate model with a smooth interaction of trend and seasonal terms (Equation 1.5) were both fitted to the data from station 15 over the time period from 1986 to 2006. An approximate F-test was then carried out to determine which of these models was more suitable given the data. The additive model assumes a constant seasonal pattern over time, while the bivariate model allows the seasonal pattern to change over time. These

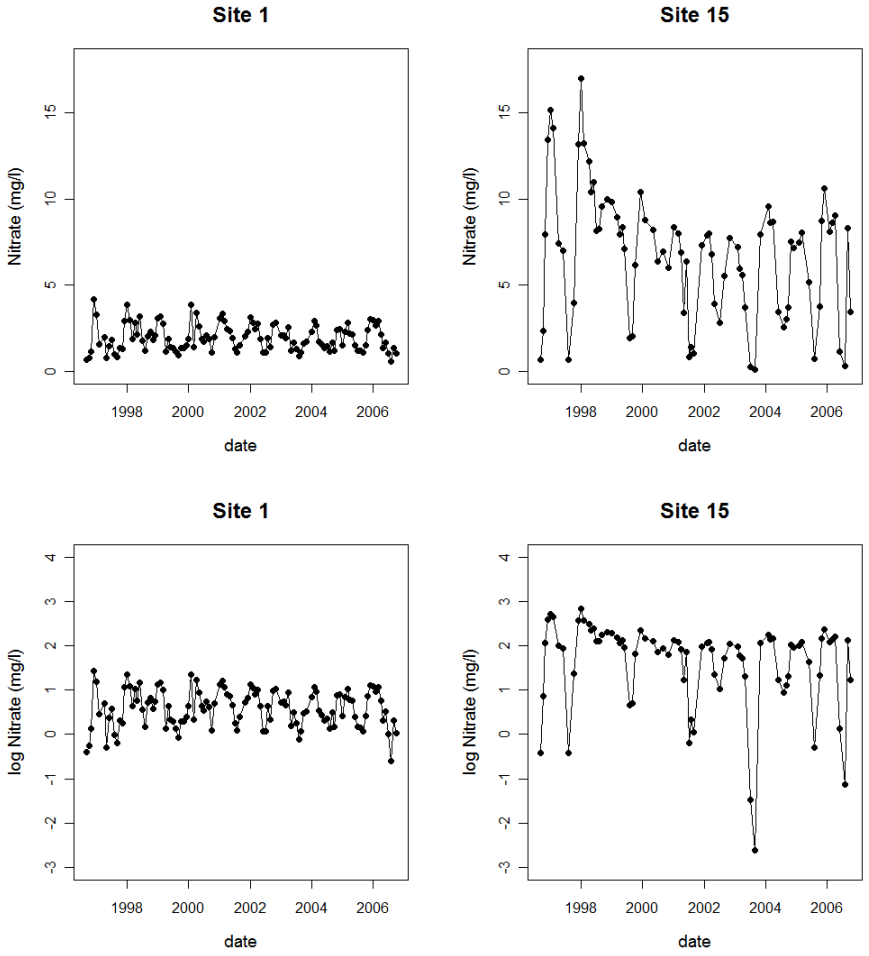
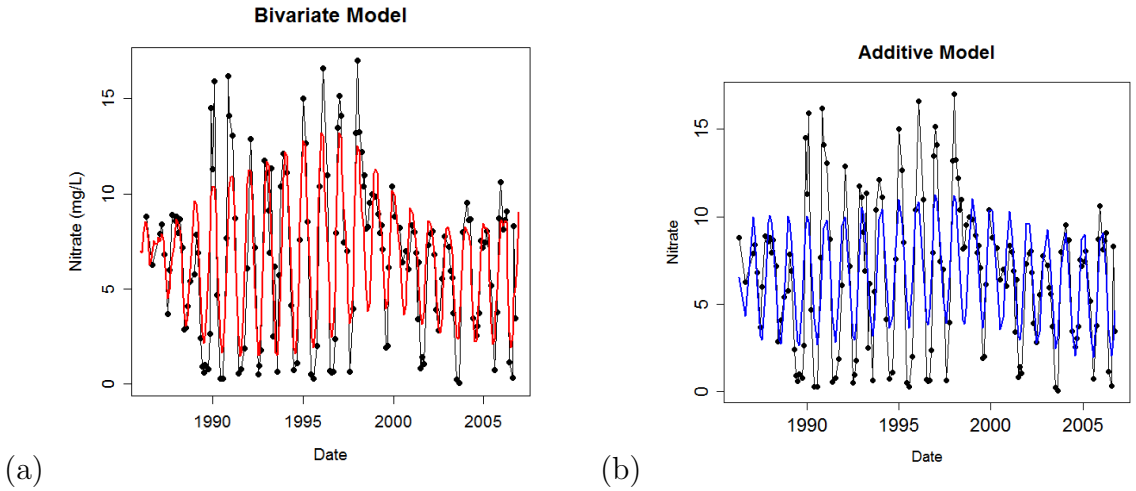


FIGURE 5.6: Observed (a) and log transformed (b) nitrate at Tweed stations 1 and 15

models were similar to those fitted in the varying seasonal simulations in Chapter 2 where the models along with the approximate F-test procedure are discussed in more detail. The presence of temporal correlation in the Tweed data at individual stations was investigated, but there was little evidence of this. This is unsurprising as the data is at a monthly frequency.

For station 15, results from the approximate F-test indicated that the bivariate model was most appropriate and so there did appear to be a change in the seasonal pattern over time. Figures 5.7(a) and 5.7(b) show the observed data from station 15 with the fitted models. It is clear that the bivariate model is more suitable as the flexible amplitude of the seasonal component captures more of the features of the data. The smoothing parameters used for each model were selected by fixing the degrees of freedom at 12 for each of the models. The results of the approximate

F-test for station 15 were typical of other stations which also displayed evidence of a change in the seasonal pattern over time.



(a) (b)

FIGURE 5.7: Plot of nitrate concentrations as Tweed station 15 showing fitted bivariate (a) and additive (b) models

The aim of our analysis is to obtain groups of stations which are similar in terms of mean levels of the determinand of interest, while taking into account any long-term trends and seasonal patterns present. After exploratory investigation of the Tweed stations, it was decided that taking the log transform of the nitrate data often 'dampened' features of the data which were thought to be of most interest in distinguishing different groups of stations, such as changes in the seasonal pattern over time. In addition, while there are differences in mean levels of the observed data at different stations, unlike the lakes data explored in Chapters 3 and 4, these differences are not particularly extreme and so it seems reasonable to compare the stations without transforming the scale of the data. Consequently, all further analysis on the Tweed data have been carried out using the raw data as it was felt this would produce more accurate groupings for this dataset.

Nitrate Vulnerable Zones

Within the River Tweed network there are areas which are designated as 'Nitrate Vulnerable Zones' under the European Union Nitrates Directive (European Parliament, 1991). The primary aim of the Nitrates Directive is to prevent, or reduce water pollution which is either induced or caused by nitrates from agricultural

sources, by controlling land-use management (SEPA, 2009). SEPA identifies regions of land as being nitrate vulnerable if they drain into waters identified as being affected by pollution, or drain into waters which may potentially be affected by pollution if no action is taken to prevent this. Under the Nitrates Directive it is essential that management programmes are both established and implemented to deal with these areas of concern. Currently all Nitrate Vulnerable Zones in the River Tweed district are subject to action programmes. These measures are reviewed by SEPA and, where necessary, are revised every four years based on assessments of their effectiveness. Clearly with the data considered in this study there are stations in the network that can be distinguished from all others due to their potentially high nitrate concentrations. It is of interest to assess if the clustering techniques proposed will be able to identify these stations within the regions which have been designated as being nitrate vulnerable, and to investigate to what extent these stations differ from those located outwith these known areas of concern.

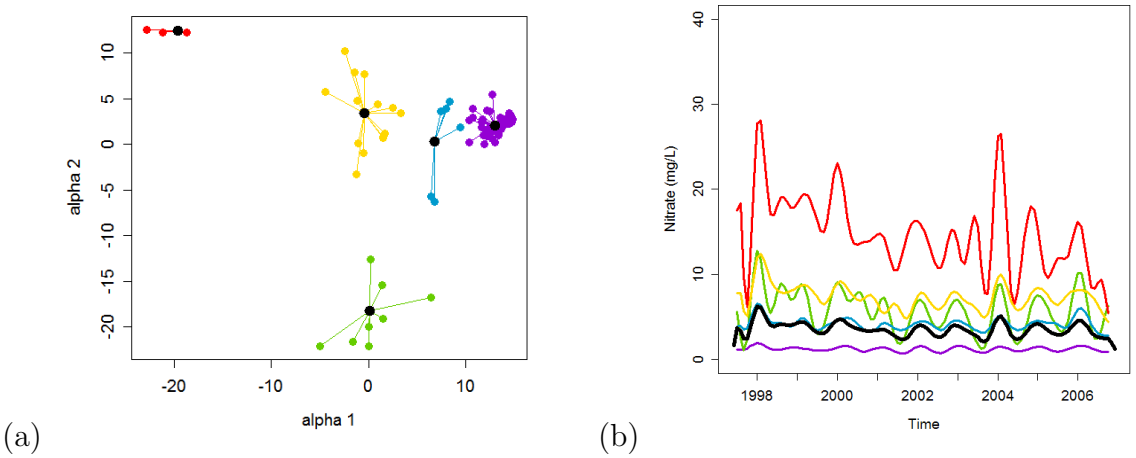
5.5.3 Clustering The River Tweed Data

This section investigates several different approaches to clustering the nitrate concentrations in the Tweed network using functional data analysis. An initial step in the investigation of whether or not there is any clear partitioning in the nitrate concentrations was to apply the functional clustering methods previously applied to the Scottish lakes data. Both the Functional Clustering Model (FCM) described in Equation 4.8 and hierarchical clustering based on distance defined by Equation 4.1 were considered. Although these methods assume that the stations are spatially independent it was thought that exploring the data using these approaches would provide a good first step to see what, if any, groups are present amongst Tweed monitoring stations. Following this, spatial covariance will be estimated and subsequently used to provide a set of weights within functional hierarchical clustering. The different clusters obtained using each of these approaches will then be compared.

Functional Model Based Clustering of the Tweed Data

The FCM was fitted using a natural cubic spline basis and a ridge parameter of 0.01, in a similar way to when this approach was used with the Scottish lakes data. The number of spline functions used was 30 and as before, this was selected using a sensitivity analysis. In order to select the statistically optimal number of clusters, the gap statistic was used. Model selection criteria such as BIC can be used to select the number of clusters for the FCM. However, it was felt that in order to compare the groups of stations obtained using the FCM and those using hierarchical clustering the same approach of determining the most suitable number of clusters should be used for all clustering methods considered. To calculate the gap statistic for the Tweed stations 500 simulated datasets were generated for the null reference distribution, which assumes there is no clustering present. This identified 5 groups to be the statistically optimal number. With this in mind, the FCM was subsequently fitted to the curves representing nitrate levels at the subset of 77 stations. The model was initially fitted with $h = 1$ however, the plot of projected curves appeared to lie in a curve, indicating that $h = 2$ was more suitable and so the model was re-fitted accordingly.

Figure 5.8 (a) shows the plot of projected cluster centres with the projected cluster means indicated by the black points, while Figure 5.8 (b) shows the estimated cluster mean curves. From Figure 5.8 (a) there appears to be a reasonable degree of separation between the clusters; the cluster represented by the red points appears to be the most distinct of the groups while the clusters represented by the blue and purple points seem to be closest together. The cluster membership probabilities for the majority of stations are high, with only three stations having membership probabilities corresponding to their predicted group of less than 0.9. It can be seen from Figure 5.8 (b) that the red points form the group with the mean curve which is far higher than any other group, while the blue and purple mean curves represent the stations with the lowest nitrate concentrations. The yellow and green curves overlap in terms of mean level, however stations which are included in the yellow group tend to have, on average, a weaker seasonal signal than those in the cluster represented by the green points. Although not shown, the predicted curves for each of the stations were also examined and for all the curves appeared to fit the data well.
In addition to the summary of the groups provided by the projected curves and estimated cluster means, Figure 5.9 shows the geographical distribution of the



(a) (b)

FIGURE 5.8: FCM projected curves (a) and estimated cluster means (b) for the River Tweed data

groups predicted using the FCM. It is clear that there is a strong spatial pattern displayed by the predicted clusters with stations which are in the same group often located close to each other. The area in the North-East of the Tweed Network has been identified as being a Nitrate Vulnerable Zone by SEPA and this has clearly been reflected within the clusters as concentrations of nitrate are far higher here than anywhere else. The stations which comprise the cluster represented by the red points on Figure 5.8 (a) are located close together within the Nitrate Vulnerable Zone. Not only is the mean level of the stations in this area much high than anywhere else, but the seasonal pattern is also much stronger and appears non-constant over time. In addition, these stations display evidence of a decreasing linear trend over the time period considered. This is possibly an indication that management plans which have been put in place by SEPA to reduce nitrate induced pollution are becoming effective. The stations which are in the South and West of the Tweed Network all appear to display relatively low nitrate concentrations with mean levels and seasonal patterns which are constant throughout time.

5.5.4 Fitting curves to the Tweed Data

Before applying hierarchical clustering techniques, the first step is to create a functional data object by fitting curves to the nitrate data at each station. There are however some problems due to the data being irregular and incomplete. The

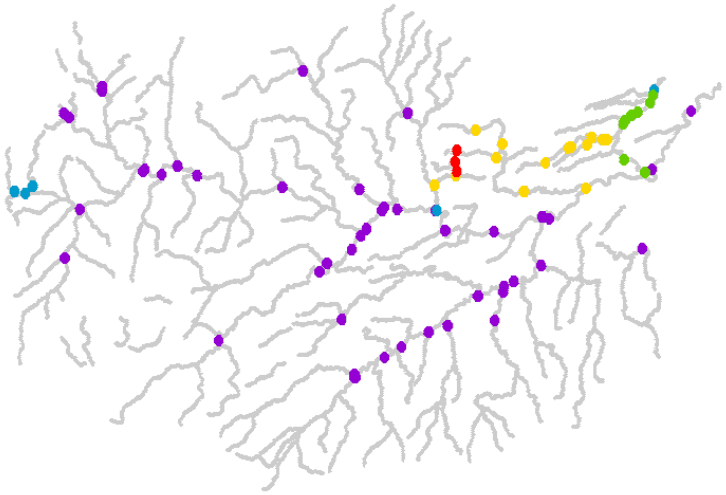


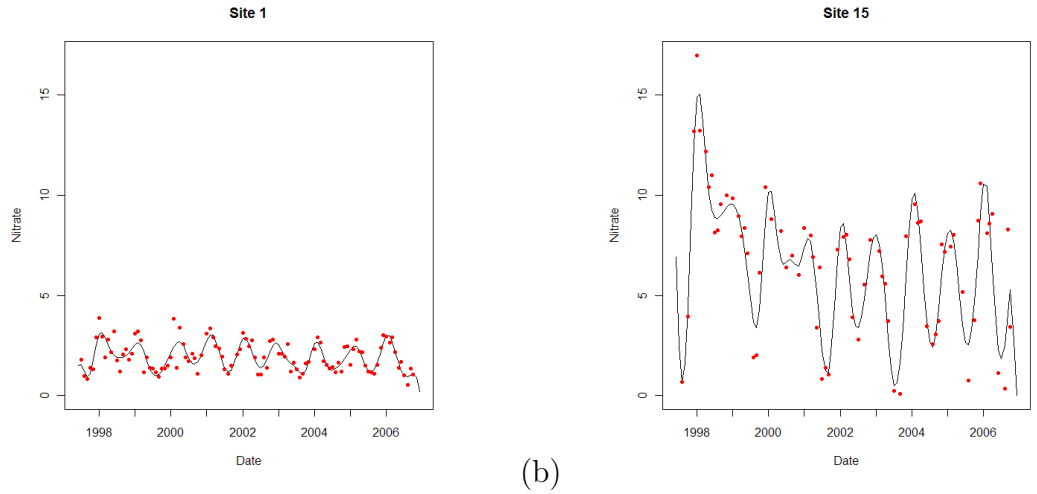
FIGURE 5.9: Map of River Tweed network showing FCM based clusters

extent of the irregularity in the data was not as pronounced as with the lakes data however, and it was thus decided that rather than fitting interpolating splines at each station, and then evaluating these at a regular grid of points to form a complete dataset, the relatively small number of missing values could be dealt with by using a ridge parameter. Using a ridge parameter proved to work well when fitting spline functions by least squares within the functional clustering model. As well as a ridge parameter a second order roughness penalty term was also used within the least squares estimation of the spline coefficients in order to ensure that the curves were not too locally variable and reflected the underlying pattern in the data accurately. If the equation for the i^{th} curve is written using the same notation as before, where $\mathbf{c}_i$ is the set of basis coefficients and $\Phi(t)$ is the spline basis function matrix then following from Equations 1.16, the estimated set of spline coefficients, $\hat{\mathbf{c}}_i$ can be estimated using least squares by,

$$\hat{\mathbf{c}}_\mathbf{i} = (\Phi(t)^T\Phi(t) + \Phi(t)\zeta + \lambda R)^{-1}\Phi(t)^T Y \tag{5.20}$$

where ζ is the ridge parameter, R is the roughness penalty matrix (defined in Equation 1.17) and Y is the matrix of observed values. A B-spline basis was used and both the number of spline basis functions, and the smoothing parameter which controls the effect of the penalty term, were selected using a sensitivity analysis. The number of spline basis functions chosen was 36 and the smoothing parameter was selected to be 0.01. As with the spline functions fitted using a B-spline basis with a roughness penalty in Chapter 4, the number of B-spline basis functions

corresponds to a knot placed approximately at 3 month time intervals. When using a ridge parameter in the estimation of the spline functions for the Scottish lake data a value of 0.01 was used, and this value also appeared to work well for the Tweed data. An example of the curves fitted is provided in Figure 5.10 which shows the observed data and curves that were fitted using the methods outlined above. It can be seen that the curves fitted are flexible enough to capture the main features of the data without being overly locally variable. After estimating sets



(a) (b)

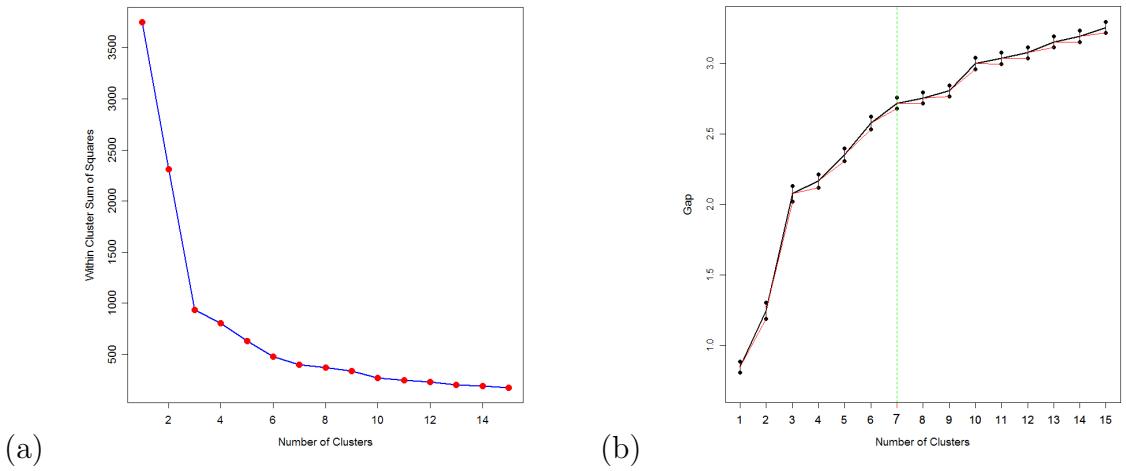
FIGURE 5.10: Fitted spline functions for nitrate data at Tweed stations 1 and 15

of basis coefficients which define each of the curves corresponding to the observed nitrate concentrations at each of the stations the next step was to use these to explore the potential presence of any groups of stations within the river Tweed network using hierarchical approaches.

Hierarchical Clustering of the Tweed Data

Using the basis coefficients which define the fitted curves in Section 5.5.4, the functional dissimilarity matrix given in Equation 4.1 was computed. As with the Lochs data, complete linkage has been used to determine the clusters. It should be noted that although different linkage measures were considered, the differences between the clusters defined by the different linkage methods were not very large. The gap statistic was used to determine the statistically optimal number of groups, again with 500 sets of data generated to create the null reference distribution, and from this the number selected was 7. Figure 5.11 (a) shows the 'L-curve' for the

Tweed hierarchical clustering with no spatial covariance incorporated and Figure 5.11 (b) shows the corresponding gap statistic plot. From the L-curve it appears that 3 groups would be most suitable, however, on inspection of the gap statistic, 7 is identified as being the most appropriate number of groups. There is evidence from the gap statistic curve that 3 groups is likely. This may suggest that there are 3 distinct clusters or 7 which are less well separated in terms of the pattern of nitrate concentrations over time.



(a) (b)

FIGURE 5.11: L curve (a) and gap statistic plot (b) for hierarchical clustering of River Tweed stations

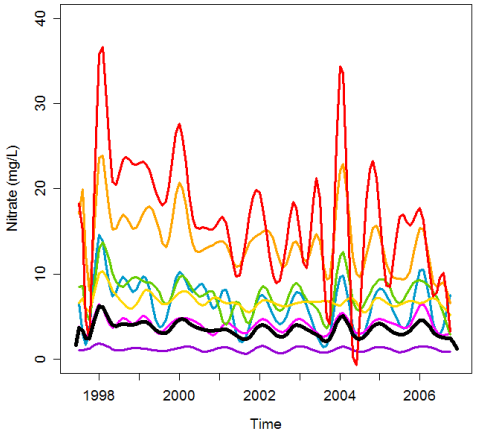


FIGURE 5.12: Cluster mean curves for Tweed nitrate data determined using hierarchical clustering (assuming no spatial covariance between stations)

Figure 5.12 shows the estimated cluster means (calculated using Equation 3.1) and Figure 5.13 shows a map of the Tweed network with the 7 different clusters

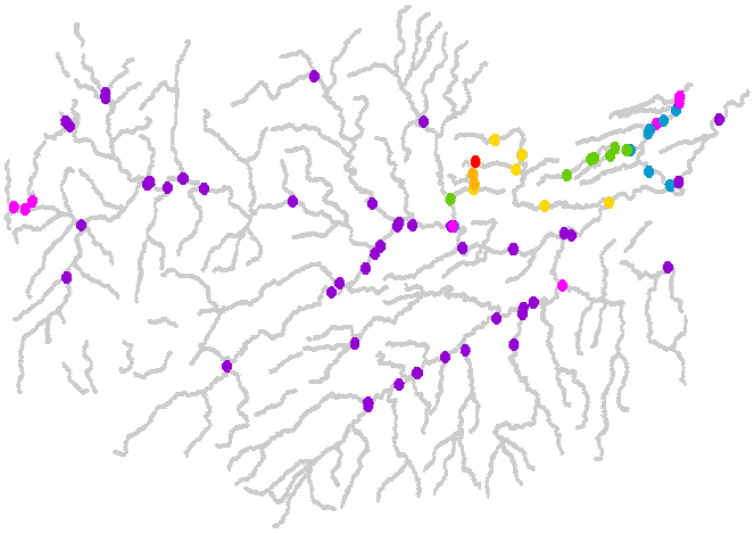


FIGURE 5.13: Map of River Tweed network showing clusters determined using hierarchical clustering (assuming no spatial covariance between stations)

of stations represented by different colours. The number identified from the gap statistic approach has been investigated here as the gap statistic approach for choosing the most appropriate number of clusters is thought to be more robust than that of the L-curve method. With the L-curve there is no reference distribution with which to compare the within cluster dispersion against the number of groups.

Although a larger number of groups was identified as being statistically optimal when hierarchical clustering was applied in comparison to when the FCM was applied, Figure 5.12 indicates that there is a considerable degree of overlap in the means of the groups identified using hierarchical clustering, particularly amongst the stations in the groups represented by the pink and green points/curves. Even though different numbers of clusters are identified, it is clear the same patterns are key in defining the different groups under both the FCM and hierarchical approaches. For the hierarchical clustering there is one single station which forms a group/cluster. This station, which is identified by the red point/curve, has a notably higher concentration over time than any of the other stations and is located in the region which is known to be a Nitrate Vulnerable Zone. The hierarchical clustering also identifies groups which highlight the geographical pattern in the nitrate concentrations over the network, with a large group of stations identified as having a low concentration and a small, constant seasonal signal in the South and West of the network. The North-East of the network displays the largest

amount of variability in terms of the mean functions of each of the groups, where as could be expected, stations closest to the Nitrate Vulnerable Zone have the highest average nitrate concentrations, although these levels decrease the further downstream you go from the high concentration area.

A cross-classification table for the hierarchical clustering of the River Tweed stations and the FCM based clusters is provided in Table 5.1. It is clear there is a large amount of agreement between the two sets of clusters in terms of the largest group of stations. In addition it can be seen that all members of hierarchical groups 4 and 5 are contained within FCM group 5.

No Spatial FCM	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
1	45	0	1	0	0	0	0	46
2	0	6	2	0	0	0	0	8
3	0	0	0	0	0	2	1	3
4	0	0	6	0	0	0	0	6
5	0	1	0	7	6	0	0	14
Total	45	7	9	7	6	2	1	77

TABLE 5.1: Cross-Classification table for Hierarchical functional clustering of Tweed Stations with no spatial weights and with FCM clusters

Based on the results of the FCM and the hierarchical clustering approaches it is clear that there does appear to be distinct groups of stations on the Tweed network which are split not only in terms of their mean level but which are also different in terms of the strength of the seasonal pattern over time. However, as with the lakes data the mean level is the driving factor underlying the distinctions between the different groups. It is also apparent that the number of groups may be overestimated using methods that ignore the spatial aspects. In view of the spatial distribution of the clusters for the two approaches using Figures 5.9 and 5.13 there is evidence of some spatial trend, with the Nitrate Vulnerable Zone located in the North-East being particularly different from the rest of the area in terms of the mean concentration. This potential inconsistency in the mean level brings into question the assumption of second order stationarity and the underlying mean of the spatial process may not be constant over the entire geographical region covered by the Tweed. In order to estimate the covariance structure it is important to first ensure the required assumption of stationarity is met and so the next section will investigate de-trending the Tweed data.

5.5.5 Estimating spatial covariance in the Tweed

The presence of longitudinal and latitudinal spatial trends on the Tweed can be explored separately in Figure 5.14 which shows the average nitrate level at each station plotted against their corresponding latitude and longitude. The red line on these plots is a loess curve (Section 1.3.2) which has been added to indicate the general trends and patterns in the station averages. Looking at the longitudinal trend in Figure 5.14 (a), it can be seen that in the middle of the area covered by the Tweed the nitrate levels are fairly low with a collection of stations in the West and East which display higher values. From Figure 5.14 (b) it can be seen there is a collection of stations towards the North of the region where the nitrate levels are high, however stations in the South and furthest North have relatively low levels. These patterns correspond to those identified in the initial clustering of the Tweed stations considered in the previous section. This initial investigation of nitrate

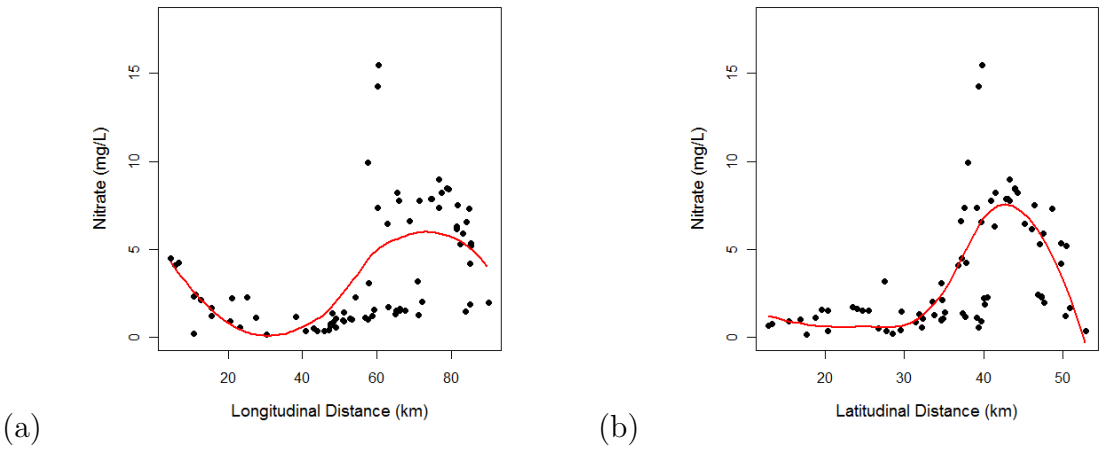


FIGURE 5.14: Plots of mean nitrate against geographical location; (a) Longitude, (b) Latitude

levels indicated that it is unlikely that the assumption of stationarity holds and so it is necessary to remove the long-term spatial trend in nitrate levels across the network before estimating the covariance structure. It is also clear from Figure 5.14 that fitting a parametric trend would not be adequate to describe the spatial patterns in the data, hence, a simple non-parametric trend was estimated using the 'mgcv' library in R. Using this package, a bivariate smooth trend was fitted to the station means using thin plate splines with a smoothing penalty applied to ensure the estimate of the trend retained the key features of the data without being overly sensitive to small changes. The model fitted was of the form shown in

Equation 1.4 where x_1 and x_2 represent the geographical co-ordinates (Latitudinal and Longitudinal distance) of the stations. A basis of thin plate splines are used to fit the model. Thin plate splines are described by Wood and Augustin (2002) as "the natural generalisation of cubic splines to any number of dimensions and almost any order of wiggliness penalty". Thin plate splines are isotropic smoothers, meaning that any rotation of the covariate co-ordinate system will not change the result of smoothing. In addition, like B-splines, thin plate splines are low-rank, which means that the number of coefficients to be estimated is far fewer than the number of observed data points. Comprehensive details of thin plate spline smoothing techniques are provided in Wood (2003).

Station means were used as the response in order to provide an estimate of the spatial trend over the time period of interest and smoothing parameters were selected using a sensitivity analysis. Functional de-trending of the data is possible and could be achieved by fitting a functional regression model with a functional response (functions of nitrate over time at each station) and a bivariate scalar covariate (the co-ordinates of the station) however it was felt that estimating the trend in this way introduced unnecessary complexity to the spatial trend estimate as the aim is to obtain a simple estimate of the trend in the nitrate levels so that the assumption of second order stationarity is met and the spatial covariance structure can subsequently be estimated.

Figure 5.14 shows the estimated spatial trend for the Tweed network as well as the observed station mean nitrate levels, which are indicated on this plot by the red points. It is clear that the estimated surface provides a good fit to the points and captures the main features of the trend. The initial impressions of the features of the spatial trend when considered in separate directions in Figure 5.14 are reinforced here and the key feature is the presence of an area of high nitrate levels in the North East. The rest of the region displays average nitrate levels which are fairly constant and relatively low. Although it may initially be concerning that the the estimated surface unrealistically falls below zero towards the South-East of the Tweed region, this is an area where there are almost no stations and hence no data have been collected here. It is possible to constrain the smooth surface to be positive, however this did not seem necessary in this situation. Figure 5.16 shows the de-trended station mean values, with a zero line shown in blue for reference. While comparing Figure 5.16 to Figure 5.14 it can be seen that although there has been a reasonable degree of improvement in terms of removing the spatial trend,

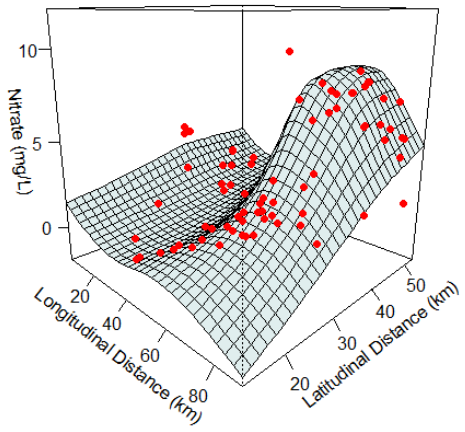
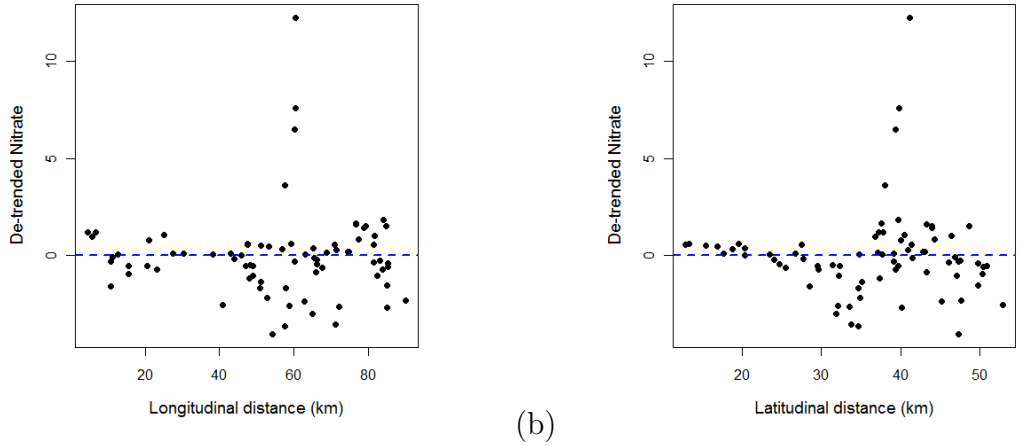


FIGURE 5.15: Estimated spatial trend for nitrate levels on River Tweed network

there continues to be some evidence that a small subset of stations located in the North and East of the network have higher average nitrate concentrations than those elsewhere. In view of this, it could be argued that there is not actually any strong spatial trend in the Tweed nitrate data, but there is in fact a discontinuity in the nitrate levels in this region. Other than the region which displays high mean levels there is almost no evidence of a spatial trend and the mean level across space seems flat, indicating that the assumption of stationarity holds outwith this small area. In fact, while there is no land-use data available surrounding the Tweed river network, it is known that the area of high nitrate levels in the North-East is in a Nitrate Vulnerable Zone. This is a region which is predominantly comprised of farming areas, while the low concentration nitrate areas are often found in more upland areas. It is possible that if suitable land use data was available, then this would be better at explaining the trend in the data rather than simply using geographical co-ordinates. Furthermore, additional exploration of the trend surface for the Tweed network could investigate the presence of discontinuities. It was thought worthwhile to estimate spatial covariance for both the original and de-trended nitrate data and the estimated trend shown in Figure 5.15 will be used for the latter estimation. While it is possible to estimate the trace-variogram using Euclidean distances, for the reasons previously discussed, when estimating stream distance based spatial covariance using the tail-up model (Equation 5.11) a covariogram should be computed to ensure a valid covariance matrix is obtained. Furthermore, in order to ensure comparisons between the clusters of stations obtained using spatially weighted hierarchical clustering are fair, it was thought that the functional dissimilarity matrix should be weighted by the same measure of



(a) (b)

FIGURE 5.16: De-trended nitrate data on River Tweed; (a) Longitude, (b) Latitude

the spatial association between stations regardless of the metric that was used to measure the distance between stations (Euclidean or stream). Using the same measure of spatial dependence for the two different distance metrics ensures that there is no disparity between the different scales of the covariograms. For this reason, for both Euclidean and stream-based distances, a functional covariogram was estimated using the procedure outline in Section 5.4.1.

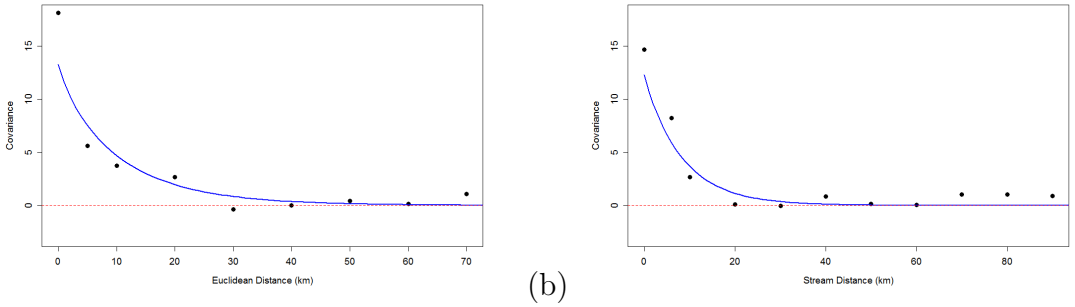
Curves were fitted to both the original, and the de-trended data at each station using the method described in Section 5.5.4. The empirical functional covariogram was next estimated using the functional covariances calculated using Equation 5.18 and the distance between stations, either Euclidean or stream-distance. The points were 'binned' at 10km intervals, with an additional point estimated at 5km since so many of the stations were separated by short lags. The covariogram was estimated up to a maximum distance of 70km since over 90% of the paired station Euclidean and stream distances were less than this. A Matérn covariance function was fitted to these empirical covariograms. As discussed earlier, weighted least squares was employed to choose the parameters θ and ν (Equation 5.3). A fine grid of different combinations of possible parameters were investigated.

Four different functional covariograms were estimated. For both the original and de-trended nitrate data a covariogram was estimated using the Euclidean and stream based distances between stations. Figures 5.17 and 5.18 shows each of these functional covariograms with the fitted Matérn covariance functions shown in blue. The parameter values of these fitted Matérn functions are shown in Table 5.2. As

	ν	θ	effective range, km ($h_{0.95}$)
Euclidean, Original	13.0	0.30	31.4
Euclidean, De-trended	2.6	3.10	17.1
Stream, Original	8.8	0.45	25.3
Stream, De-trended	4.2	0.80	15.4

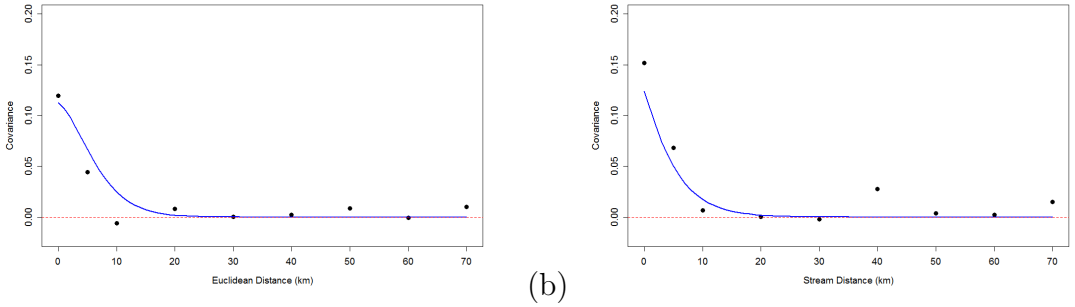
TABLE 5.2: Fitted Matérn covariogram parameter estimates

the Matérn family covariance functions have an infinite range and approach the sill asymptotically, the effective range, $h_{0.95}$, as defined in Cressie (1993) has been calculated (in addition to the value of the parameter θ). This is the distance which corresponds to 5% of the maximum covariance, $Cov(0)$.



(a) (b)

FIGURE 5.17: Estimated and fitted covariograms for original Tweed nitrate data. The fitted Matérn covariance functions are shown in blue. ((a) Euclidean (b) Stream)



(a) (b)

FIGURE 5.18: Estimated and fitted covariograms for (de-trended) Tweed nitrate data. The fitted Matérn covariance functions are shown in blue.((a) Euclidean (b) Stream)

From Table 5.2 it can be seen that for both the Euclidean and stream distance metrics the covariograms for the original data have a far greater effective range than those for the de-trended data. This could be expected as de-trending was carried out in an attempt to remove the long-range spatial trend across the river

network. However, even after the removal of the spatial trend there continues to be evidence of spatial correlation between the stations. For stream distance, the covariogram fitted indicates that after de-trending, flow-connected stations are spatially correlated until they are separated by a distance of more than 15.4km. For Euclidean distance it is thought stations separated by a distance of more than 17.1km will be spatially uncorrelated. The covariances estimated for stream and Euclidean distances are very similar in terms of both the shape of the covariogram estimated, and in terms of the effective ranges. This is particularly true for the de-trended data where the effective ranges are different by a relatively small distance of 2.3km.

Using the estimated covariograms a set of covariance matrices were obtained. In line with Equation 5.11, to ensure the covariance matrices were valid, the stream distance matrices obtained were multiplied through by the square root of the flow weights. In the next section the covariance matrices obtained using the functional covariograms were used to weight the functional dissimilarity matrix as shown in Equation 5.19.

5.5.6 Spatial Functional Clustering Approaches

For each of the clustering methods, the statistically optimal number of clusters as determined using the gap statistic with 500 reference distributions is shown in Table 5.3. It can be seen that for both Euclidean and Stream distance weighted hierarchical clustering fewer groups are optimal than when no spatial weighting is applied. Also, for each distance metric, the number of groups chosen using the de-trended and original data was the same. The results of the hierarchical clustering approaches with these optimal numbers of clusters will now be discussed in more detail.

Figures 5.19 and 5.20 display maps of the clusters for the de-trended Euclidean and Stream based covariance and the corresponding plots of the cluster mean curves. On each of the cluster mean plots the black solid line represents the overall mean for all stations. The clusters from the Euclidean distance weighted clustering were identical regardless of whether or not the original or de-trended data was used. This is not the case for stream distance weighted clustering, where although the number of groups selected as being most suitable was the same for the de-trended and original data, the stations contained in each of the groups are

slightly different. The results of the original stream distance covariance weighted clusters are shown in Figure 5.21. It is clear there are similarities between all clustering approaches considered; the Nitrate Vulnerable Zone in the North-East of the region is clearly distinct from the other areas.

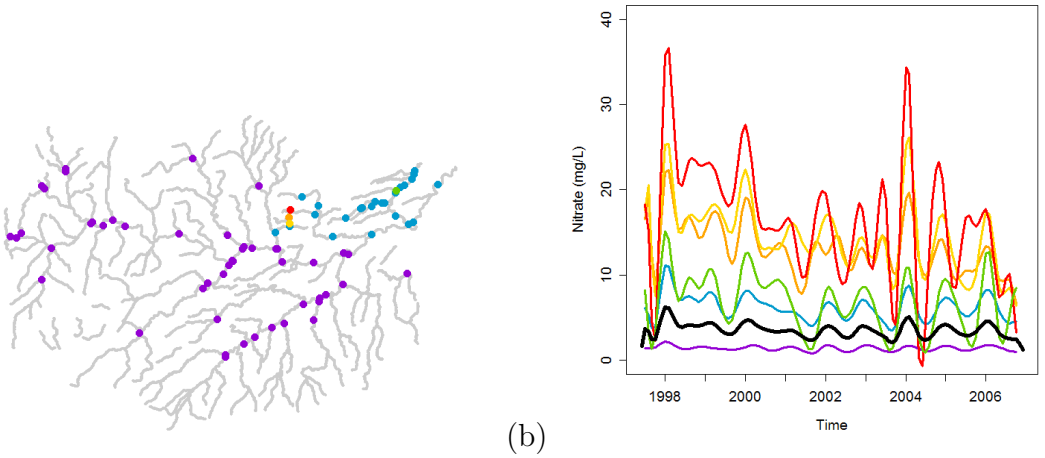
The results of the clusters which are weighted by Euclidean distance based covariances indicate that the 6 groups identified are effectively 2 large groups, and 4 individual stations which are distinct from these 2 groups, and one another. From Figure 5.19 it can be seen that the cluster represented by the purple points is made up of a large number of stations which all have relatively low nitrate concentrations and display very little evidence of a seasonal signal. The cluster represented by the blue points is comprised of stations which have, on average, both a mean nitrate concentration across time which is higher than the overall average and a seasonal signal which is moderate. The 4 stations which do not fall in these groups all have mean levels which far exceed the overall average, and display seasonal signals which are exceptionally strong. There is very little evidence however of a difference between the stations represented by the orange and yellow points and both seem very similar to one another.

In comparison to all other clustering methods investigated, fewer clusters were identified as being optimal when using stream distance based spatial covariance to weight the dissimilarity matrix. The gap statistic indicated that only 3 groups were required in order to adequately capture the differences in the nitrate concentrations amongst the stations. For both the covariance based on the de-trended data and the original data it can be seen from Figure 5.20 and Figure 5.21 that there is one group of low concentration, low seasonal signal stations (shown in purple), one group of moderate concentration, moderate seasonal signal stations (shown in blue) and one group of higher concentration, higher seasonal signal stations (shown in green). As mentioned earlier there are differences between the clusters found using the de-trended and the original stream distance covariances. When the de-trended covariance is used there are only 3 stations in the high concentration group, and as shown in the cluster means there is far more disparity between the moderate and the high group when compared with the results based on the original stream distance data.

Cross-classification tables for the three pairwise combinations of hierarchical functional clustering with no spatial weight, de-trended Euclidean spatial weights and de-trended Stream distance based spatial weights are displayed in Tables 5.5,

Method	Number of Clusters
FCM, no spatial	5
Hier, No spatial	7
Hier, Euclidean, original	6
Hier, Euclidean, de-trended	6
Hier, Stream, original	3
Hier, Stream, de-trended	3

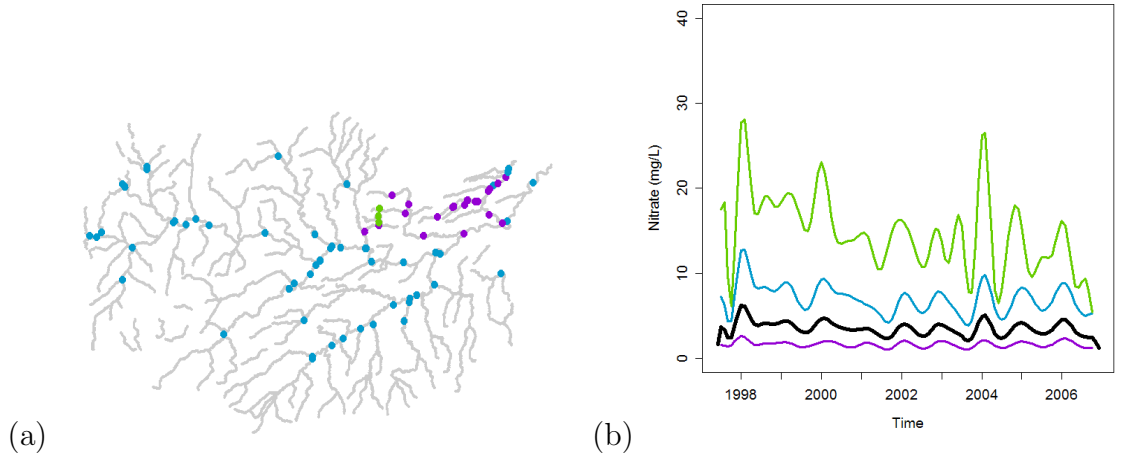
TABLE 5.3: Number of clusters for functional nitrate data chosen using the gap statistic



(a) (b)

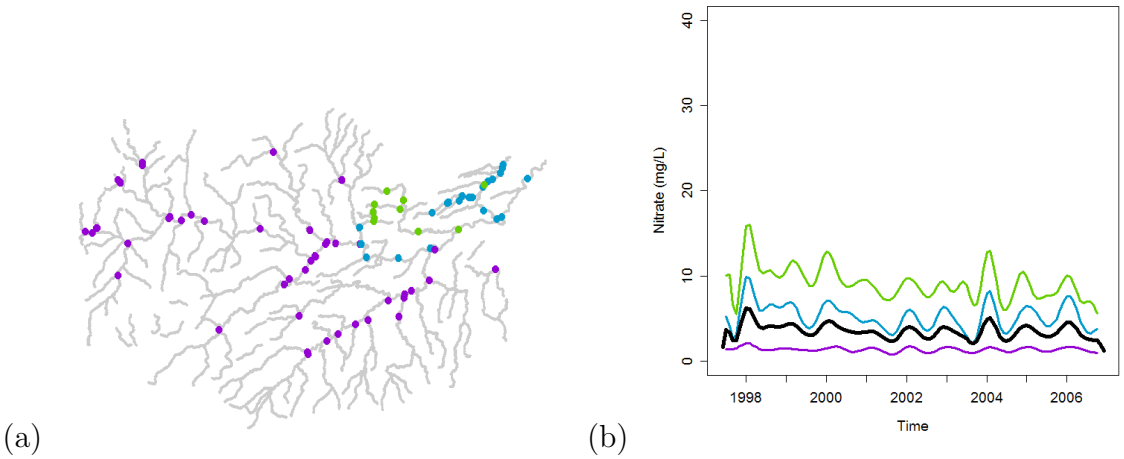
FIGURE 5.19: Plots showing de-trended data Euclidean covariance weighted hierarchical clustering results. (a) Tweed network showing different groups, (b) Group mean curves

5.5 and 5.6. In each of the pairs it can be seen that there is general agreement between the largest groups which consist of stations that have relatively low mean levels of nitrate. It is also of interest to note that three of the four stations identified as being in clusters with only one station under de-trended Euclidean distance weighted hierarchical clustering are grouped together under stream distance weighted clustering. Further comparisons of the agreement in the partitions based on different clustering approaches are provided in the next section.



(a) (b)

FIGURE 5.20: Plots showing de-trended data stream distance covariance weighted hierarchical clustering results, (a) Tweed network showing different groups, (b) Group mean curves



(a) (b)

FIGURE 5.21: Plots showing stream distance covariance weighted hierarchical clustering results. (a) Tweed network showing different groups, (b) Group mean curves

Detrend Euc No Spatial	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
1	2	43	0	0	0	0	45
2	6	0	0	0	0	1	7
3	4	5	0	0	0	0	9
4	7	0	0	0	0	0	7
5	6	0	0	0	0	0	6
6	0	0	1	1	0	0	2
7	0	0	0	0	1	0	1
Total	25	48	1	1	1	1	77

TABLE 5.4: Cross-Classification table for Hierarchical functional clustering of River Tweed Stations with no-spatial weights and with de-trended Euclidean distance based spatial weights

Detrend Stream No Spatial	**1**	**2**	**3**	**Total**
1	45	0	0	45
2	0	7	0	7
3	9	0	0	9
4	0	7	0	7
5	0	6	0	6
6	0	0	2	2
7	0	0	1	1
Total	54	20	3	77

TABLE 5.5: Cross-Classification table for Hierarchical functional clustering of River Tweed Stations with no-spatial weights and with de-trended Stream distance based spatial weights

Detrend Stream Detrend Euc	**1**	**2**	**3**	**Total**
1	6	19	0	25
2	48	0	0	48
3	0	0	1	1
4	0	0	1	1
5	0	0	1	1
6	0	1	0	1
Total	54	20	3	77

TABLE 5.6: Cross-Classification table for Hierarchical functional clustering of Tweed Stations with de-trended Euclidean distance spatial weights and de-trended Stream distance based spatial weights

5.5.7 Comparing the Partitions

Although there is no single number of clusters identified as being statistically optimal, there are undoubtedly broad similarities between the results of all of the different clustering methods investigated here. The key feature common to all clustering methods, whether spatial correlation is incorporated or not, is the identification of stations within the Nitrate Vulnerable Zone as being distinct to the stations outwith this area. However, the difference between each of the clustering approaches is the extent to which the partitioning of the stations within the Nitrate Vulnerable Zone differ from one another. For Euclidean distance and spatially independent functional hierarchical clustering approaches the number of groups identified as most appropriate using the gap statistic results in a partition of the stations such that stations within the North East of the region are assigned to different groups. These groups are often small, or consist of only a single station. The gap statistic for stream distance weighted clustering suggests there are only three groups, and for the de-trended stream covariance weighted clusters, there are only 3 stations which are identified as having markedly higher nitrate concentrations in comparison to all the other stations.

In order to attempt to quantify how similar the results of each method were the Adjusted Rand coefficient (Hubert and Arabie, 1985) has been computed for all pairs of clustering approaches. The Rand coefficient (Rand, 1971) is a measure of agreement between two partitions. It is an index which is based upon counting the pairs of points on which two clusterings agree or disagree. An extension of this is the Adjusted Rand Index (ARI) which was developed in Hubert and Arabie (1985) and is corrected for the possibility that agreement between two sets of clusters may simply be due to chance. Milligan and Cooper (1986) evaluated several different methods for measuring the agreement between different sets of clusters and recommended the use of ARI as an index for comparing clusters which performed well. Table 5.7 shows the ARI for all pairs of clusters obtained for the Tweed data. The maximum value of the ARI is one, which corresponds to perfect agreement between two partitions. Conversely, if the ARI is zero, the two partitions are mutually independent.

There is clearly strong agreement between the sets of clusters obtained from the two methods where the stations are assumed to be spatially independent.

In addition, there is reasonably strong agreement between the hierarchical partitions of the stations which incorporate de-trended Euclidean and de-trended stream covariance (ARI=0.71). The results indicate that although different numbers of groups are identified as being statistically optimal, there is actually not a huge difference between the results obtained using the original and de-trended data and the two different distance metrics to estimate spatial covariance. This is perhaps unsurprising since there were strong similarities between the stream and Euclidean distance based covariograms. It should be noted however that although the covariograms look similar, the stream and Euclidean covariance matrices will differ due to the effects of the flow-weight and flow-connectedness, which are incorporated into the stream covariance matrix, but do not have any influence on the Euclidean covariance matrix. Using the ARI to compare the methods which assume spatial independence to those with spatial covariance incorporated it can be seen that there is moderate agreement between the partitions determined by the FCM and hierarchical clustering with de-trended spatial covariance based weights (ARI=0.68). There is however slightly less agreement between the spatially weighted clusters and the hierarchical clusters which assumed independence where the ARIs are 0.61 (Euclidean) and 0.64 (stream).

It should be noted, however, that the ARI should be viewed with caution and should not be considered without also comparing the distribution of the partitions via either cluster means or geographical maps. An example of why the ARI should not be viewed in isolation in order to compare partitions is the value computed for the ARI between hierarchical clustering with original scale stream distance based covariance and de-trended stream based covariance (ARI=0.54). In view of the similarities between the results of the two clustering methods in terms of the number of clusters and the cluster means (shown in Figure 5.20 and 5.21) this relatively low ARI is possibly unexpected. One possible explanation of this is the small number of clusters and the fact that the number of clusters is the same for both approaches being compared. This means that any disagreements are likely to have a large effect in the computation of the ARI. The small number of clusters, combined with the relatively small number of lakes meant that the ARI was not optimal for quantifying agreement between the different partitions identified in the investigation of the lakes data in Chapter 4.

	FCM	**Hier, NS**	**Hier, Euc**	**Hier, Str**	**Hier, Euc (d)**
Hier, NS	0.90				
Hier, Euc	0.68	0.61			
Hier, Str	0.57	0.56	0.73		
Hier, Euc (d)	0.68	0.61	1.00	0.73	
Hier, Str (d)	0.68	0.64	0.71	0.54	0.71

TABLE 5.7: Adjusted Rand Index for partitions obtained using different clustering approaches

5.6 Summary

This chapter has discussed how spatial covariance can be estimated for functional data and how this estimate of the spatial relationships between locations can subsequently be incorporated into hierarchical clustering techniques.

The functional data considered were from stations situated along a river network and so two different distance metrics were investigated; standard Euclidean distance and stream distance which takes into account the unique features of river network data. The use of stream distance to estimate covariance introduced several additional complications which were required to be addressed in order to ensure the stream distance covariance matrix was valid. It was decided when using stream distance that the tail-up model was suitable for the Nitrate data on the Tweed. This model incorporates both flow weight and flow-connectedness and assigns a covariance of zero to stations which do not flow into one another. The structure of the tail-up model however precipitates the need for a single value measure of covariance between curves so that a covariogram can be estimated. Using the standard measure of covariance and ideas from the trace variogram, covariance was defined using areas between pairs of curves. This is the first time that the tail-up model has been applied to functional data on a river network using stream distance.

After estimation of both Euclidean and stream distance based covariance matrices these were used as weights within hierarchical functional clustering techniques to develop methods which have already been employed in previous literature. In addition to the hierarchical techniques, the functional clustering model already discussed in detail for the lakes data was also applied to the river network data. Using the gap statistic, the smallest number of clusters identified as

statistically optimal was using the hierarchical clustering weighted by stream distance, whilst the largest was for hierarchical clustering which assumed the stations were spatially independent. Although there is no single correct number for how many groups is best, the stream distance covariance weighted clusters were all distinct in terms of their cluster means. There was no overlap in the means and the clusters appeared to be internally homogeneous. For the other hierarchical approaches there appears to be a great deal of overlap in the cluster means, which suggests that hierarchical methods which do incorporate spatial correlation are best. All clustering approaches identified the presence of a Nitrate Vulnerable Zone in which stations have a markedly higher mean concentration as well as a much higher seasonal pattern than the other stations. The number of clusters of stations within the Nitrate Vulnerable Zone, however, is where the methods differ. The de-trended stream distance covariance weighted clustering identified only one group of stations whilst the Euclidean distance approaches identified four groups consisting of either one or two stations each.

The functional clustering model appeared to provide a good estimate of the underlying patterns of nitrate over time and identified a moderate number of groups which was greater than the number used with stream distance based covariance but less than that used with hierarchical clustering with no spatial weights. Even without incorporating any measure of spatial correlation between stations, the model based clustering approach performed well in terms of separation of groups and the cluster membership probabilities were all relatively high. It would be interesting to investigate the potential for the functional clustering model to incorporate measures of spatial covariance. Due to the structure of this model based approach where sets of basis coefficients are clustered rather than stations it is likely that the spatial covariance matrix incorporated within the functional clustering model would need to be defined in terms of covariance between individual spline coefficients. The use of river network data and stream distance would add further complications with the unique difficulties of also requiring flow-connectedness and flow weights to be defined in terms of the spline coefficients.

The different approaches taken were compared by looking at the partitions of the stations and the different cluster means under each of the methods. The Adjusted Rand Index was also used to quantify the agreement between each pair of partitions, however it was noted that this should only be used in combination with other descriptions of the results such as maps which show the geographical

distribution of the clusters. The ARI indicated that there was reasonably strong agreement between the Euclidean and stream distance spatially weighted hierarchical clusters.

Chapter 6

Conclusions, Discussions and Future Work

The overall aim of this thesis has been to employ statistical analysis to evaluate the design and efficacy of commonly used environmental monitoring programs that are used to inform evidence based policy such as the WFD. Methods have been applied and developed to investigate how sites are grouped for classification. The techniques considered have been extended to the multivariate setting so that multiple determinands can be incorporated simultaneously within the formation of the groups. Existing techniques for grouping locations on river networks have also been adapted to include spatial correlation, if required. In the examples presented, particular focus has been placed on monitoring networks for water quality in both lakes and rivers, but the techniques explored could potentially be applied in many environmental contexts.

6.1 Assessing Statistical Power to Detect Change

The simulation study presented in Chapter 2 was motivated by the importance of being able to detect underlying changes in environmental data. It was designed in order to provide an insight into the likely effectiveness of current monitoring programmes, and to act as a useful guide to the relative power associated with different patterns and forms of underlying change. The results of the simulation study in Chapter 2 indicated that resources may be best used to sample at a reduced number of locations on a more frequent basis in order to detect change with

a suitable level of power. For many of the forms and magnitudes of change investigated within this thesis it was found that monthly sampling, which is the frequency most commonly used for monitoring water bodies for WFD classification, was often inadequate in terms of the power to detect such changes. Many practitioners will not appreciate the length of time series required for an appropriate power with monthly data.

For the most simple scenario, where the change was a a constant decrease over time it was found that after around 15 or more years of data, weekly and fortnightly sampling generally worked well in terms of detecting fixed linear changes, even when those changes were relatively small. If the underlying trend in the data was strong, monthly sampling also reached an adequate level of power with around 15 years of data. However, with less frequent sampling such as bi-monthly and annual sampling, time series in excess of around 20 years are required to detect even relatively large changes of 10 and 20% each year. For non-linear patterns weekly and fortnightly sampling frequencies again performed well, while annual sampling was noticeable in its inability to detect underlying change, even when the time series of data considered were relatively long.

The results of the varying amplitude simulation indicated that even when there is a large change in the amplitude of the seasonal signal, corresponding to a 60% reduction, that monthly sampling often does not have a sufficient level of power to detect this change. The ecological consequences of changing seasonal signals are potentially far reaching since changes in the seasonal pattern of a single determinand can affect an entire ecosystem. It is therefore important to employ a sampling programme which can detect such changes in seasonal signals. This simulation study presented in this thesis has provided evidence that while current sampling programmes may be capable of detecting pronounced linear increases or decreases in the concentration of a determinand of interest over time, they may fail when trying to detect more subtle or complex features of the underlying systems.

6.2 Grouping Sites for Monitoring

In view of conclusions drawn from the simulation study on ability to detect long term change it is clear there is a need for increased levels of sampling in some

situations. However, there are often limitations on resources that prevent continuous monitoring of all stations of interest within a monitoring network. There was consequently some question of where the most suitable locations to implement increased monitoring should be. The grouping feature of the WFD means that fewer sites than before can be sampled in order to produce chemical classification of water bodies, although it is vital that the groups defined are appropriate and internally homogenous.

Existing SEPA Groups

The investigation of the groups of lakes currently used by SEPA for WFD classification presented in Chapter 3 indicated that the current group structure, which is based on broad categories of altitude and alkalinity, rather than observed determinands of interest, does perform reasonably well in capturing the variability of the lakes. Using penalised regression splines provided a computationally efficient and flexible way to estimate the true functions underlying the observed data and, after imputing the small quantity of missing data by first fitting interpolating splines, the functions fitted using this approach were a good fit to the data.

From the exploratory functional analysis it was evident that there is a large degree of overlap in the existing SEPA groups. The SEPA groups were most distinct in terms of observed patterns of alkalinity although this is unsurprising as the current groups used are primarily based on broad categories of alkalinity, and hence this determinand would likely drive any differences between them. For both phosphorus and chlorophyll there was less evidence of differences between the group means, considering only at the estimated curves for each lake. Some groups of sites appeared to be more distinct from other current groups for all determinands but in particular, in terms of alkalinity. The geographical location of some of sites was thought to be one potential explanation for the differences since the surrounding environment and land use around the lakes may be quite distinct in different places.

Functional regression was used for the estimation of group specific effects that summarised the group data. One drawback of the functional regression approach, however, was the degree of uncertainty associated with each of the group effects due to the relatively small number of functional observations. Despite the large standard errors for each of the group effects, permutation F-tests proved for all

determinands that there are at least some differences between the current groups. Further to this the results of t-tests reinforced that not all of the current groups were distinct from one another and that fewer groups may be sufficient.

The effects of correlation

The effects of correlation on our ability to detect different forms of underlying change are investigated in Chapter 2, while our ability to distinguish between different groups of lakes in the presence of correlated observations is explored in Chapter 3. When there is strong correlation present, the power of the different sampling schemes to detect underlying change considered in Chapter 2 diminishes. The effects of correlation were found to be particularly evident within the simulation scenario which considered the power to detect a change in the amplitude of a seasonal component over time. Autocorrelation appears to have more of a detrimental effect on the power to detect a varying seasonal pattern than any of the other forms of change considered here. It is therefore important to assess the presence and strength of autocorrelation in the data is the change of interest is a change in the seasonal signal over time.

A novel simulation study was presented in Chapter 3 which was designed to assess the effect of correlated errors on our ability to identify differences between groups of lakes. Although the presence of temporal correlation did not have a large effect on our ability to distinguish between different groups of lakes, there was some evidence to suggest that there may be a limited effect if the temporal autocorrelation is strong. The study presented considered only monthly observations since this was the sampling frequency of interest for the Scottish lakes data. In other contexts, if the observations were more frequently collected it is likely the correlation may have a bigger impact and should be taken into account when interpreting the results of permutation tests.

Forming new group structures

In order to ensure accurate classification of all lakes when using the groups approach for monitoring, it is of key importance that the groups are formed from sites which are similar to one another in terms of the levels and temporal dynamics of the determinands of interest.

Hierarchical functional clustering was first applied to each individual determinand and provided a good initial step that enabled us to visually assess if there was any clear underlying group structure in the Scottish lakes data. Following this, model based clustering was used. This method had the benefit of being compatible with standard model comparison techniques that can be used to select the optimal number of groups, and is an approach that can also be used to provide a level of confidence for the classification of each lake. Furthermore, the functional clustering model proposed by James and Sugar (2003) had the additional advantage of being able to deal with irregular and sparsely sampled data, which was a problem in the Scottish lakes data.

While it is often difficult to identify a group structure looking solely at the observed data for each of the sites alongside the predicted group means, particularly in the multivariate case, the FCM overcomes this difficulty through the parameterization of the group effect used. This set-up of the group effect allowed low-dimensional projections of the curves to be computed and plots of these could then be used to clearly identify clusters of sites, if they exist. The projected cluster values can also be used to select the representative sites which perform well in terms of representing the key features of all members of the group.

The extension of the FCM to the multivariate setting enables information from several determinands of interest to be used within formation of groups. This new approach is something which is of particular importance in view of the fact that the WFD classification encompasses a range of different variables. While the WFD is an extremely complex piece of legislation, in which the classification of sites encompasses a huge range of different determinands, both chemical and biological, grouping the sites using a functional clustering approach provides a statistical basis for determining a group structure which is based on a selection of the variables of most interest. Differences in univariate groups for the the Scottish lakes data were primarily based on mean concentrations while the multivariate model did indicate a split in the groups underpinned by the strength of seasonal signals at the lakes.

As the multivariate model can be used to determine a group structure that is based on several determinands it is an extremely useful tool in the design of a monitoring network. If the aim is to explore potential groupings for classification then the multivariate model is more suitable than a univariate approach. However, while the multivariate model can be used to obtain a group structure for a set of

variables, it is often difficult to interpret the estimated lake curves and cluster mean curves for each of the determinands separately in the multivariate model. Consequently, if there is an additional interest in temporal patterns for individual variables at each site this can be investigated by fitting relevant curves using penalised regression splines. Alternatively, the univariate models can be used. Another approach may be to combine the variables in some way via a trophic index and proceed with this in a univariate setting. The idea of applying the univariate data to trophic index data is explored in Pastres et al. (2011). The key benefit of the multivariate model is that there is no question as to how to combine the variables of interest. The functional clustering approach provides a statistical framework for investigation of water quality across multiple sites in an efficient and cost effective way.

Identifying the optimal number of groups

Analysis of new group structures explored in Chapter 4 suggested that a smaller number of groups would be sufficient in capturing the differences between the lakes. For the hierarchical and model based clustering methods explored, the statistically optimal number of clusters identified were fewer than the number currently used. The existing SEPA groups are in many cases combined to form larger groups within the new group structures. This was the case both when each determinand was considered independently, and when all three were considered simultaneously in the multivariate functional clustering model. The gap statistic provided an approach for selecting the optimal number of clusters which produced consistent results that were sensible given the earlier exploratory analysis of the current SEPA groups for the Scottish lakes. For the univariate models there was agreement between the number of clusters identified as being best using both methods. Furthermore, the number of clusters determined using the functional hierarchical approach also agreed with the number of clusters identified for each of the univariate determinand FCMs. In practical terms, one of the key implications of reducing the number of groups is that the number of sites which are required to be monitored can be reduced, while ensuring that variability amongst the lakes is account for, not only with regards to mean levels, but also for trends and seasonal signals. Reduction of monitoring networks to include a smaller number of sites could be balanced with an increase in the sampling frequency at each of these sites

so that potentially small or complex changes that could affect all group members can be observed if they are occurring.

6.3 Identifying Spatially Homogenous Groups

There was evidence in from the results of the functional clustering models that groups of lakes determined using the functional clustering model had a spatial pattern. However, limitations in the quantity of data available hindered any further investigation of incorporating spatial covariance in groups of lakes used for classification. Following from this the River Tweed nitrate dataset was used to investigate how spatial variability amongst locations could be included when forming groups of stations. Investigating river network data introduced a set of new challenges for functional clustering which went beyond the inclusion of Euclidean distance based spatial correlation. Using the standard measure of covariance and ideas from the trace variogram, a novel method of defining covariance using areas between pairs of curves was developed. The development of this method enabled the tail-up covariance model to be applied to functional data on a river network using stream distance.

All of the functional clustering approaches applied to the Tweed, hierarchical and the model based approach, were able to clearly identify the area in the north-east of the Tweed Network which is known to be a Nitrate Vulnerable Zone. The mean concentration of nitrate at stations in this region far higher than elsewhere and the seasonal signal was far stronger. However the methods differed in terms of the number of clusters of stations within the Nitrate Vulnerable Zone. The detrended stream distance covariance weighted clustering identified only one group of stations covering the Nitrate Vulnerable Zone, whilst the Euclidean distance approaches identified four groups consisting of either one or two stations each.

As before the gap statistic was used to select the statistically optimal number of clusters. The smallest number of clusters identified as statistically optimal was using the hierarchical clustering weighted by stream distance, whilst the largest was for hierarchical clustering which assumed the stations were spatially independent. Although there is no single correct number for how many groups is best to describe the Tweed stations, including information about the spatial variability in nitrate levels based on stream distance resulted in clusters which were all distinct

in terms of their cluster means. This suggests that the stream distance weighted hierarchical clustering may be the best approach to take if the aim is to identify groups of stations on a river network, which are similar in terms of the temporal dynamics of the determinand of interest.

6.4 Future Work

There are several possible extensions to the statistical analysis of environmental monitoring networks that has been carried out within this thesis. The nature of possible future work involves not only direct extensions of the analysis of the problems presented, but also could involve additional statistical challenges.

There are a variety of extensions which could be applied to the simulation study presented in Chapter 2. The simulation patterns chosen within the study were designed to be indicative of the types of underlying patterns which may be of interest when investigating changes in water quality. Clearly there are a number of different of forms of change which could be investigated using a similar approach to that taken within this simulation study. Alternative non-linear patterns of change could be considered in order to explore the ability of monitoring networks to detect particular changes of interest. Furthermore, it may be of interest it to investigate the ability to detect a change in the phase of a seasonal pattern over time, rather than a change in the amplitude.

For the lake groups data missing data were first imputed using interpolating splines and after this penalised regression splines were fitted. An alternative way of dealing with the problem of missing observations in estimation of functional data would be to apply the ridge parameter approach which was successfully applied to the lakes and river data in Chapters 4 and 5. Other possible extensions to the initial functional data analysis of the lakes would be to consider a multivariate functional regression as discussed in Ramsay and Silverman (1997) in order to consider all of the determinands when examining the existing SEPA groups. However, in this context, as the work presented within Chapter 3 was to explore if there were any differences between the existing groups in terms of at least one determinand, rather than all of the determinands available, this seemed to be unnecessary here.

The hierarchical clustering approach could be extended to the multivariate setting using the method described in Henderson (2006) where, after a suitable

standardization of each of the determinands, the overall squared functional distance matrix can be defined as the sum of the squared distance matrices for the individual determinands. In the context of the lakes data, the multivariate functional clustering model had several advantages over the hierarchical approach, and so the application of the hierarchical method did not go beyond the univariate cases presented. With the river Tweed data, data on only a single determinand was available and so there was no scope to develop a multivariate technique. Where there are more data available, and it is of interest to combine information across several determinands, multivariate hierarchical functional clustering methods could be used in conjunction with the spatial weights as demonstrated in Chapter 5 to explore grouping sites.

Although no covariate data were available for the lakes or the Tweed dataset explored here, in situations where such data were available it may be advantageous to incorporate this into the functional data analysis techniques considered. One method to include covariate data would be to create functions of the response variable where the argument is the covariate rather than time, and subsequently use these functions for any analysis. Alternatively, assuming functions of the response variable and any covariates of interest are both functions of time, and cover the same time period, a concurrent functional regression model can be used to relate the value of the functional response to the value of the functional covariates at the corresponding time points. In the context of water quality, response functions of a particular chemical determinand over time could be modelled using patterns of land use or climate variables over time. Groups of these modelled functions could be then be compared using permutation F-tests or t-tests, or the functions could be used within a functional cluster analysis. Covariate data can more simply be included within functional clustering approaches using a multivariate model. In addition to functional covariates, scalar covariates can be included by including them as a constant function over time.

While spatial covariance has been successfully incorporated within hierarchical clustering approaches, due to the advantages of model based functional clustering over non-probabilistic methods it would be of interest to investigate the potential for the functional clustering model to incorporate measures of spatial covariance. As noted in Chapter 5, due to the structure of the functional clustering model it is likely that the spatial covariance matrix incorporated within the functional clustering model would need to be defined in terms of covariance between individual

spline coefficients. Additional complexities are also presented if the model were to be applied ro river network data since the stream distance and flow weights would also need to be defined in terms of spline coefficients. If a suitable measure of covariance could be defined in terms of the basis coefficients, this term could subsequently be used as a constraint within the EM algorithm in a similar way to that proposed by Soares et al. (1996).

The methodology applied and developed within this thesis provides ideas for the design of effective and cost efficient monitoring networks for environmental data that are underpinned by a solid statistical basis. The consequence of this is that classifications made using data collected by these monitoring programmes will be reliable and accurate. This thesis develops and explores a new class of models for grouping spatiotemporal data and provides a platform for a variety of future research.

Bibliography

Abraham, C., Cornillion, P. A., Matzner-Lober, E. and Molinari, N. (2003), 'Unsupervised curve clustering using b-splines', *Scandinavian Journal of Statistics* **30**, 581–595.

Akaike, H. (1973), 'Information theory and an extension of maximum likelihood principle', *Second International Symposium on Information Theory, Akademia Kiado* pp. 267–281.

Akita, Y., Carter, G. and Serre, M. L. (2007), 'Spatiotemporal nonattainment assessment of surface water tetrachloroethene in new jersey', *Journal of Environmental Quality* **36**(2), 508–520.

Banfield, J. D. and Raftery, A. E. (1993), 'Model-based gaussian and non-gaussian clustering', *Biometrics* **49**(3), 803–821.

Bates, B., Kundzewicz, Z., Wu, S. and J.P., P. (2008), Climate change and water. technical report, Technical report, Intergovernmental Panel on Climate Change (IPCC) Secretariat, Geneva.

Bourgault, G. and Marcotte, D. (1991), 'The multivariate variogram and its application to the linear coregionalization model', *Mathematical Geology* **23**, 899–928.

Bourgault, G., Marcotte, D. and Legendre, P. (1992), 'The multivariate (co)variogram as a spatial weighting function in classification methods', *Mathematical Geology* **24**, 463–478.

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

Bowman, A. W. and Azzalini, A. (2010), *R package* `sm`*: nonparametric smoothing methods (version 2.2-4)*, University of Glasgow, UK and Università di Padova,

Italia.
URL: *"http://www.stats.gla.ac.uk/ adrian/sm"*

Bowman, A. W., Giannitrapani, M. and Scott, E. M. (2009), 'Spatiotemporal smoothing and sulphur dioxide trends over europe', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **58**(5), 737–752.

Brockwell, P. J. and Davis, R. A. (1991), *Time Series: Theory and Methods*, 1991 edn, Springer-Verlag, New York, NY.

Brunsdon, C., Fotheringham, A. S. and Charlton, M. E. (1998), 'Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis', *Environment and Planning A* **30**, 1905–1927.

Calinski, T. and Harabasz, J. (1974), 'A dendrite method for cluster analysis', *Communications in Statistics* **3**, 1–27.

Carvalho, L. and Kirika, A. (2003), 'Changes in shallow lake functioning: response to climate change and nutrient reduction', **506-509**, 789–796.

Chalwa, R. and Hunter, P. R. (2005), 'Classification of bathing water quality based on the parametric calculation of percentiles is unsound', *Water Research* **39**, 4552–4558.

Chiou, J. and Li, P. (2007), 'Functional clustering and identifying substructures of longitudinal data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(4), 679–699.

Chudova, D., Hart, C., Mjolsness, E. and Smyth, P. (2004), Gene expression clustering with functional mixture models., *in* S. Thrun, L. K. Saul and B. SchÃlkopf, eds, 'NIPS', MIT Press.

Clement, L., Thas, O., Vanrolleghem, P. A. and P., O. J. (2006), 'Spatio-temporal statistical models for river monitoring networks', *Water Science & Technology* **3**(1), 9–15.

Cleveland, R. B., Cleveland, W. S., Mcrae, J. E. and T., I. (1990), 'Stl: A seasonal-trend decomposition procedure based on loess', *Journal of Official Statistics* **6**(1), 3–73.

Cleveland, W. S. and Devlin, S. J. (1988), 'Locally weighted regression: An approach to regression analysis by local fitting', *Journal of the American Statistical Association* **83**(403), 596–610.

Craven, P. and Wahba, G. (1979), 'Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation.', *Numerische Mathematik* **31**, 377–403.

Cressie, N. (1985), 'Fitting variogram models by weighted least squares.', *Mathematical Geology* **17**, 563–586.

Cressie, N. (1993), *Statistics for Spatial Data*, John Wiley & Sons.

Cressie, N., Frey, J., Harch, B. and Smith, M. (2006), 'Spatial prediction on a river network', *Journal of Agricultural, Biological, and Environmental Statistics* **11**, 127–150.

Cressie, N. and O'Donnell, D. (2010), 'Statistical dependence in stream networks', *Journal of the American Statistical Association* **105**(489), 18–21.

de Boor, C. (1978), *A Practical Guide to Splines*, revised edition. edn, Springer-Verlag Berlin and Heidelberg GmbH & Co. K.

Delicado, P., Giraldo, R., Comas, C. and Mateu, J. (2010), 'Statistics for spatial functional data: some recent contributions', *Environmetrics* **21**(3-4), 224–239.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society, Series B* **39**(1), 1–38.

Di Stefano, J. (2001), 'Power analysis and sustainable forest management', *Forest Ecology and Management* **154**, 141–153.

Diggle, P. J. and Hutchinson, M. F. (1989), 'On spline smoothing with autocorrelated errors', *Austral. J. Statist* **31**(1), 166–182.

Diggle, P. and Ribeiro, P. (2007), *Model-based Geostatistics.*, Springer Series in Statistics, Springer.

Eastoe, E. F., Halsall, C. J., Heffernan, J. E. and Hung, H. (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 Artic', *Atmospheric Environment* **40**, 6528–6540.

EEA (2010), The european enviroment soer: Freshwater quality, Technical report, European Environment Agency.

Eilers, P. H. C. and Marx, B. D. (1996), 'Flexible smoothing with b-splines and penalties', *Statistical Science* **11**(2), 89–121.

Engels, J. M. and Diehr, P. (2003), 'Imputation of missing longitudinal data: a comparison of methods', *Journal of Clinical Epidemiology* **56**(10), 968 – 976.

Environment Agency (2007), The Unseen Threat to Water Quality. Diffuse water pollution in England and Wales report., Technical report, Environment Agency.

EPA, U. (1976), Quality criteria for water, Technical report, United States Environmetnal Protection Agency.

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* pp. 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.

European Parliament (2006), 'Directive 2006/7/EC. of the European Parliament, concerning the management of bathing water quality and repealing directive 76/160/EEC', *Official Journal of the European Communities* **64**, 37–51.

Fan, J. (1992), 'Design-adaptive nonparametric regression', *Journal of the American Statistical Association* **87**(420), 998–1004.

Fan, J. and Gijbels, I. (1992), 'Variable bandwidth and local linear regression smoothers', *The Annals of Statistics* **20**(4), 2008–2036.

Faraway, J. J. (1997), 'Regression Analysis for a Functional Response', *Technometrics* **39**(3), 254–261.

Ferguson, C. A., Bowman, A. W., Scott, E. M. and Carvalho, L. (2007), 'Model comparison for a complex ecological system', *Journal of the Royal Statistical Society Series A* **170**(3), 691–711.

Ferguson, C. A., Carvalho, L., Scott, E. M., Bowman, A. W. and Kirika, A. (2008), 'Assessing ecological responses to environmental change using statistical models', *Journal of Applied Ecology* **45**(1), 193–203.

Field, S. A., O'Connor, P., Tyre, A. J. and Possingham, H. P. (2007), 'Making monitoring meaningful', *Environmental Monitoring and Assessment* **32**, 485–491.

Field, S. A., Tyre, A. J. and Possingham, H. P. (2005), 'Optimizing allocation of monitoring effort under economic and observational constraints', *Journal of Wildlife Management* **69**(2), 473–482.

Fraley, C. and Raftery, A. E. (1998), 'How many clusters? which clustering method? answers via model-based cluster analysis', *The Computer Journal* **41**(8), 578–588.

Fraley, C. and Raftery, A. E. (2002), 'Model-based clustering, discriminant analysis, and density estimation', *Journal of the American Statistical Association* **97**, 611–631.

Garcia-Escudero, L. and Gordaliza, A. (2005), 'A proposal for robust curve clustering', *Journal of Classification* **22**, 185–201.

Gerrodette, T. (1987), 'A power analysis for detecting trends', *Ecology* **68**(5), 1364–1372.

Giannitrapani, M., Bowman, A., Scott, E. and Smith, R. (2005), 'Sulphur dioxide in europe: the relationship between emissions and measured concentrations.', *Atmospheric Environment* **40**, 2524–2532.

Giraldo, R., Delicado, P., Comas, C. and Mateu, J. (2010), 'Hierarchical clustering of spatially correlated functional data'.
URL: *www.ciencias.unal.edu.co/unciencias/datafile/ estadistica/RepInv12.pdf*

Giraldo, R., Delicado, P. and Mateu, J. (2011), 'Ordinary kriging for function-valued spatial data', *Environmental and Ecological Statistics* **18**, 411–426.

Goodwin, T. H., Young, M. G., Homes, H., Musgrave and Pitson, D. (2004), Development and assessment of methods to estimate flow statistics at the ungauged site for use within lf2000, Technical report, Wallingford HydroSolutions/ SEPA.

Goulard, M. and Voltz, M. (1992), 'Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix', *Mathematical Geology* **24**, 269–286.

Goulard, M. and Voltz, M. (1993), 'Geostatistical interpolation of curves: A case study in soil science', *Geostatistics Troja 92* **2**, 805–816.

Green, P. J. and Silverman, B. W. (1993), *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*, Chapman and Hall/CRC.

Gu, C. (2002), *Smoothing Spline ANOVA Models*, Springer, New York.

Guttorp, P., Meiring, W. and Sampson, P. D. (1994), 'A space-time analysis of ground-level ozone data', *Environmetrics* **5**(3), 241–254.

Hall, P. and Johnstone, I. (1992), 'Empirical functionals and efficient smoothing parameter selection', *Journal of the Royal Statistical Society. Series B (Methodological)* **54**(2), 475–530.

Hartigan, J. (1975), *Clustering Algorithms*, John Wiley and Sons, New York.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, first edn, London: Chapman and Hall.

Helsel, D. R. (1990), 'Less than obvious: statistical treatment of data below the detection limit.', *Environmental Science Technology* **24**, 1766–1774.

Helsel, D. R. (2005), *Non-detects and Data Analysis*, first edn, New York: John Wiley.

Henderson, B. (2006), 'Exploring between site difference in water quality trends: a functional data analysis approach', *Environmentrics* **17**, 65–80.

Hirsch, R. M., Slack, J. R. and Smith, R. A. (1982), 'Techniques of trend analysis for monthly water quality data', *Water Resources Research* **18**, 107–121.

Hoerl, A. E. and Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**(1), pp. 55–67.

Houseman, E. A. (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.

Howden, N. J. K., Burt, T. P., Worrall, F. and Whelan, M. J. (2011), 'Monitoring fluvial water chemistry for trend dectection: hydrological variability masks trends in datasets covering fewer than 12 years', *Journal of Environmental Monitoring* **13**, 514–521.

Hubert, L. and Arabie, P. (1985), 'Comparing partitions', *Journal of Classification* pp. pp. 193–218.

Ignaccolo, R., Ghigo, S. and Giovenali, E. (2008), 'Analysis of air quality monitoring networks by functional clustering', *Environmetrics* **19**(7), 672–686.

James, G. M. and Hastie, T. J. (2001), 'Functional linear discriminant analysis for irregularly sampled curves', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **63**(3), 533–550.

James, G, M. and Sugar, C. A. (2003), 'Clustering for Sparsely Sampled Functional Data', *Journal of the American Statistical Association* **98**(462), 397–408.

Kaufman, L. and Rousseeuw, P. (1987), Clustering by means of medoids, *in* Y. Dodge, ed., 'Statistical Data Analysis Based on the Ll Norm', North-Holland, Amsterdam, pp. 405–416.

Kaufman, L. and Rousseeuw, P. J. (1990), *Finding Groups in Data An Introduction to Cluster Analysis*, Wiley Interscience, New York.

Keizer-Vlek, H. E., Verdonschot, P. F. M., Verdonschot, R. C. M. and Goedhart, P. W. (2012), 'Quantifying spatial and temporal variability of macroinvertebrate metrics', *Ecological Indicators* **23**, 384–393.

Krzanowski, W. J. and Lai, Y. T. (1988), 'A criterion for determining the number of groups in a data set using sum-of-squares clustering', *Biometrics* **44**(1), pp. 23–34.

L., L. (2012), *NADA: Nondetects And Data Analysis for environmental data*. R package version 1.5-4.
URL: *http://CRAN.R-project.org/package=NADA*

Legg, C. J. and Nagy, L. (2006), 'Why most conservation monitoring is, but need not be, a waste of time', *Journal of Environmental Management* **78**, 194–199.

Little, R. and Rubin, D. (1987), *Statistical Analysis of Missing Data*, New York, Wiley.

Luan, Y. and Li, H. (2003), 'Clustering of time-course gene expression data using a mixed-effects model with b-splines', **19**(4), 474–482.

MacQueen, J. B. (1967), Some methods for classification and analysis of multivariate observations, *in* L. M. L. Cam and J. Neyman, eds, 'Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, University of California Press, pp. 281–297.

Maher, W. A., Cullen, P. W. and Norris, R. H. (1994), 'Framework for designing sampling programs', *Environmental Monitoring and Assessment* **30**, 139–162.

Malfait, N. and Ramsay, J. (2003), 'The historical functional linear model', *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **31**(2), 115–128.

Mann, H. B. (1945), 'Nonparametric tests against trend', *Econometrics* **13**, 245–259.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1980), *Multivariate Analysis (Probability and Mathematical Statistics)*, Academic Press.

McMullan, A., Bowman, A. W. and Scott, E. M. (2007), 'Water quality in the river clyde: a case study of additive and interaction models', *Environmetrics* **18**(5), 527–539.

McNicholas, P. D. and Murphy, T. B. (2010), 'Model-based clustering of longitudinal data', *Canadian Journal of Statistics* **38**(1), 153–168.

Milligan, G. W. and Cooper, M. C. (1985), 'An examination of procedures for determining the number of clusters in a data set', *Psychometrika* **50**(1), 159–179.

Milligan, G. W. and Cooper, M. C. (1986), 'A study of the comparability of external criteria for hierarchical cluster analysis.', *Multivariate Behavioral Research* **21**, 441–458.

Morton, R. and Henderson, B. L. (2008), 'Estimation of nonlinear trends in water quality: An improved approach using generalized additive models', *Water Resources Research* **44**.

Nagelkerke, L. A. J. and van Densen, W. L. T. (2007), 'Serial correlation and interannual variability in relation to the statistical power of monitoring schemes to detect trends in fish populations', *Environmental Monitoring and Assessment* **125**(1-3).

Nerini, D., Monestiez, P. and Mant, C. (2010), 'Cokriging for spatial functional data', *Journal of Multivariate Analysis* **101**(2), 409 – 418.

Nicholson, M. D. and Fryer, R. J. (1992), 'The statistical power of monitoring programmes', *Marine Pollution Bulletin* **24**(3).

O'Donnell, D. (2012), Spatial Prediction and Spatio-Temporal Modelling of River Network Data, PhD thesis, University of Glasgow.

Oliver, M. and Webster, R. (1989), 'A geostatistical basis for spatial weighting in multivariate classification', *Mathematical Geology* **21**, 15–35.

Opsomer, J., W. Y. and Yang, Y. (2001), 'Nonparametric Regression with Correlated Errors', *Statistical Science* **16**(2), 134–153.

Pastres, R., Pastore, A. and Tonellato, S. F. (2011), 'Looking for similar patterns among monitoring stations. venice lagoon application', *Environmetrics* **22**(6), 712–724.

Pebesma, E. (2004), 'Multivariable geostatistics in s. the gstat package.', *Computers and Geosciences* **30**, 683–691.

Plaia, A. and Bondi, A. (2006), 'Single imputation method of missing values in environmental pollution data sets', *Atmospheric Environment* **40**(38), 7316–7330.

Raftery, A. (1995), 'Bayesian model selection in social research', *Sociological Methodology* **25**, 111–163.

Ramsay, J., G., H. and Graves, S. (2009), *Functional Data Analysis with R and MATLAB (Use R)*, Springer.

Ramsay, J. O. and Dalzell, C. J. (1991), 'Some Tools for Functional Data Analysis', *Journal of the Royal Statistical Society. Series B (Methodological)* **53**(3), 539–572.

Ramsay, J. O., Wickham, H., Graves, S. and Hooker, G. (2010), *fda: Functional Data Analysis*. R package version 2.2.1.
URL: *http://CRAN.R-project.org/package=fda*

Ramsay, J. and Silverman, B. W. (1997), *Functional Data Analysis (Springer Series in Statistics)*, 1st edn, Springer.

Ramsay, J. and Silverman, B. W. (2003), *Applied functional data analysis : methods and case studies (Springer Series in Statistics)*, Springer.

Rand, W. M. (1971), 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical Association* **66**(336), pp. 846–850.

Romano, E., Balzanella, A. and Verde, R. (2010), Clustering spatio-functional data: A model based approach, *in* 'Classification as a Tool for Research', Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin Heidelberg, pp. 167–175.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581–592.

Ruppert, D. and Wand, M. P. (1994), 'Multivariate locally weighted least squares regression', *The Annals of Statistics* **22**(3), 1346–1370.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**, 461–464.

Secchi, P., Vantini, S. and Vitelli, V. (2011), Spatially clustering of functional data, *in* F. Ferraty, ed., 'Recent Advances in Functional Data Analysis and Related Topics', Contributions to Statistics, Physica-Verlag HD, pp. 283–289.

SEPA (2009), The river basin management plan for the solway tweed river basin district 2009-2015, Technical report, Scottish Environment Protection Agency/Environment Agency.

Shaddick, G. and Wakefield, J. (2002), 'Modelling daily multivariate pollutant data at multiple sites', *Applied Statistics* **51**(3), 351–372.

Shaikh, M., McNicholas, P. D. and Desmond, A. F. (2010), 'A pseudo-EM algorithm for clustering incomplete longitudinal data', *The International Journal of Biostatistics* **6**(1).

Shen, Q. and Faraway, J. (2004), 'An f test for linear models with functional responses', *Statistica Sinica* **14**, 1239 – 1257.

Shumway, R., Azari, R. and Kayhanian, M. (2002), 'Statistical approaches to estimating mean water quality concentrations with detection limits', *Environmental Science Technology* **36**, 3345–3353.

Singh, A., Singh, A. and Engelhardt, M. (2007), The Lognormal Distribution in Environmental Applications, Technical report, United States Environmental Protection Agency. Technology Support Center Issue Paper.

Smith, E. P., Rheem, S., Holtzman, G. I., Patil, G. P. and Rao, C. R. (1993), *Multivariate assessment of trend in environmental variables*, Multivariate Environmental Statistics, Elsevier Science Publishers B.V.

Smith, V. H., Tilman, G. D. and Nekola, J. (1999), 'Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems', *Environmental Pollution* **100**, 179–196.

Soares, A., Gomez-Hernandez, J. and Froidevaux, R., eds (1996), *Clustering of Spatial Data by the EM Algorithm. In geoENV - Geostatistics for Environmental Applications*. Quantitative Geology and Geostatistics, vol. 9.

Sugar, C. A. and James, G. M. (2003), 'Finding the number of clusters in a dataset: An information-theoretic approach', *Journal of the American Statistical Association* **98**(463), 75–0–763.

Sugiura, N. (1978), 'Further analysis of the data by akaike' s information criterion and the finite corrections', *Communications in Statistics - Theory and Methods* **7**(1), 13–26.

Sun, Y. and Genton, M. G. (2011*a*), 'Adjusted functional boxplots for spatio-temporal data visualization and outlier detection', *Environmetrics* .

Sun, Y. and Genton, M. G. (2011*b*), 'Functional boxplots', *Journal of Computational and Graphical Statistics* **20**, 316–334.

Tibshirani, R., Walther, G. and Hastie, T. (2001), 'Estimating the number of clusters in a data set via the gap statistic', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **63**(2), pp. 411–423.

van Belle, G. and Hughes, J. (1984), 'Non Parametric Tests for Trend in Water Quality', *Water Resources Research* **20**(1).

Ver Hoef, J. and Cressie, N. (1993), 'Multivariate spatial prediction', *Mathematical Geology* **25**(2), 219–240.

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

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

Ward, J. H. (1963), 'Hierarchical grouping to optimize an objective function', *Journal of the American Statistical Association* **58**(301), pp. 236–244.

Webster, R. and Oliver, M. (2007), *Geostatistics for environmental scientists*, John Wiley & Sons.

Winder, M. and Schindler, D. E. (2004), 'Climate change uncouples trophic interactions in an aquatic ecosystem', *Ecology* **85**(8), pp. 2100–2106.

Wood, S. (2003), 'Thin plate regression splines', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(1), 95–114.

Wood, S. and Augustin, N. (2002), 'GAMs with integrated model selection using penalized regression splines and applications to environmental modelling', *Ecological Modelling* **157**(2-3), 157–177.

Wood, S. N. (2006), *Generalised Additive Models - An Introduction with R (Texts in Statistical Science)*, 1st edn, Chapman & Hall/CRC.

Wood, S. N. (2011), 'Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(1), 3–36.

Yamanishi, Y. and Tanaka, Y. (2003), 'Geographically weighted functional regression.', *Journal of Japanese Society of Computational Statistics* **15**, 307–317.

Yan, M. and Ye, K. (2007), 'Determining the number of clusters using the weighted gap statistic', *Biometrics* **63**(4), pp. 1031–1037.

Yue, S. and Wang, C. (2004), 'The mann-kendall test modified by effective sample size to detect trend in serially correlated hydrological series', *Water Resources Management* **18**, 201–218.