Spatial Prediction and Spatio-Temporal Modelling on River Networks

David O'Donnell

A Dissertation Submitted to the
University of Glasgow
for the degree of
Doctor of Philosophy

School of Mathematics and Statistics

January 2012

Abstract

The application of existing geostatistical theory to the context of stream networks provides a number of interesting and challenging problems. The most important of these is how to adapt existing theory to allow for stream, as opposed to Euclidean, distance to be used. Valid stream distance based models for the covariance structure have been defined in the literature, and this thesis explores the use of such models using data from the River Tweed.

The data span a period of twenty-one years, beginning in 1986. During this time period, up to eighty-three stations are monitored for a variety of chemical and biological determinands. This thesis will focus on nitrogen, a key nutrient in determining water quality, especially given the Nitrates Directive (adopted in 1991) and the Water Framework Directive(adopted in 2002). These are European Union legislations that have set legally enforcable guidelines for controlling pollution which national bodies must comply with.

The focus of analysis is on several choices that must be made in order to carry out spatial prediction on a river network. The role of spatial trend, whether it be based on stream or Euclidean distance, is discussed and the impact of the bandwidth of the estimate of nonparametric trend is explored. The stream distance based "tail-up" covariance model structure of Ver Hoef and Peterson (2010) is assessed and combined with a standard Euclidean distance based structure to

form a mixture model. This is then evaluated using crossvalidation studies in order to determine the optimum mixture of the two covariance models for the data. Finally, the covariance models used for each of the elements of the mixture model are explored to determine the impact they have on the lowest root mean squared error, and the mixing proportion at which it is found.

Using the predicted values at unobserved locations on the River Tweed, the distribution of yearly averaged nitrate levels around the river network is predicted and evaluated. Changes through the 21 years of data are noted and areas exceeding the limits set by the Nitrates Directive are highlighted. The differences in fitted values caused by using stream or Euclidean distance are evident in these predictions.

The data is then modelled through space and time using additive models. A novel smoothing function for the spatial trend is defined. It is adapted from the tail-up model in order to retain its core features of flow connectivity and flow volume based weightings, in addition to being based on stream distance. This is then used to model all of the River Tweed data through space and time and identify temporal trends and seasonal patterns at different locations on the river.

Acknowledgements

I would like to acknowledge the fantastic help and support of my supervisors Marian Scott and Adrian Bowman. Throughout my PhD their encouragement and guidance have been invaluable to me. The last four years would have been much more difficult without their contribution. I would also like to acknowledge the Scottish Environment Protection Agency for providing the data for the project. In particular I would like to thank Mark Hallard, Ted Schlicke and Fiona Carse for their expertise and input, as well as everyone behind the scenes that worked to provide me with and make sense of the data. I would also like to acknowledge the Engineering and Physical Sciences Research Council (EPSRC) for providing funding for this project.

I would like to thank Professor Noel Cressie for allow me the opportunity to work with him at the Ohio State University for three months. I had a fantastic time and learned a lot on both a personal and intellectual level. A special thanks must go to Peter, Matthias and Michael for making my stay so much more enjoyable by introducing me to so many people, and involving me in so many of your social activities. I gratefully acknowledge the Jim Gatherall scholarship for providing the generous funding for my visit.

To the (former) Department of Statistics at Glasgow University, it is fair to say that I would not still be here eight years after I first arrived were it not for

the warm and friendly atmosphere that was evident from the very first day. I sincerely hope that the stats group is able to retain its independence and identity into the future. I have made so many friends in this department and, in particular I'd like to thank Claire, Robin, Ally, Nicola, James, Claire and Gillian for all the lunches and after uni drinks, and the odd bit of stats advice (but mainly the lunches and drinks)!

I would also like to thank my friends from outside uni, in particular Alan and Gary, for the welcome distractions from my work. All the quizzes, games of snooker, football matches and nights out are what have kept me sane through the years.

To my family, mum, dad, Susan, granny, grandad and many more, I couldn't have done this without you, thanks to your never-ending support and belief. You are the reason that I have been able to accomplish so much. This is for all of you.

Finally to Ruth, thanks for everything over the last couple of years. Writing up has meant that I have not always been the easiest person to deal with, and I can't thank you enough for always being there for me.

Contents

List of Tables

List of Figures

Chapter 1

Background

The routine monitoring of water quality has been carried out on rivers around the world for decades. Increasing industrialisation and intensive farming make it more relevant than ever before to monitor the effects these activities have on our water bodies and the aquatic life they support.

The Scottish Environment Protection Agency (SEPA) is responsible for the routine collection and evaluation of water quality data from Scotland's lochs, rivers and estuaries. One of its roles is to identify any areas that exceed predetermined safe levels of certain pollutants and, if necessary, take steps to bring these areas under control. European Union directives such as the Nitrates Directive (European Parliament, 1991) and the Water Framework Directive (European Parliament, 2000) have made this even more of a necessity, as EU members have committed to meet the targets outlined in such legislation. This requires regular reporting of the water quality and the state of the surrounding environment in general, for examples see Robson et al. (1996), Tweed River Purification Board (1996) and Scottish Environment Protection Agency (2009b).

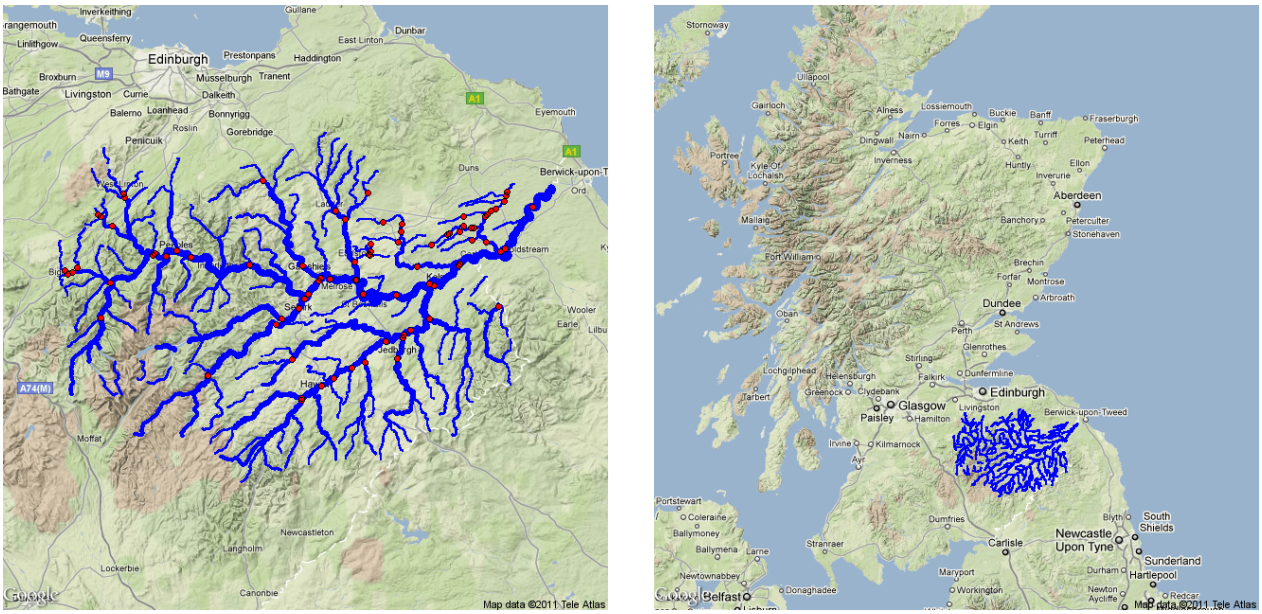
Rivers in particular have been subject to increasing levels of pollution over

the last century from both diffuse pollutants, seeping from nearby land and fields, and point sources such as sewage treatment facilities. The monitoring of river networks, no matter where they are, is generally carried out using reasonably few monitoring stations in order to minimise the time and money required to maintain such a network. This means that the monitored locations must be used to provide inference on the remaining locations on the river, and methods must be used so that such predictions are accurate.

1.1 Data

The data that will be used are supplied by SEPA and cover the River Tweed, which runs through the borders in the south of Scotland, and some of its tributaries. The Tweed is surrounded by mainly arable land and passes through relatively few built-up areas, when compared to other major rivers such as the Forth and the Clyde. In 1976, the Tweed was designated a 'Site of Special Scientific Interest', and has since been a focus of the Land Ocean Interaction Study research program, with studies at certain river sites carried out between 1994 and 1997 (Robson and Neal, 1997; Clayton, 1997). As of 2008, SEPA classifies 51% of river water bodies in the Tweed catchment as "good or better", but aim to increase this to 98% by 2027 (Scottish Environment Protection Agency, 2009a).

Data are available from January 1986 to October 2006 for up to one hundred and thirty nine unique monitoring sites on the river, although not all stations have measurements over the same time period. Figure 1.1 shows the River Tweed and its tributaries superimposed over maps of Scotland, firstly with just the local area and then on the map of Scotland (Loecher, 2011). Monitoring data are available at some point in time for each of the points marked in red. A plethora of different chemical and biological determinands were measured, although some



(a) (b)

Figure 1.1. Maps of the River Tweed. Line thickness is proportional to flow volume so larger streams are more prominent.

were not monitored for the entire 21 year period. A variety of land use data was also made available, but the variables, such as number of cattle in the surrounding area, are just based on one particular time point and so do not change over the monitoring period.

Figure 1.2 shows a picture of the River Tweed and its tributaries. Monitoring data are available at some point in time for each of the points marked in red. The river width in the plot is not necessarily to scale (as no data are available for this) but it does correspond to the estimated average flow, so the thicker stretches of river on the plot are likely to be the wider stretches on the actual river. The main river is therefore the slightly thicker line that runs from the far South-West to the North-East. The picture shown here is incomplete, as water bodies such as minor lochs and some river systems are not shown, though all monitoring stations are at locations which do feed into the Tweed itself eventually. This means there are some locations on Figure 1.2 where the stream segments do not seem to join

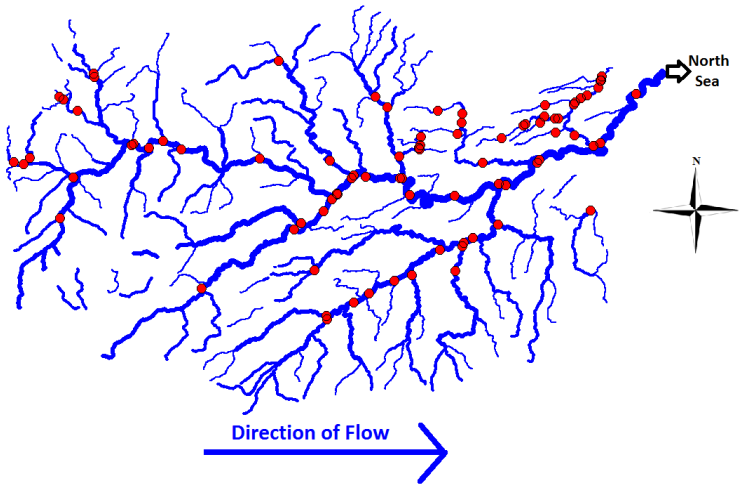


Figure 1.2. Available monitoring stations on the Tweed

up. However, some of these gaps are filled in by the map backgrounds in Figure 1.1(a). These stretches are assumed to have the same effect on the water as if it were simply traveling down a stream segment, but it is possible that the water may be in the loch for some time.

Flow data are available at twenty-one locations on the river, with daily average flows for the entire time period (and beyond). However, this was not sufficient to provide estimated flow volumes at all the 298 possible stream segments on the Tweed network. In order to provide this, the available flow data were used with a computer package called Low Flows 2000 (Goodwin et al., 2004) to produce estimated average flows at one location on all of these segments. This means that the available flow data, as represented by the widths of stream segments in Figure 1.2 are static and do not change over time. This means that irregular events a heavy rainstorm, which is likely to cause a spike in pollutant levels, cannot be reflected in the flow data.

1.1.1 Phosphorous

There are several chemical determinands available for use in the analysis, but the two that were identified as the most interesting, in terms of the way they reflect the changing health of the river network, were phosphorous and nitrate. The available phosphorous data consists of two types of measurement, Soluble Reactive Phosphorous (SRP) and Total Phosphorous (TP), both measured in milligrams per litre. TP is the sum of SRP and several other measures of phosphorous. Phosphorous levels tend to be very heavily influenced by point sources of pollution, such as those found in industrial areas, as well as runoff from fertilisers.

From looking at the phosphorous data, it is clear that there are several limit of detection issues that would have to be addressed prior to analysis. Figure 1.3 shows log transformed TP and SRP plotted over time for the Norham Gauging Station, and provides a typical example of the wider problem. It is obvious that between 1996 and 2002 the SRP levels are consistently falling below detection limits, but closer inspection shows this is also the case for some readings before 1996. From the figure it seems that, prior to 1996, the logged SRP levels seem to take only certain values in the range especially when compared to the post-2002 period, where the levels do seem to take any value in the range. There are two likely explanations for this occurrence. The first is that earlier in the time period, measuring equipment was not sensitive enough to detect minor changes in SRP. This lack of precision would then be more obvious when these values were logged, especially for observations at the lower end of the scale. The other possibility is that these values are all falling below detection limits, but the detection limits themselves are changing from reading to reading. This is quite a common occurrence in environmental data.

It is very likely that a combination of these two explanations is at work. The limit of detection does seem to be lower before 1996, but comes with the by-product of fewer decimal places in the observations, leading to lower precision and less of a difference between them. Between 1996 and 2002, the observations seem to have the same problem with precision but with a higher limit of detection than before, possibly due to new equipment. After 2002, this problem seems to disappear, probably because the equipment was upgraded to more precise and sensitive detection systems. Discussions with SEPA regarding this topic did not provide any definite conclusions about the likely cause. They confirmed that it was likely that the detection limits would change frequently as different equipment could be used for measurement. Despite this, they stated that there was no specific record of what equipment was used at certain points in time.

While phosphorous is a very interesting variable to look at in terms of how it reflects the health of the river network as a whole, the issues with limits of detection require a lot of attention before analysis could begin. Limit of detection issues are a nuisance rather than the focus of the analysis, and so phosphorous data was deemed to be not as suitable for further study as nitrate.

1.1.2 Nitrate

Available nitrate data also consists of two types of measurement: Nitrate (N) and Total Oxidised Nitrate (TON), both measured in milligrams per litre. TON is the sum of Nitrate and Nitrite levels but, since the latter tends to be very small, SEPA regards N and TON as being equivalent. This will be assumed in all analysis conducted in this thesis, so that the gaps that would otherwise be seen in the data do not provide further complication. Sewage effluent and runoff from fertilisers used on agricultural land are among the largest contributors to nitrate

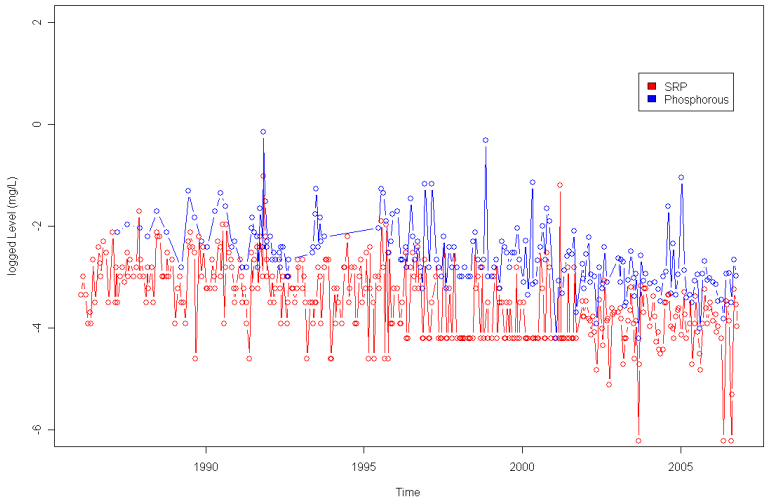


Figure 1.3. Plot of logged phosphorous levels at Norham Gauging Station

levels, so that nitrate is considered generally as a diffuse pollutant rather than due to point sources. This makes it appropriate for analysis on a mainly arable catchment such as the Tweed.

The Nitrates Directive was designed to reduce current levels of nitrate in surface and groundwater. Management options include implementing Action Program Measures in either the whole area or specially selected "Nitrate Vulnerable Zones". The Action Program Methods involve the education of farmers in the "best practice in the use and storage of fertiliser and manure". In terms of this analysis, the Nitrates Directive states that 50 mg/l of NO_3-N, or 40 mg/l with evidence of an upward trend, are the limits above which action should be taken. This translates to 11.3 mg/l and 9.04 mg/l in the available data using atomic weights.

There are several reasons why low nitrate levels are important in maintaining a healthy water body. High nitrate levels are known to lead to stress and stress-related diseases in fish and marine life, and in extreme cases to death (Romano

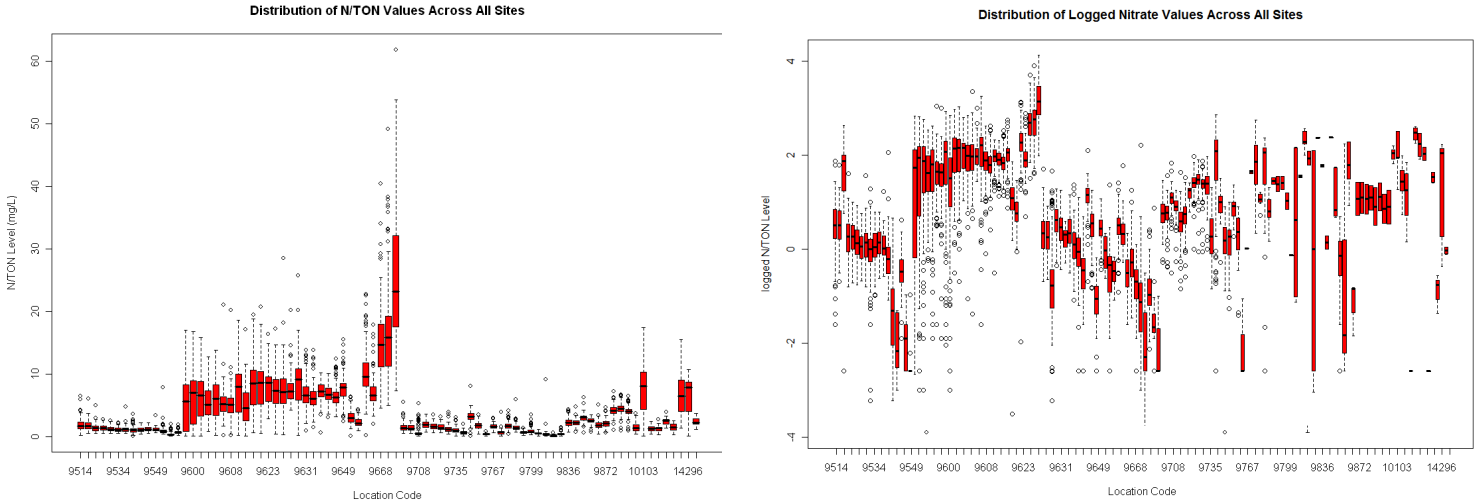
and Zeng, 2007). In humans, high levels in drinking water can also lead to long term health problems and death, especially in very young children (Addiscott and Benjamin, 2004). For these reasons, it is important to control discharges into our water bodies.

Initial plots of the distributions of nitrate values across all available sites indicated that a log transformation may be necessary to stabilise the variance. Figure 1.4(a) shows boxplots of the nitrate levels at each of the monitoring locations, across the whole time period. It is clear that the variance changes markedly across the stations, with some of the boxes covering a range of one or two milligrams per litre and others covering a range of more than 50 mg/l. Figure 1.4(b) shows the same data, but this time using the log transformed nitrate levels. This improves substantially on the previous plot, as the variance seems to be much more consistent across stations. Therefore, all further analysis will use the log transformed nitrate level. It is also worth noting that a number of stations have a very small number of observations available, and these were removed before formal analysis.

Given that nitrate is such an important indicator of water quality and that not a single nitrate observation seems to fall below a detection limit, it is the obvious choice for initial analysis of the River Tweed area.

1.2 Nitrate Data Exploration

The nitrate data will now be examined for evidence of temporal trends and seasonal patterns in order to identify the features of the data that may have to be accounted for in later analysis.



(a) Nitrate Distributions, No Transformation

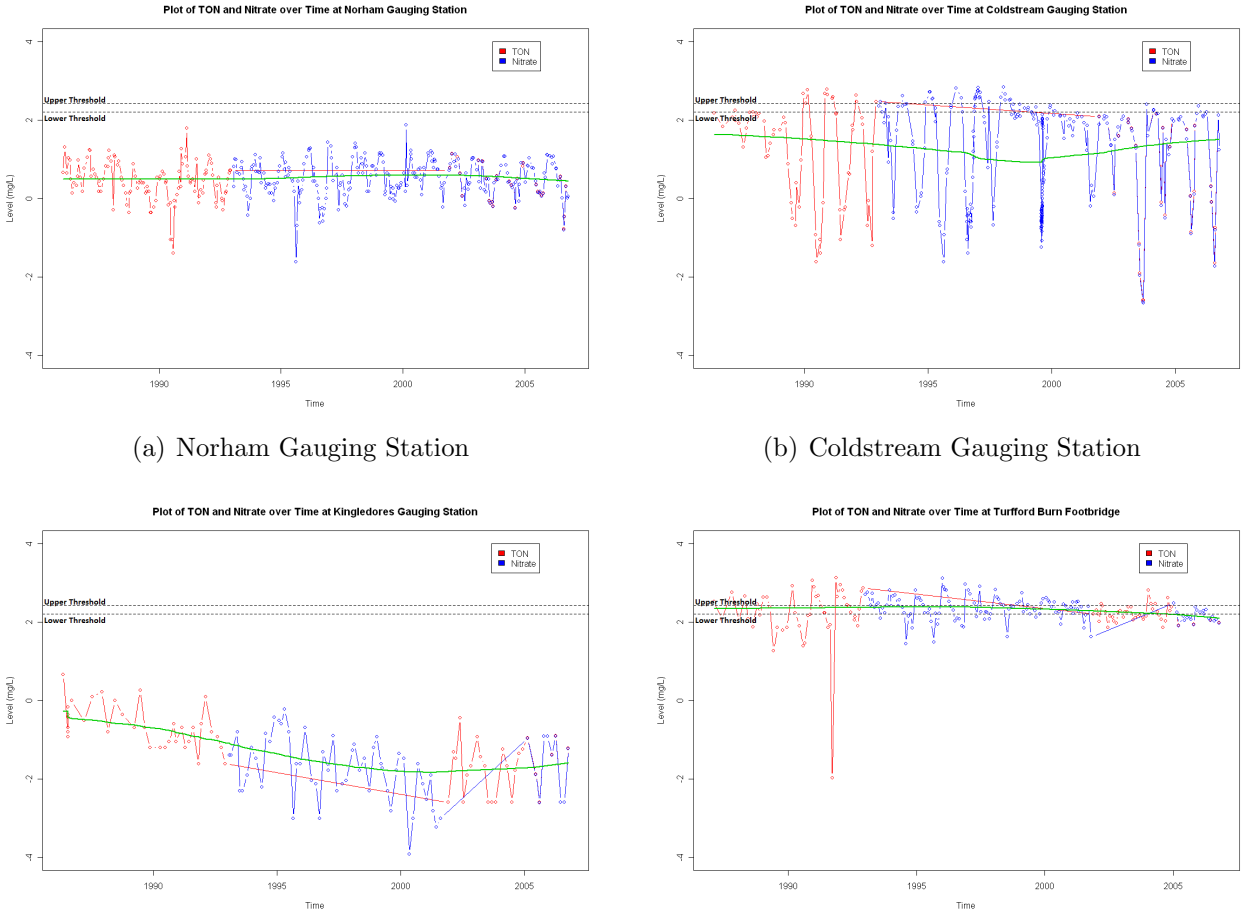
(b) Nitrate Distributions using Log Transformation

Figure 1.4. Plots showing distributions of TON/N at each Station, with and without using log transformation. Plots use same x axis and so can be directly compared.

1.2.1 Trends

To identify any possible trends in the data, the Nitrate (N) and Total Oxidised Nitrate (TON) were plotted at each station. These plots were the same as the one shown for the phosphorous data in Section 1.1.1. Only eighty three of the one hundred and thirty nine monitoring stations had sufficient observations to give a reasonable assessment of any underlying trend. A "sufficient" number of observations was defined as six or more observations in at least two years of the twenty-one year period. In all further analysis, this subset of eighty three stations is used instead of the full dataset. A Loess smoothed curve has been added to the data as well as cut-off bands to denote the two limits of 11.3 and 9.04 mg/l (transposed onto the log scale) as defined in the Nitrates Directive (Section 1.1.2).

A loess smooth identifies the k nearest neighbours to the point at which the smoothed estimate is to be obtained, x^*. The distance from x^* to the furthest away of its k nearest neighbours, $\Delta(x^*)$, is computed and weights w_i are assigned



(a) Norham Gauging Station

(b) Coldstream Gauging Station

(c) Kingledores Gauging Station

(d) Turfford Burn Footbridge

Figure 1.5. Time series plots of N/TON at selected locations with loess smooth added. Dotted lines denote the upper and lower thresholds defined by the Nitrates Directive

to each data point x_i in the neighbourhood of x^* using the tri-cube weight function, as defined in (1.1) and (1.2). The smoothed value $s(x^*)$ at point x^* is the fitted value from the least squares fit of y to x, with weights w_i, confined to the neighbourhood $N(x^*)$.

$$w_i = W\left(\frac{|\, x^* - x_i \,|}{\Delta(x^*)}\right) \tag{1.1}$$

$$W(u) = \begin{cases} (1-u^3)^3 & 0 \leq u < 1 \\ 0 & \text{otherwise} \end{cases} \tag{1.2}$$

Figure 1.5(a) shows the Nitrate and Total Oxidised Nitrate at the Norham Gauging Station. This station is one of the most interesting, as it is the closest to the mouth of the River Tweed and so provides the best estimate of the quality of water being discharged into the North Sea. Apart from a seasonal component, there is very little evidence of a significant change over time. There is a slight decline after 2002 but it is not very large at this scale. Norham is quite representative of most of the other stations on the network in that it shows little evidence of a trend, as only the Leet tributary and Kingledores Gauging Station show a trend that is much more pronounced than the trend here. It is also worth noting that the data for Norham, like many of the other monitoring stations, does not cross either Nitrates Directive threshold at any point. The lack of evidence for a trend at Norham, and at some other stations, is affected by the use of the log scale, which reduces the size of the high valued trends in the data, and accentuates the trend at Kingledores (Figure 1.5(c)). The log transformation has somewhat disguised the trend that is present at this, and several other stations, as an increase is "flattened out" by the transformation at this end of the scale. This does not mean that there is no trend present in the data, rather it means that it is not so obvious visually until the reverse transformation is taken.

Norham is also useful to illustrate the difference between TON and N. Until early 1993 only the Total Oxidised Nitrate was recorded and between 1993 and late 2001, only Nitrate was recorded. After that, one or both of these were recorded at each visit. The latter observations at Norham illustrate that there is very little difference between TON and N as there are many time points where both have been measured, and these observations are almost exactly the same,

with a mean difference of just 0.01 mg/l. This justifies the decision to treat them as equivalent, so that TON will be used to fill in the gaps in nitrate data.

Figure 1.5(b) shows N and TON at Coldstream Gauging Station, which lies very close to the point where the Leet Water meets the River Tweed. This station is quite representative of stations on and around the Leet. These stations exceed both thresholds, though mostly just the lower, on a number of occasions but seem to show the most pronounced seasonality as Summer/Autunm values are very low, while Winter/Spring values are very high in this area. A reason for this could be that this region is used much more for farming than other regions around the Tweed, meaning that around the time of harvest fewer fertilisers will be used on the crops. At these times it is likely that there will also be less rain to wash such chemicals into the surrounding river systems.

An interesting feature of the data from Coldstream can be found in August and September 1999, where readings were taken on several occasions each day, and on an almost daily basis. Sampling as dense as this was not carried out at any other station on the network at any time during the twenty one year period and no reason has been found for why observations are so densely sampled at Coldstream in this brief time period. It is likely that this could be due to some event (that SEPA were aware of) taking place in the surrounding area and the more extensive monitoring was to make sure this event did not have a detrimental effect on the water quality. Another, slightly less likely, explanation is that it was not carried out to monitor a known 'event', but to check that there were not large changes in nitrate levels outside of normal monitoring hours, namely in evenings, Saturdays and Sundays. However, no definitive explanation for this unusual monitoring has been found.

The dense sampling at Coldstream in 1999 seems to be heavily influencing the Loess smooth on the plot. The smoothed line appears to decrease between 1986

and 1999 after which it starts to rise again, but this appears to be a result of the dense monitoring in 1999 has had on the line itself. Looking at the larger values in the data, as opposed to the smoothed line, the opposite seems to be the case. There is actually a slight rise between 1986 and the mid to late 1990s, before a slight decrease. This seems to be consistent with the trends from other stations on the Leet. This is backed up by the fact that after about 1998 no values exceed the upper limit despite this being a regular occurrence between 1990 and 1998. This may be evidence that measures which have been taken after the nitrate levels repeatedly crossed the upper threshold were starting to have a beneficial effect. On the other hand, it may simply indicate that changes were happening in the winter months, where higher nitrate levels are traditionally observed, as it is unclear whether the summer observations are undergoing a similar decreasing trend.

Figure 1.5(c) shows N and TON at Kingledores, the furthest upstream of all stations lying on the River Tweed (as opposed to a tributary). The Kingledores nitrate levels show the most pronounced trend of all stations sampled, with a steady decline until around 2000, at which point it seems to level off. This again seems to be an artefact of the log scale the data are now on, as on the original scale the trend was less dramatic. While the log scale reduces the visible trend at many stations, the very low values at Kingledores cause the trend seen here to be very much more pronounced.

Figure 1.5(d) shows N and TON at the footbridge at Turfford Burn, the only area in the Tweed that is consistently above the thresholds given in the Nitrates Directive. The Turfford Burn area is a very minor tributary but it is just as important to control nitrate levels here as it is on larger rivers. All five Turfford monitoring sites give significant cause for concern, breaking both limits on a regular basis. Even the loess smoothed fit, which gives an impression of long

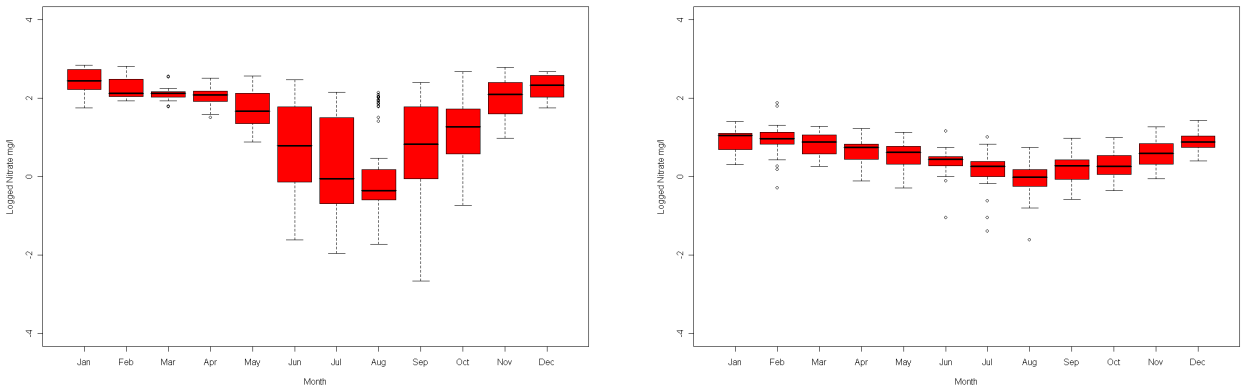
term trend without seasonality, is consistently above the lower (in four out of five) and the higher (three out of five sites) of the thresholds and so the higher observations in the winter are even worse than this. Only two other sites on the entire network have their loess smooth estimate breaking these limits at *any* point in time, and even then it is only very briefly, so Turfford is by far the most concerning location on the Tweed. However, this seems to be changing for the better. Figure 1.5(d) seems to indicate a slight reduction in nitrate levels since the late 1990s, and this pattern is replicated across each of the other monitoring stations around Turfford Burn. These sites would definitely have been identified as "Nitrate Vulnerable Zones" as soon as the Nitrates Directive was adopted in 1991, and so measures would have been put in place to reduce nitrate levels.

1.2.2 Seasonality

As well as identifying trends, it is important to determine at this stage whether the nitrate levels change throughout the year. This will be examined on a large scale by looking at the nitrate levels for each month, before moving to a smaller scale to see if day of the week or even hour of the day seem to have an effect.

Figure 1.6 shows the distribution of nitrate levels for each month of the year over the entire time period, using the Norham and Coldstream Gauging stations as examples.

From Figure 1.5, it was expected that certain stations (mainly those in the Leet) would show much greater seasonal pattern, as they were much more variable than others (such as Norham). Figure 1.6 shows that this is the case when comparing the Coldstream and Norham stations (Coldstream being a good example of the Leet data).



(a) Coldstream Gauging Station (b) Norham Gauging Station

Figure 1.6. Monthly distribution of nitrate levels at two monitoring stations

Figure 1.6(a) shows that there is a fairly large change in both the median and variance of the data over the course of the year at Coldstream. Nitrate levels seem to peak around December/January and fall to their lowest levels towards the end of summer in July/August. The variance also appears to increase in the summer months too, possibly as a by-product of unpredictable Scottish summers (and predictably wet winters), as rain is likely to increase nitrate levels by washing fertilisers from farms into the surrounding rivers. This shape of plot seems fairly consistent with others from areas of high nitrate levels, such as the Leet and Turfford Burn.

Figure 1.6(b) shows that the seasonal pattern that was evident at Coldstream is also evident at Norham, albeit to a lesser extent. The lowest levels again appear to be around August, while the highest levels appear to be around January/February. Norham is also fairly representative of most of the rest of the stations, in that there is not a huge difference between the months, but there is still a slight peak in winter and a trough in summer. The variance appears much more consistent across the months. This perhaps reflects the fact that this station is the furthest downstream and every other station feeds into it, and very few of these stations show much variability over time.

Although it is unlikely that day of the week should have a very significant effect, it is possible that weekend days may have slightly lower levels of nitrate since industrial sources of pollution may be closed (though this is not as likely to affect nitrate, which is mainly diffuse pollution). The problem with trying to assess any potential difference between different days is that there are very limited data available for Fridays, Saturdays and Sundays, and the data for these days is limited to just one or two stations. Even when it comes to other days of the week, some stations have limited available data making it very difficult to assess whether the effect differs across sites. Looking at the stations that have sufficient data, there does not appear to be much difference between days. However, a pattern may become more evident if the entire set of data were used as in Figure 1.7, instead of just a single station each time.

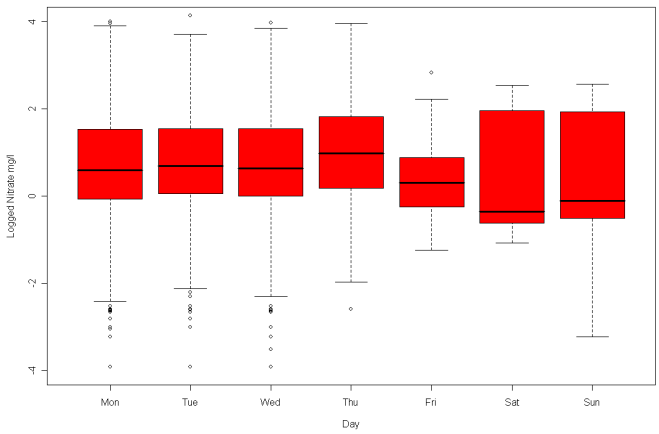


Figure 1.7. Nitrate distribution by day of the week, aggregated over all monitoring sites

From Figure 1.7, there do not appear to be huge differences between the days of the week. The boxes corresponding to Monday to Thursday are almost identical and although the boxes for Friday through Sunday look slightly different, they are made up from far fewer observations than other days. Therefore, while the weekend seems to differ slightly from the week, it is likely that this is attributable

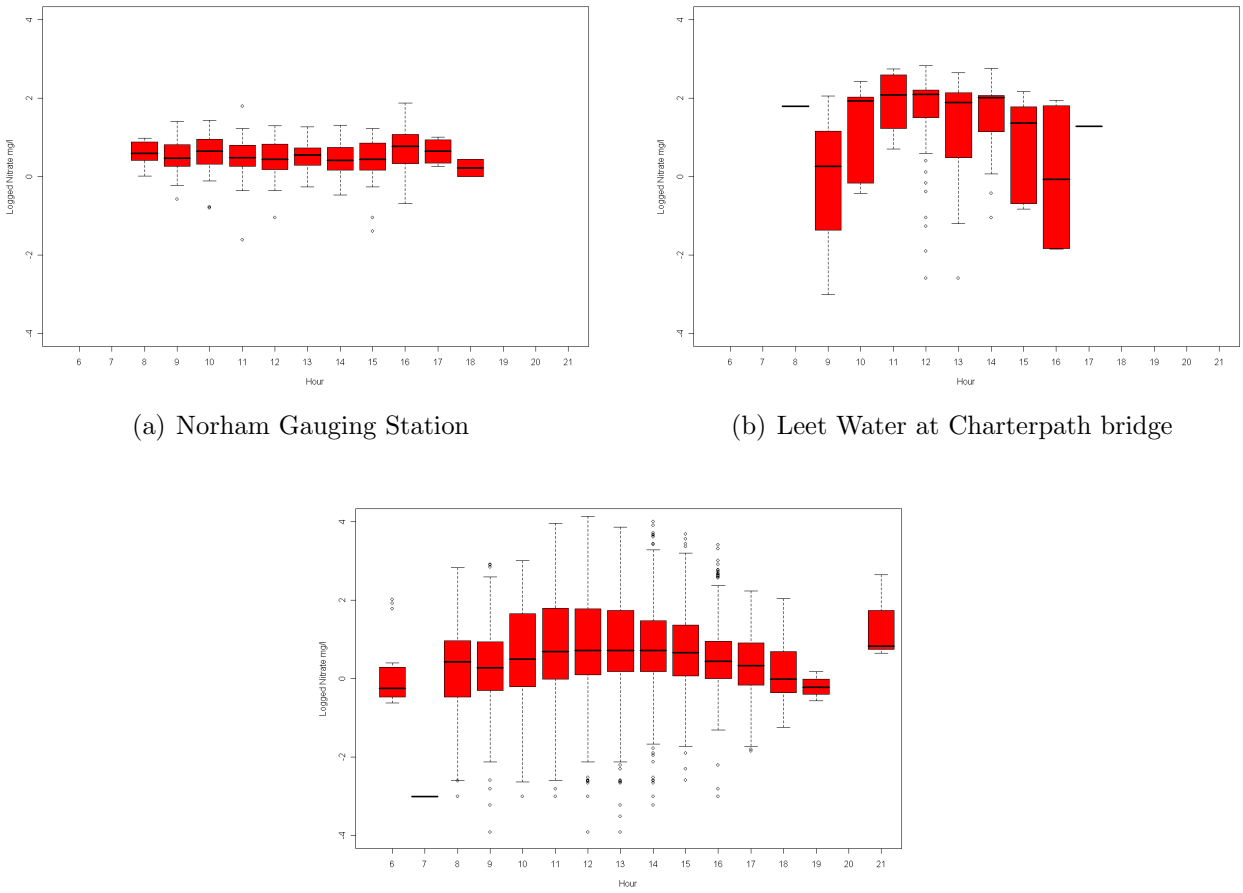
to lack of data for these days. Further analysis will therefore assume that there is no impact of day of the week on nitrate levels.

It is perhaps more natural to think of a daily cycle of a pollutant than a weekly one. Just as there was a lack of data at weekends, there is a similar problem with no observations having been taken later than 9pm or earlier than 8am. There is also a potential bias due to the fact that stations tend to be measured at roughly the same times on every day that measurements are taken, along semi-regular 'routes'. Therefore there are very few stations with anything close to a full complement of times available, and an overall plot may well be biased by the fact that a limited subset of stations provide all the information for certain time points. Figure 1.8 shows a plot of this type, as well as plots for Norham Gauging Station and the Charterpath Bridge station on the Leet.

The Norham gauging station data shown in Figure 1.8(a) provide one of the most complete sets of time points out of all stations on the Tweed. It is also one of the majority of stations that seem to exhibit little change in nitrate level throughout the day.

The data from the Charterpath Bridge station, shown in Figure 1.8(b) and located on the Leet, and is one of the few examples of reasonably well defined change in nitrate throughout the day; rising to a peak just after midday. In general, the Leet does not exhibit much hourly change, as only one other Leet monitoring station shows much evidence of a change. It also does not appear that areas of high variability in nitrate levels exhibit more of a change within a day than others, in contrast to what one might have conjectured. It is also interesting to note that for those stations which do show an hourly change, the time of the peak seems to consistently be around 1-2pm.

The overall plot in Figure 1.8(c) suggests that there may well be a sinusoidal



(a) Norham Gauging Station

(b) Leet Water at Charterpath bridge

(c) Overall distribution for all Stations

Figure 1.8. Daily distributions of nitrate levels

pattern running through the day, and suggests that this pattern is more clear-cut than could be seen from the plots from individual locations. This Figure also suggests that a peak does indeed occur sometime between midday and 2pm. The only hour exhibiting a slightly anomalous distribution would seem to be the one at 9pm. However, only two stations have had measurements taken this late (Coldstream Gauging Station and a station on a tributary of Turfford Burn) and there are very few observations at each, meaning that it is probably best to disregard these when looking for a daily cycle.

The problems with potential bias discussed previously may have something to do with the observed pattern. It is possible that areas with high nitrate, such

as the Leet and Turfford Burn, are observed more frequently around the middle of the day while areas of low nitrate are observed more at the beginning and end of the day. Looking at the times of observations, there is evidence to suggest that the observations are made along rough 'routes'- presumably to keep travel time to a minimum. Therefore, although it does seem that time of day may affect nitrate levels, the lack of data at a complete set of time points for each station would make its use very problematic.

Subsequent analysis will attempt to model the annual cycle, and day of the week and time of day will not be investigated any further.

1.3 Spatial Prediction Over A Network

Unlike most other spatial problems, a river network setting contributes an additional challenge to spatial modelling: more than one possible measure of distance is available. Euclidean distance is probably the most commonly used distance measure for spatial problems and has been used in the analysis of problems on a river network too, for example see Cressie et al. (2006). However the "stream distance", defined to be "the shortest distance between two locations where distance is only computed along the stream network" (Ver Hoef et al., 2006), could also be used in such a context; either on its own or possibly in conjunction with Euclidean Distance. Stream distance obviously has the drawback that it may not be applicable to two stations that are not "flow connected" in the river network. Two stations are described as being flow connected if the water at either location flows into the water at the other location. While finding such distances does not pose too great a problem given the GIS software currently available, Ver Hoef et al. (2006) suggest that using them is not as simple as plugging them in to existing models that were developed for Euclidean Distance.

Consider the stream distance $d(x, y)$ between two points, x and y, on a river. The stream distance has the following properties. It will never be negative (i.e. $d(x, y) \geqslant 0$)); $d(x, y) = 0$ if and only if $x = y$; it is symmetrical, so that $d(x, y) = d(y, x)$; and finally, for some other point on the river z, $d(x, z) \leqslant d(x, y) + d(y, z)$.

Euclidean distance is the simplest measure of distance available and has the advantage that there are many existing methods of modelling spatial correlation structures using Euclidean distance. Despite its simplicity, Euclidean distance may not be the most appropriate measure of distance for a river network as it takes no account of the river network structure and may not adequately reflect the distances between locations. It is, however, a very good place to start the initial investigation of spatial correlation.

1.3.1 Spatial Dependence: Euclidean Distance

Spatial correlation based on Euclidean Distance will be assessed using covariograms. A covariogram is constructed by finding the covariance between each possible set of observations at a particular point in time, and plotting these against the 'lag' – the distance between each pair of locations – to produce a covariogram cloud. The plot is then binned by averaging at regular intervals to produce the final covariogram, and it is from this plot that the underlying covariance structure can be estimated. There are several valid covariance models that can be used here and the formulae for three such models, the exponential, linear with sill and spherical are given in (1.3), (1.5) and (1.6) respectively. The exponential model is a special case of a wider set of "Matern" covariance models, defined in (1.4). Here, $\nu > 0$ is shape parameter, $\Gamma(.)$ is the gamma function,

$K\nu(.)$ is the modified Bessel function of the second kind and order ν. The exponential model is the case where $\nu = 0.5$.

$$C_{exp}(h|\theta) = \begin{cases} \theta_0 + \theta_1 & \text{if } h = 0 \\ \theta_1 \exp(-\frac{h}{\theta_2}) & \text{otherwise} \end{cases} \tag{1.3}$$

$$C_{mat}(h|\theta) = \theta_1(\frac{2^{1-\nu}}{\Gamma(\nu)}(\frac{h}{\theta_2})^\nu K_\nu(\frac{h}{\theta_2})) \tag{1.4}$$

$$C_{linsil}(h|\theta) = \begin{cases} \theta_0 + \theta_1 & \text{if } h = 0 \\ \theta_1(1 - \frac{h}{\theta_2}) & \text{if } h \leq \theta_2 \\ 0 & \text{if } h > \theta_2 \end{cases} \tag{1.5}$$

$$C_{sph}(h|\theta) = \begin{cases} \theta_0 + \theta_1 & \text{if } h = 0 \\ \theta_2\{1 - (\frac{3h}{2\theta_2} - \frac{1}{2}(\frac{h}{\theta_2})^3)\} & \text{if } h \leq \theta_2 \\ 0 & \text{if } h > \theta_2 \end{cases} \tag{1.6}$$

Here, h is the distance (in this case we are assuming Euclidean) between the points s_i and t_j on stream segments i and j and θ_0, θ_1 and θ_2 respectively are the 'nugget', 'partial sill' and 'range' parameters. The nugget effect takes account of microscale variation causing "a discontinuity at the origin" (Webster and Oliver, 2001), but is more commonly thought of as allowing for measurement error. The nugget,θ_0, combined with the partial sill, θ_1, make up the 'sill', the maximum covariance given by the model, while the range parameter θ_2 gives the distance after which there is little (for the exponential model) or no (for the other two models) change in covariance.

While covariances will be the main focus when it comes to describing the spatial dependence of the data, it is still worth noting the relationship between

the covariogram and the variogram, the most common means of describing spatial dependence. A simple way of picturing the relationship is that "a graph of the variogram is simply a mirror image of the covariance function about a line or plane parallel to the abscissa" (Webster and Oliver, 2001). The relationship is defined according to (1.7), where $\gamma(h)$ is the value of the variogram for two points a distance h apart. Combining (1.7) with the three covariance models shown previously will give the formulation of the variogram under each of the models. As an example, the exponential variogram model is shown in (1.8) and can be obtained from (1.3) using the relationship in (1.7).

$$\gamma(h) = C(0) - C(h) \tag{1.7}$$

$$\gamma(h) = \theta_0 + \theta_1(1 - \exp(-\frac{h}{\theta_2})) \tag{1.8}$$

The variogram is far more commonly used than the covariogram in spatial statistics. However, due to the fact that it is necessary to model the covariogram (as will be discussed in Chapter 2) in later analysis, the Euclidean distance based analysis will also be conducted by modeling the covariogram, to ensure that there is consistency in the theory and results.

1.3.2 Spatial Dependence: Stream distance

The potential unsuitability of Euclidean distance for describing the spatial dependence structure in the context of a river network has already been mentioned. It seems reasonable to believe that the 'stream distance' – the "shortest distance between two locations, where distance is only computed along the stream network" – has the potential to be a more appropriate distance metric. Stream

distance is an additive measure of distance, in that if the shortest stream distance between A and B is d_{ab} and the shortest stream distance between B and C is d_{bc} then the shortest distance between A and C is $d_{ac} = d_{ab} + d_{bc}$.

However, Ver Hoef et al. (2006) demonstrate that standard autocovariance models such as those detailed in Section 1.3.1 (with the exception of the exponential model) are not valid when using the stream in place of Euclidean distance. They have shown that the spherical and linear with sill models, when used with stream distance rather than Euclidean, produced covariance matrices that were not positive definite. To demonstrate this, they use an idealised stream network and simulate the covariance matrices obtained by using standard covariance models with stream distance instead of Euclidean. The resulting matrices are shown to have negative eigenvalues under certain range parameters for the linear with sill and spherical models. In order to avoid such issues, Ver Hoef and Peterson (2010) define two types of model, termed 'tail-up' and 'tail-down', to model the spatial dependence across river networks using stream distance. The names 'tail-up' and 'tail-down' refer to the direction in which the tail of the moving average process moves (either upstream or downstream).

Tail-Up Models

Tail-up models were first described in Ver Hoef et al. (2006) and Cressie et al. (2006), and take the general form shown in (1.9). The notation used here is the same as that used by Ver Hoef et al. (2006), so that $Z(s_i)$ is the value of a random variable at point s_i which is located on stream segment i. The upper and lower ends of stream segment i are denoted by u_i and l_i respectively, while the set of all stream segments upstream of i (but excluding i itself) is denoted U_i. The moving average models are then defined as the integration of the moving average process, $g(.)$, over a white noise random process, $W(.)$. Barry and Ver Hoef (1996)

state that the choice of moving average process, $g(.)$, is free but must have finite volume in order to create a stationary process. For example, the moving average processes for common structures such as the exponential and spherical are e^{-x} (for $I(0 \leq x)$) and $1 - x$ (for $I(0 \leq x \leq 1)$) respectively (Ver Hoef et al., 2006). Finally, $B_{s_{i,[j]}}$ is the set of all stream segments between segments i and j (including the furthest upstream segment but excluding the furthest downstream).

$$\begin{aligned} Z(s_i) &= \int_{s_i}^{u_i} g(x_i - s_i|\theta)W(x_i)dx_i \\ &\quad + \sum_{j \in U_{s_i}} (\prod_{k \in B_{s_{i,[j]}}} \sqrt{\omega_k}) \int_{l_i}^{u_i} g(x_j - s_i|\theta)W(x_j)dx_j \end{aligned} \tag{1.9}$$

This equation states that the value of the random variable at point s on stream segment i is composed of two different elements. We will define those two elements as:

$$\begin{aligned} a(s_i) &= \int_{s_i}^{u_i} g(x_i - s_i|\theta)W(x_i)dx_i \\ b(s_i) &= \sum_{j \in U_{s_i}} (\prod_{k \in B_{s_{i,[j]}}} \sqrt{\omega_k}) \int_{l_i}^{u_i} g(x_j - s_i|\theta)W(x_j)dx_j \end{aligned}$$

The element $a(s_i)$ integrates the moving average function $g(x_i - s_i|\theta)$ between s_i and u_i (the top of stream segment i) over the white noise process $W(x_i)$. The moving average function, a function of the distance $x_i - s_i$, can be chosen from one of the standard models to ensure that it is a stationary process. Stationarity means that the covariance between two points does not change depending on the spatial location of the points, and only the lag (or distance) h between the two will affect the covariance (Priestley, 1981). In general, element $a(s_i)$ shows the contribution of stream segment i, the segment on which the random variable is

being measured, to the value of the random variable.

The second element, $b(s_i)$, is again made up of integrals of moving average functions over white noise processes, but each is now supplemented by a weighting parameter calculated as $\prod_{k \in B_{s_{i,[j]}}} \sqrt{\omega_k}$. The variable ω_k will be discussed in greater depth later on, but for now it will be regarded simply as a weight for each stream on the network. The set $k \in B_{s_{i,[j]}}$ is the set of all stream segments on the river network that are between segment i and segment j (inclusive of j but excluding i). Therefore the weight is the product of the weights on all the stream segments that are upstream of segment i *and* downstream of segment j. Overall, $b(s_i)$ is the sum of the contributions of all stream segments upstream of i, weighted according to their likely contribution to the random variable.

From this general form, it is then possible to define a class of autocovariance models which will be suitable for use with stream instead of Euclidean distance. Ver Hoef et al. (2006) and Ver Hoef and Peterson (2010) state that these models take the form shown in (1.10), where h_{str} is the stream distance between points s_i and t_j.

$$C(h_{str}|\theta) = \begin{cases} 0 & \text{if s and t are not flow connected} \\ \Pi_{k \in B_{s_i,t_j}} \sqrt{\omega_k} C_u(h_{str}) & \text{otherwise} \end{cases} \tag{1.10}$$

$$C_u(h_{str}|\theta) = \begin{cases} \theta_0 + \theta_1 & \text{if } h_{str} = 0 \\ \theta_1 \exp(-\frac{h_{str}}{\theta_2}) & \text{if } h_{str} > 0 \end{cases} \tag{1.11}$$

In this model $C_u(h_{str})$ is the standard formulation (used in Euclidean based models) of the chosen covariance function. For example, the exponential function is shown in equation (1.11). The weights ω_k are the only remaining part of the

model which have not been mentioned in any detail. Ver Hoef et al. (2006) suggest that the best way to define ω_k is to use flow volume. If there is a 'join' in the river network, where two streams join to create just one (presumably) larger stream, then if the flow volume is recorded at both of the feeder streams ω can be defined at each of the feeder streams as the proportion of the contribution they make to the overall volume after the join. For example if stretches 1 and 2 have volumes μ_1 and μ_2 respectively, then $\omega_1 = \frac{\mu_1}{\mu_1+\mu_2}$. However, comprehensive flow data, whether observed or simulated, may be quite difficult to acquire, and so Ver Hoef et al. (2006) suggest that in that case a proxy variable such as the 'Stream Order', which will be defined in Section 1.3.2, could be used.

The autocovariance structure in (1.10) has some properties which, at least intuitively, seem very desirable in the context of a river network. Firstly a covariance of zero is assigned when the two locations in question are not flow connected, meaning that we assume that the value of a random variable at one location will not affect the value at another if neither flows into the other. In addition to this, the weighting $\Pi_{k\in B_{s_i,t_j}}\sqrt{\omega_k}$ used to define the covariance between two flow connected stations in different locations is constructed so that less weight will be given if the upstream station is on a relatively minor stretch of water. Specifically the weightings are constructed by assessing each point of confluence on the network, by calculating what proportion of the water after the join comes from each of the meeting streams and then taking the square root. To calculate the weighting between two points on the river, the product of the square root of the proportions on all streams that lie between the two points in question is taken. This means that, while models developed for use with Euclidean distance specify the covariance as a function of the distance between stations, the tail-up model will reduce the covariances if a station is on a very small stretch, even if the stations are quite close together in terms of stream distance.

Weighting in the "Tail-Up" Model

The weighting structure $\Pi_{k \in B_{s_i,t_j}}$ used in the tail-up model, shown in (1.10), allows existing covariance models to be used in conjunction with stream distances. However, the weighting is more than just an elegant solution to that particular problem as it allows a very mechanistic and intuitive way of describing the flow of water on a river network. Consider two streams i and j that meet and form a single river k. Now consider the points s_i on i and t_j on j located just before the streams join, and point u_k on stream k just after the point of confluence. When dealing with the raw total of a pollutant at a certain time (i.e. total weight of nitrate, for example) then it makes sense for the total at u_k to be the sum of the totals at s_i and t_j. On the other hand, if the data were the concentration of a pollutant, it would seem much more intuitive to have a weighted average of the concentrations at s_i and t_j. This is essentially what the weighting in (1.10) allows.

The interesting decision to make here is what to base the weights on. Several variables could be used, such as basin volume or cross sectional areas at s_i and t_j; however, the most suitable option would appear to be flow. Flow is a measure of the volume of water passing a point in unit time and this makes it a more informative choice than either the volume of water or the velocity of water at a point. The impact of flow data would be even greater in the context of spatio-temporal modelling if it were possible to dynamically alter the weightings used in the autocovariance structures over time according to flow data. This would be very expensive in both financial and computational terms. The use of flow data seems the most desirable option, and indeed Ver Hoef et al. (2006) suggest flow volume as ideal weights with surrogates such as basin volume or stream order suggested as possible proxies if flow data were unavailable.

Stream order has two different definitions. The first is known as 'Strahler Stream Order' (Strahler, 1952), and it designates as first-order the smallest or "finger-tip" channels (the streams that are unconnected at one end), as second-order the streams "formed by the junction of any two first-order streams", and as third-order the stream "formed by the joining of any two second-order streams", and so on. The second is known as 'Shreve's Stream Order' (Shreve, 1967) and it again designates as first order the "finger-tip" channels but this time stream orders are subsequently added at the points of confluence. For example, a tenth order stream meeting a fifth order stream creates a fifteenth order stream. Shreve's stream order is used for analysis in Cressie et al. (2006) in lieu of information on flow or stream topography, and this approach is shown by Ver Hoef and Peterson (2010) to be equivalent to the weighing scheme given in Equation (1.9). Indeed, Ver Hoef and Peterson (2010) point out that the stream order approach has the benefit of being less computationally intensive as only n stream orders need be computed and stored as opposed to an $n \times n$ matrix of weightings, a subset of which must then be multiplied together when using flow data or its equivalent. Preliminary analysis on the data in this thesis was carried out using Shreve's Stream Order in place of flow data, and the results did not differ substantially from the results subsequently obtained using flow volumes. Both this and the fact that Shreve's Stream Order is very different from Strahler Stream Order, imply that Shreve's stream order is a much better surrogate for flow data, and will produce results quite close to those that would be obtained using actual flow data.

The use of flow data, while slightly more computationally intensive, has the potential to be far more descriptive than stream order. Its potential in both the modelling of tail-up moving average models and in other modelling techniques does not yet seem to have been fully realised, and any potential development of a more mechanistic class of model for river networks will surely use flow data

even more prominently. For instance, if the tail-up model were to be extended into a space-time setting, it seems intuitive that the tail-up weightings would be able to change over time and depend on flow volumes. This would depend on flow data being available as a time series in all necessary locations. This may prove prohibitive in practice, but an alternative would be to model the flows to fill in monitoring gaps. This demonstrates the potential of flow data in this area of analysis.

Tail-Down Models

Tail-down models, introduced in Ver Hoef and Peterson (2010), allow correlation between flow-unconnected locations while still using stream distance. These are most effective when used in conjunction with tail-up models to create variance component models (Ver Hoef and Peterson, 2010; Cressie and O'Donnell, 2010). The tail-down class of models seem to be of most practical use when dealing with scenarios where it would be desirable to have correlation between flow-unconnected locations, such as when dealing with organisms that can swim against the flow. This makes it suitable for use with data on fish, or chemical determinands, for example a waste product such as ammonia, that are linked to fish and other living organisms. The tail-down model therefore seems unsuited to the context of nitrates data.

The definition of stream distance for these models is the same as before and so the distance between flow unconnected sites is calculated by finding the distance from one site to the nearest common point of confluence and back upstream to the other site. For flow connected sites the covariance is the same as in the tail-up model but without the weighting structure $\Pi\sqrt{\omega_k}$. This can be seen in (1.12), which shows, as an example, the covariance under the tail-down exponential

model.

$$C_d(h_{str}|\theta) = \begin{cases} \theta_0 + \theta_1 & \text{if } h_{str} = 0 \\ \theta_1 \exp(-\frac{h_{str}}{\theta_2}) & \text{if } h_{str} > 0 \end{cases} \tag{1.12}$$

Unlike the tail-up models, tail-down models have limited use on their own (Ver Hoef and Peterson, 2010; Cressie and O'Donnell, 2010) as "flow-unconnected sites have more autocorrelation than flow-connected sites". This means that two locations that are a stream distance of h apart but are not connected by flow would have a *higher* correlation under the tail-down model than two streams a distance of h apart that *were* connected. This would seem to be unrealistic for most if not all determinands and so Ver Hoef and Peterson (2010) suggest using the tail-down model as part of a variance component model in order to allow for a wider range of ratios between flow-connected and flow-unconnected autocorrelation.

1.3.3 Variance Component Models

The use of variance component models incorporating stream distance based covariance structures such as the tail-up and tail-down models was first suggested in Cressie et al. (2006). This paper proposed a covariance structure such as that shown in (1.13), which is a weighted average of the tail-up covariance structure $C_u(s_i, t_j|\theta)$, as shown in (1.10), and the Euclidean distance based covariance structure $C_{euc}(s_i, t_j|\theta)$, as shown in (1.3).

Cressie (1991) states that "if $C_1(\cdot)$ and $C_2(\cdot)$ are two valid covariograms in $\mathbb{R}^d$ then ... $C_1(\cdot) + C_2(\cdot)$ is a valid covariogram in $\mathbb{R}^d$ ", and "for $b > 0$, $bC_1(\cdot)$ is a valid covariogram in $\mathbb{R}^d$". This means that a weighted average of two valid

covariograms will itself be a valid covariogram.

$$C_{mix}(s_i, t_j|\theta) = \lambda C_u(s_i, t_j|\theta) + (1-\lambda)C_{euc}(s_i, t_j|\theta), \quad \lambda \in [0, 1]. \tag{1.13}$$

Ver Hoef and Peterson (2010) use the concept of basing a covariance structure on multiple covariance models in order to build a more widely applicable model incorporating the tail-down covariance structure. On its own, the tail down structure means that "flow-unconnected sites have more autocorrelation than flow-connected sites" (Ver Hoef and Peterson, 2010). However, when used in conjunction with the tail-up model, a much wider range of flow connected to flow unconnected covariance ratios can be obtained. This model is not limited to two components, and Peterson and Ver Hoef (2010) fit mixtures that include tail-up, tail-down and Euclidean components. Ver Hoef and Peterson (2010) show the necessity for variance component models when using the tail-down covariance structure in order to appropriately balance out the covariances of connected and unconnected locations. In addition, Peterson and Ver Hoef (2010) and Cressie et al. (2006) suggest that a mixture of covariance structures in a variance component model may increase the accuracy of predictions too.

1.3.4 Spatial Prediction Via Kriging

Kriging is a form of spatial prediction that interpolates between previously observed locations in order to predict at new locations. There are different types of kriging, which make different assumptions about the spatial trend underlying the data. Ordinary kriging makes the assumption that there is a constant but unknown mean underlying the spatial process. Simple kriging could be used in

the case where we assume that the data has mean zero, however ordinary kriging will be used to give a little more flexibility. Another option would be to use universal kriging, which assumes that the underlying trend of the data follow a simple polynomial function. This is the method used by Ver Hoef et al. (2006), but a simple polynomial may be too general to account for the trend in the Tweed data. The polynomial function will be based on Euclidean rather than stream distances, and so other detrending options will be explored in Section 2.1 and then predictions will be calculated from the detrended data using ordinary kriging. The equations quoted below (all taken from Cressie 1991) are defined in terms of the covariance function in order to remain consistent with the stream distance model theory discussed in Section 1.3.2. It is more common to see these formulae defined in terms of the variogram, but they are equivalent. In fact, if (1.7) is used to transform the covariogram into a variogram then it is possible to use the variogram based kriging equations even with stream distance based data.

Given data $Z \equiv (Z(s_1), ..., Z(s_n))'$, collected at known spatial locations $s_1, ..., s_n$, denote the generic predictor of $g(Z(.))$ by $p(Z; g)$. Ordinary kriging is defined as spatial prediction under the assumption that the data Z come from model 1.14 and have predictor 1.15, where B is the one or two dimension space we wish to predict in.

$$Z(s) = \mu + \delta(s), \quad s \in D, \mu \in R, \text{ and } \mu \text{ unknown} \tag{1.14}$$

$$p(Z; B) = \sum_{i=1}^{n} \lambda_i Z(s_i), \quad \sum_{i=1}^{n} \lambda_i = 1 \tag{1.15}$$

The optimal predictor $p(.; B)$ will minimise the mean squared prediction error (1.16) over the class of linear predictors $\sum_{i=1}^{n} \lambda_i Z(s_i)$ that satisfy $\sum_{i=1}^{n} \lambda_i = 1$.

This can be written as (1.17), as long as $\sum_{i=1}^{n} \lambda_i = 1$.

$$\sigma^2 \equiv E(Z(B) - p(Z;B))^2 \tag{1.16}$$

$$\begin{aligned}(Z(s_0) - \sum_{i=1}^{n} \lambda_i Z(s_i))^2 = \quad & (Z(s_0) - \mu)^2 \\ & + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j (Z(s_i) - \mu)(Z(s_j) - \mu) \\ & -2\sum_{i=1}^{n} \lambda_i (Z(s_0) - \mu)(Z(s_i) - \mu)\end{aligned} \tag{1.17}$$

Now we suppose that (1.14) holds, with $\delta(.)$ a zero mean second order stationary process having covariogram $C(h), h \in R^d$. Then we minimise (1.18) with respect to $\lambda_i : i = 1, ..., n$ and m in order to obtain the ordinary kriging equations, given in (1.19) and (1.19). Here, λ' is defined as in (1.21), while m is defined in (1.22).

$$C(0) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(s_i - s_j) - 2\sum_{i=1}^{n} \lambda_i C(s_0 - s_i) - 2m(\sum_{i=1}^{n} \lambda_i - 1) \tag{1.18}$$

$$\hat{p}(Z; s_0) = \lambda' Z \tag{1.19}$$

$$\sigma_k^2 = C(0) - \lambda' c + m \tag{1.20}$$

$$\lambda' = \left(c + 1\frac{(1 - 1'\Sigma^{-1}c)}{1'\Sigma^{-1}1} \right)' \Sigma^{-1} \tag{1.21}$$

$$m = \frac{1 - 1'\Sigma^{-1}c}{1'\Sigma^{-1}1} \tag{1.22}$$

In these equations, $c \equiv (C(s_0 - s_1), ..., C(s_0 - s_n))'$ and Σ is an $n \times n$ matrix whose (i, j)th element is $C(s_i - s_j)$.

Therefore, predictions are made using a weighted average of the observed stations, with the weights proportional to the distance between the observed and new locations. Using the covariogram-based kriging equations rather than the variogram requires the stronger assumption that the process $Z(\cdot)$ be second-order stationary (Priestley, 1981). This assumption means that in addition to the assumption associated with first order stationarity, it is also assumed that there exists a covariogram function and thus $Var(Z(s)) = C(0)$. First order stationarity assumes the existence of the variogram, but the assumption of the existence of the covariogram is a stronger assumption. Ordinary kriging also assumes that the weights shown in (1.21) sum to 1, to guarantee "uniform unbiasedness" (Cressie, 1991), however this assumption is automatically satisfied if a valid autocovariance structure, such as those discussed in Sections 1.3.1, 1.3.2 and 1.3.3, is used.

1.4 Thesis Overview

The investigation to be undertaken will aim to explore the predictive accuracy of the tail-up model, and variance component models involving it, using the River Tweed dataset. It will also aim to adapt the tail-up model in order to fit a stream distance based spatio-temporal model for the first time.

Chapter 2 will focus on trend in the data. It will discuss methods of extracting a nonparametric trend, based on either Euclidean or stream distance, in order to better account for the trend than would be done by using ordinary or universal kriging with the original data. The remainder of the chapter will look at how to estimate the covariance parameters needed for kriging, and will discuss the potential problems in this process posed by use of the tail-up model.

Chapter 3 will explore which factors have the most effect on the predictive accuracy from kriging by carrying out four studies into different elements of the models. The first study will investigate whether a mixture of covariance structures will provide more accurate predictions for the Tweed data, as has already been suggested in the literature, and what mixture is likely to provide the lowest prediction errors. The second study will investigate the role that different covariance models and parameters have in the prediction and will decide whether significant decreases in prediction error can be obtained with different combinations of these. The third study will then investigate the trend surfaces fitted in Chapter 2, and will carry out a study to see whether the bandwidth used in the smooth function used to generate them will affect the prediction accuracy. The final study will simulate data using the River Tweed geography as a template. From this, it will try to understand better the impact that different sampling schemes would have on the prediction errors and the conclusions for the real dataset.

Chapter 4 will use the covariance parameters and models estimated in previous chapters to predict the nitrate levels at unsampled locations on the River Tweed using the yearly averages for each of twenty-one years worth of data. It will discuss the impact that using Euclidean or stream distance for either the trend or covariance structure has had on the fitted nitrate value, and then assess how the average nitrate levels are changing over the years.

Finally, Chapter 5 will adapt the tail-up model to allow a stream distance based smoothing function to be used for the spatial component of a spatio-temporal additive model. Additional space-season and space-time interaction terms will be defined and the best combination of these terms will be assessed.

Chapter 2

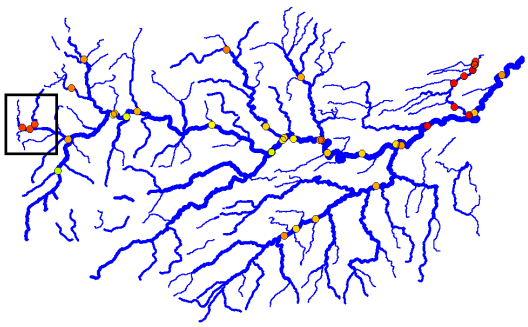
Estimating Trend and Covariance

This chapter will attempt to estimate a nonparametric trend for nitrate levels and will allow the trend to be based on stream distance as well as the standard Euclidean distance. This will involve the use of a novel smoothing function based on the tail-up model. Prediction using kriging requires the covariance structure to be estimated. This chapter will detail the procedure used to estimate covariance parameters for the River Tweed data, and the problems that are faced when the covariance structure is based on stream, rather than Euclidean distance.

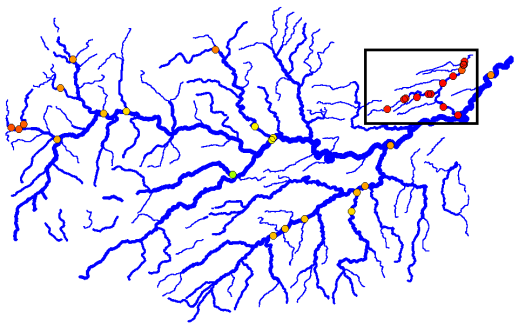
2.1 Spatial Trend

As discussed in Section 1.3.4, Ordinary Kriging using covariogram models requires the assumption of second order stationarity. Second order stationarity implies that the underlying mean of the spatial process on the Tweed is constant over the entire river and that the covariance $C(s_i, t_j)$ depends just on h, the distance (or 'lag') between s_i and t_j and not the positions of s_i or t_j. It is difficult to assess the accuracy of the latter of these two assumptions, so it may be

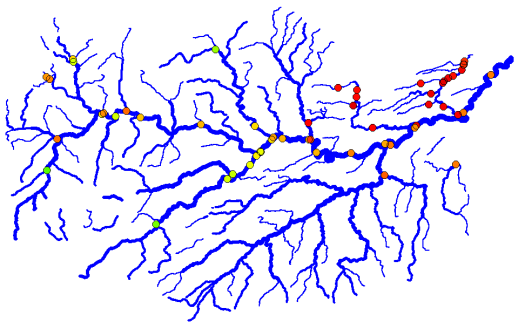
necessary to pragmatically accept this and say that it would be expected that the covariance structure underlying the data would be constant. The first assumption of a constant (but possibly non-zero) mean over the whole river network is easier to assess, and we can examine this using Figure 2.1.



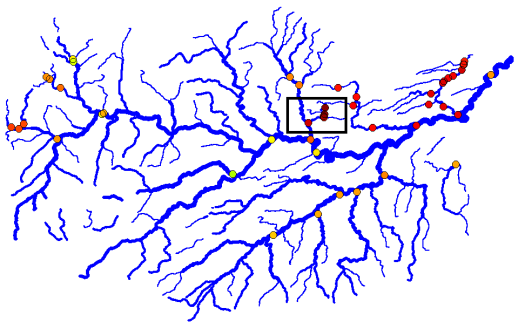
(a) March 1987



(b) May 1988



(c) February 2002



(d) November 2005

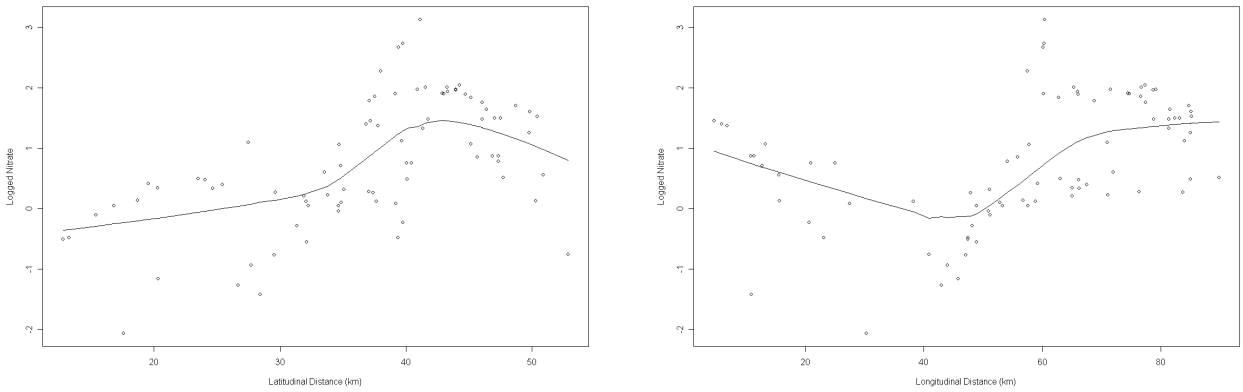
Figure 2.1. Spatial plot of nitrate levels at selected dates

Figure 2.1 shows the Nitrate levels at available stations at four time points. It is clear, even from just four plots, that certain river stretches have consistently higher values than certain others no matter what time point is examined. The Leet (the stream segments shown inside the box in Figure 2.1(b)) contains the most consistently high values, something that looked to be the case when the trends in the data were analysed in Section 1.2.1. While the Leet has generally

quite high levels of Nitrate, they do not cross the Nitrates Directive thresholds on many occasions. This is in stark contrast to Turfford Burn and its tributary (shown in the box in Figure 2.1(d)), which consistently crosses the upper and lower thresholds when data are present. The only other area of consistently high nitrate levels is the cluster of stations to the very far west of the network (shown in the box in Figure 2.1(a)). These areas have nitrate values which are consistently at the higher end of the scale, no matter what month of data is examined, implying that the high mean values are due to spatial trend rather than the seasonal cycle. These areas of high values, contrasting with the generally low values elsewhere on the network, suggest that the assumption of a constant mean over the entire network is not valid unless the underlying spatial trend is removed prior to analysis. Trend would not need to be removed if it could be modelled as a parametric function via universal kriging. This is an unrealistic model for the trend observed on the River Tweed data and would not be able to provide a stream distance based trend without being adapted to suit this scenario.

The assumption of constant trend can be examined in more detail by using the plots shown in figure 2.2, which show the average logged Nitrate values of each of the stations over the entire 21 year period plotted against their latitude and longitude respectively. A loess smooth of the data has been added to the plots to make it easier to identify patterns.

If no spatial trend were present, these graphs would show a scatter of points around a mean line. Since there is a definite pattern to both it would appear that this is not the case. In Figure 2.2(a) it seems that there is an increase in nitrate levels from south to north. Similarly Figure 2.2(b) suggests that the higher nitrate values are found in the far east and west of the network, with stations in the middle having the lowest values. Looking at these plots is merely an exploratory technique to identify whether there is a spatial trend present in



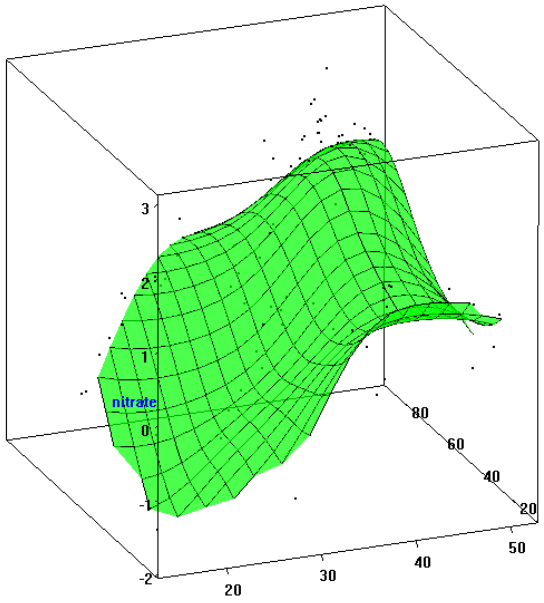
(a) North/South trend evaluation (b) East/West trend evaluation

Figure 2.2. Nitrate against spatial location, used to detect presence of spatial trend.

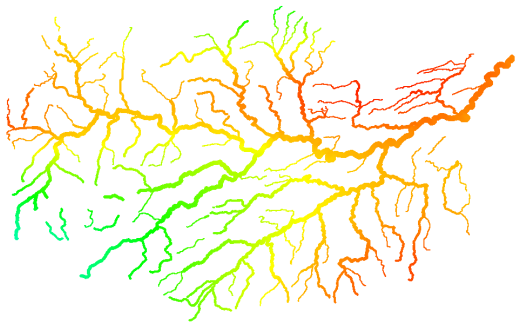
the data, and the evidence suggests that there is a spatial trend that should be removed before any analysis is performed.

The trend that has been observed here may possibly be explained by the land use in each of the areas. Even without any formal land use data, it is possible to identify most of the high nitrate stations as being in areas with a lot of farming activity, and that lower levels are generally found in the more upland areas. In the absence of such a formal classification, the spatial trend will be removed by fitting quite general models, however it is worth noting that land use may be better able to explain the trend.

It was decided that a parametric trend would not capture the features of the data as well as a nonparametric trend. Universal kriging would fit a parametric trend to the data and so this decision means that this method of prediction is unsuitable. Consequently, it was decided to create a simple non-parametric smooth estimate using R's 'sm' library (Bowman and Azzalini, 2007), and then predict at unsampled locations using ordinary kriging on the detrended data. A local linear method of trend estimation (Bowman and Azzalini, 1997) was used, with isotropic smoothing and with the effective degrees of freedom set to 12. This



(a) Plot showing smooth fitted to averaged data



(b) Plot of smoothed values on river network

Figure 2.3. Estimated Euclidean nonparametric trend

offers a reasonable compromise between flexibility over space and the retention only of large scale effects. The results are robust against modest changes in the degrees of freedom. Using this method, the estimator for the spatially smoothed data $\hat{m}_{euc}(x)$ for some point x is the least squares estimator $\hat{\alpha}$ shown in (2.1). Here, $w(x_{ji} - x_j; h_j)$ is a normal kernel density as given by (2.2) where x_{ji} and x_j are the coordinates of points x_i and x in the j^{th} dimension.

$$\min_{\alpha,\beta,\gamma} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_{1i} - x_1) - \gamma(x_{2i} - x_2)\}^2 w(x_{1i} - x_1; h_{s_1}) w(x_{2i} - x_2; h_{s_2}) \quad (2.1)$$

$$w(x_{ji} - x_j; h_j) = exp(-\frac{1}{2}\frac{(x_{ji} - x_j)^2}{h_j^2}) \quad (2.2)$$

$$\min_{\alpha,\beta,\gamma} \sum_{i=1}^{n} \{y - \alpha 1_n - X\beta\}^T W \{y - \alpha 1_n - X\beta\} \tag{2.3}$$

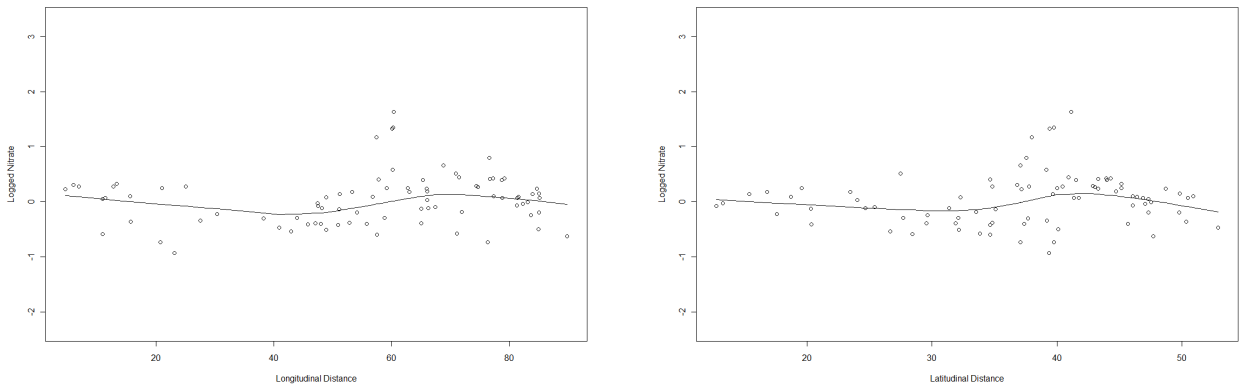
Equation (2.1) can be rewritten in vector matrix form as shown in (2.3). Here, X is a matrix with three columns consisting of a column of 1's, a column of the difference in coordinates in one dimension $(x_{1i} - x_1)$ and a column of the difference in coordinates in the other dimension $(x_{2i} - x_2)$ respectively. The matrix W is a diagonal matrix with elements $w(x_{1i} - x_1; h_{s_1})w(x_{2i} - x_2; h_{s_2})$ down its diagonal and zeros elsewhere. Reformulating the model as $Y = X\theta + \epsilon$ and fitting this locally allows the system to be solved using (2.4). The smoothed value at point x is then given by the value in the first row, and first column of the matrix $\hat{\theta}$.

$$\hat{\theta} = (X^T W X)^{-1} X^T W y \tag{2.4}$$

The bandwidths in each dimension are allowed to differ and are defined as h_{s_1} and h_{s_2}, where s_1 and s_2 are the sample standard deviations in each dimension. This means that the bandwidths are a function of the standard deviations, with the same function used in each dimension. The estimated smooth function uses a bandwidth of around 14km in the East-West dimension and around 6km in the North-South dimension, corresponding to 12 degrees of freedom. A thorough study of the effect of different bandwidths will be conducted in Section 3.3, but this set of parameters was used as it offers a reasonable compromise between picking out the individual 'hotspots' and producing a more generally smooth estimate. A range of other bandwidths were tested, but visually these seemed to produce the best trade-off between too general and too well-fitting a trend. Being a Euclidean distance based trend, the structure of the network is not taken into account and thus Figure 2.3(a) shows the fitted smooth curve without reference to the network, while Figure 2.3(b) shows how that is expressed as values on the

river network.

First compare the plots of the original data in the North-South and East-West directions in Figure 2.2 to those with the detrended data, obtained by subtracting the estimated trend from the original data, in Figure 2.4. Figure 2.4 is shown on the same scale as Figure 2.2 in order to highlight the reduction in trend that has been brought about by detrending. The Euclidean distance based detrending seems to have reduced the spatial trend present in the data. The loess smoothed curves in both plots are much flatter for the detrended plots and this suggests that the detrending process has been reasonably successful in eliminating trend. While neither plot is perfectly flat, suggesting that there is still some residual trend left in the data, they are both an improvement on the plots in Figure 2.2. In fact, it is the influence of just one or two outliers (located at Turfford Burn) that drag up the loess line and make it seem that there is still residual trend. If these points were excluded, the assumption of zero trend in both directions would seem very plausible. The detrended data is therefore more likely to satisfy the assumption of isotropy necessary for kriging than the data for which no trend has been removed.



(a) Nonparametric fit: East/West trend (b) Nonparametric fit: North/South trend

Figure 2.4. Plots to detect spatial trend still present in detrended data

There is a very fine line between what constitutes trend and what is actually

part of the covariance structure we are seeking to describe. For example, if there is a point source discharging nitrate into a river then it will cause high values close to the source, before being diluted as you move further downstream. This raises the question of whether this be incorporated into the fitted trend? While it could easily be regarded as such, another view to take is that the high values are being caused by the point source and thus are due to the close proximity of the sampling locations to the discharge. That would suggest that it was the autocovariance structure at work. In this example, the distinction is almost certainly made depending on whether the point source outputs a consistent amount of nitrate. If the levels at the point source were observed at ten time points and a huge amount of variation was seen then the conclusion would probably be that this formed part of the covariance structure rather then trend. However, if the ten values were all roughly the same, then the best course of action would seem to be to regard this as trend. The key idea here is that spatial trend is likely to persist over time while random fluctuations will not. Therefore a reasonably general trend, such as that fitted here, will avoid modelling the data too closely, and thus be less susceptible to picking up some of the covariance structure. Later analysis will show that this is a moot point, as trend seems to balance out with the covariance structure so that certain elements of the underlying process are picked up by each.

Given how unsuitable a parametric trend seems in this context, and the importance of detrending in the statistical theory that underpins this type of analysis, it is likely that one would want to make at least some attempt to address this issue. However, this issue has not been described in the literature that uses the tail-up model. The literature has tended to use the parametric trends that form part of universal kriging predictions. Given the data observed on the Tweed, and indeed many of the example datasets used in the literature, the assumption that a parametric surface will adequately account for the trend in the data seems dubious. However, in many of these examples this could be explained by the lack

of data over a long period of time. For example, Ver Hoef et al. (2006), Ver Hoef and Peterson (2010), Peterson and Ver Hoef (2010), Garreta et al. (2009) and Guillaume Blanchet et al. (2008) use data from a single time point, while Cressie et al. (2006) consider just two years worth of yearly averaged data. The lack of long term data is likely to make trend identification problematic and so this could be the reason that detrending is not considered in these examples. A notable exception in this is Clement (2007), which use data collected over a fourteen year period. The focus of their analysis here though is on spatiotemporal modelling of data at individual sites of a small catchment of a river network, rather than spatial prediction as seen in much of the rest of the literature. Spatial trends are not the focus of their analysis as a very small catchment is used.

2.1.1 Stream Distance Based Trend

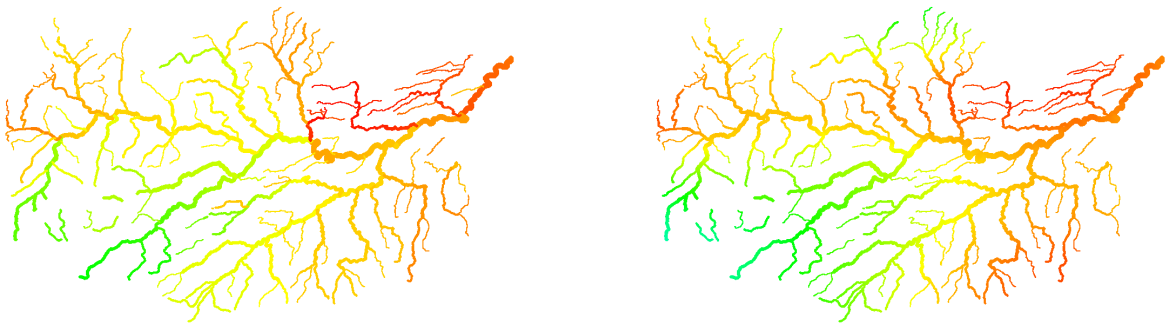
As literature in the area of modelling river network data has focused on using stream distance to come up with a more intuitive covariance structure, it is a natural extension to use stream distance in order to fit the trend. This avenue of investigation has not been explored before, as the literature has not considered the option of detrending using either distance metric prior to analysis. This section will detail a novel approach to using stream distance to construct a trend based on the tail-up model. The tail-up stream distance based covariance model was designed with the intention of being able to incorporate characteristics such as stream distance, relative size of rivers (via flow data) and whether locations were flow connected, in order to produce a more accurate reflection of the processes going on in that particular river. However this kind of structure is not accounted for in the detrending process that was used in the previous section. It would seem more appropriate to detrend the data using stream distance as the distance metric and to take into account the flow-connectivity network.

In order to detrend the data in a way that would remain more faithful to the ethos of the tail-up model, a novel non-parametric smooth estimate was constructed by estimating the trend at point x according to (2.5). This is based on a very simple local mean smooth function. A local linear approach was also considered for constructing this trend, but it had the drawback that it was not possible to estimate a trend on any streams more than one bandwidth distance away from all of the monitoring stations. There is an argument that one would not wish to make predictions at such streams anyway as there is little information on which predictions can be based. However, in order to have complete results on the network the trend was estimated using (2.5). Here, $w(d;h)$ is a Normal kernel density function N(0,h) and is given by (2.2); y_i is the overall average nitrate level at station i; d is the stream distance between point x and station x_i; $\delta_i(x)$ is an indicator function which takes value 1 if station i is flow connected to point x and 0 if it is not. The h parameter in the weight equation corresponds to the desired bandwidth.

$$\hat{m}_{riv}(x) = \frac{\sum_{i=1}^{n} y_i w(d;h)\delta_i(x)}{\sum_{i=1}^{n} w(d;h)\delta_i(x)} \tag{2.5}$$

This predictor can also be formulated in a similar, though simpler fashion to that shown for the Euclidean distance trend in (2.1) and can again be rearranged into vector matrix form so that it can be solved using (2.4). In this equation, when using stream distance, X now denotes a vector of 1's, while W has elements $w(d;h)$ on the diagonal where d is the stream distance between point x and point x_i. This will be explicitly stated and discussed in Section 5.1.1. It is worth noting that here, both stream distance and the indicator function are symmetric, so that the stream distance between points A and B is equal to the distance between points B and A, while the indicator function will just depend on whether the two sites are flow connected and not where the points lie in relation to one another.

It should be noted that neither the indicator function, $\delta_i(x)$, nor the weight, $w(d;h)$, contain the flow related weightings included in the tail-up model. It was initially considered that these could form part of the indicator function, and such an approach is adopted for the smoothing function used in the additive models in Section 5.1.1, but this approach was ultimately rejected for defining the trend here. Including the flow based weightings seemed to go against the wish to keep the trend fairly general. This is because the smooth estimate is much more dominated by nearby observations as further away observations tended to be after one or several points of confluence and the flow weighting then attaches much less importance to them. Fitted values incorporating flow based weightings tend to track the data too closely and so it seemed inappropriate for the estimated trend. This property may be desirable as part of the covariance structure in kriging, but is not ideal when the object is to keep the trend general.



(a) Stream distance based trend (b) Euclidean distance based trend

Figure 2.5. Estimated stream and Euclidean distance based trends

Figure 2.5(a) shows the stream distance based trend fitted to the Tweed data with a bandwidth of 15km. This was chosen to provide a reasonable balance between generality and accuracy. As mentioned previously, Section 3.3 explores the effect of changing bandwidth in more detail, and suggests that the results of kriging are not very sensitive to small adjustments in bandwidth.

Comparing Figure 2.5(a) to the estimated Euclidean distance based trend, replicated in Figure 2.5(b), characterises the differences that are seen when using a flow connected stream distance, as opposed to a Euclidean based approach. The Euclidean approach produces a surface that takes no account of the structure of the river. In the stream distance based approach, the effect of the connectivity can be clearly seen, with some minor streams in the centre of the plot tending to the mean value (an orange colour) despite all the surrounding streams being a much lower (green) value. This reflects the fact that the low nitrate levels that cause the surrounding streams to have low values are observed at locations that are not flow connected to these minor streams. Also, the stream distance based trend has some quite sharp changes in nitrate levels at points of confluence. This is another feature that is brought about due to the flow connectivity information, and is likely to better reflect the behaviour of a river network, rather than the smooth changes seen in the Euclidean trend that occur regardless of the locations of the joins in the stream segments.

Neither nonparametric trends nor trends based on stream distance have been considered in the literature involving prediction on river networks using the tail-up model. Literature such as Ver Hoef and Peterson (2010) has tended to account for tend in the data using universal kriging to fit a simple polynomial function. Given the irregular nature of environmental data in general, a nonparametric trend would seem to be an obvious extension of existing work. A trend based on stream distance would also seem an intuitive direction in which to extend this work. Later chapters will investigate furtherthe possibility of using a nonparametric, stream distance based trend for the first time, and will aim to assess the benefits it would bring over a Euclidean distance based trend.

2.2 Choosing a Covariance Structure

As discussed in Section 1.3.3, it is possible to model the covariance structure of the data using a tail-up, tail-down or Euclidean model as well as a mixture model including two, or all three of these constructs. In deciding which structure or mixture to use, it is important to keep in mind that it is Nitrate data that will be analysed, as the most appropriate model will undoubtedly depend on the properties of the variable in question.

Both Peterson and Urquhart (2006) and Cressie et al. (2006) use a model based on Euclidean distance to model dissolved organic carbon and change in dissolved oxygen respectively. Peterson and Urquhart (2006) only consider the Euclidean model, as the tools for analysing using the tail-up model were still being developed at the time of publication. However, Cressie et al. (2006) consider, for the first time, a mixture of tail-up and Euclidean covariance models. They conclude that the best mixture of covariances (determined by likelihood methods) put all of the weight in the model on the Euclidean covariance structure, meaning that the covariance structures used in both Peterson and Urquhart (2006) and Cressie et al. (2006) are identical.

Peterson and Ver Hoef (2010) analyse data on "the proportion of native fish species expected (PONSE), which is simply the ratio of observed to expected native freshwater fish species richness", and test a variety of different covariance constructs. The data are from a river network located in South East Queensland, Australia. For each model on its own (tail-up, tail-down and Euclidean) as well as each possible mixture model, the parameters for four different types of covariance model (linear with sill, exponential, spherical and MARIAH (Ver Hoef et al., 2006)) were estimated. The lowest root mean squared prediction error (RMSPE), generated using leave-one-out cross-validation, was reported for each of the three

different structures and the four possible mixtures of them. The lowest RMSPE (0.2088) was found using a mixture of the tail-up and tail-down models, while the next lowest (0.2094) used all three and the next lowest after that (0.2103) used a mixture of tail-down and Euclidean. All of these models use the tail-down construct, and that is not surprising given that the data being analysed here are based on the number of fish at various points on the river. The tail-up model would assign a covariance of zero to flow-unconnected (i.e. not directly connected by the flow of the river, but connected via a point of confluence further downstream by going back upstream against the flow) locations on the river and this is not realistic in a scenario where the fish are able to swim between such locations.

Ver Hoef and Peterson (2010) use two examples to demonstrate the use of a mixture of tail-up and tail-down models in a covariance structure. These examples use pH and conductivity ("the ability of a solution to carry an electrical charge based on ion concentration and temperature") data, again collected from a river network in South East Queensland but on a different stretch of river to that analysed in Peterson and Ver Hoef (2010). Despite having data from the same part of the world these papers vary considerably in their approach, with Ver Hoef and Peterson (2010) fitting just a mixture of tail-up and tail-down models to each of the examples, and focusing more on describing the estimated parameters in the model and what they mean in terms of the data. The conductivity data is used to explain the process of estimating the covariance parameters and, despite commenting that there is "relatively strong autocorrelation among flow-connected sites with somewhat weaker, but still substantial, autocorrelation among flow-unconnected sites", there is no indication of how important the tail-up and tail-down elements of the model were in relation to one other. The pH example goes slightly further in the analysis of the data on this network by obtaining the covariance structure and then subsequently using it to fit models

including a range of environmental covariates. For the pH data, it is stated that the tail-up model accounts for around 85% of the variation explained in the model of the spatial covariance and that there is "only slight evidence of spatial autocorrelation between flow-unconnected sites" from the empirical semivariogram. The justification that is offered for the greater importance of the tail-up component is that pH is strongly influenced by the direction of flow in the river and so it is unlikely that there would be a high correlation between flow-unconnected sites.

Garreta et al. (2009) use temperature and nitrate data from the Meuse and Moselle basins in the North East of France to further investigate possible covariance structures. For each outcome, a tail-down model using two different covariance structures (exponential and spherical), a tail-up exponential model and a mixture model with tail-up and tail-down elements were fitted and analysed using AIC and MSPE. For the temperature data the optimum AIC and MSPE are obtained using the mixture model, with the tail-up model not too much worse and the tail-down models much poorer. This final conclusion is to be expected given the comments in Ver Hoef and Peterson (2010); Cressie and O'Donnell (2010) regarding the use of the tail-down structure on its own. The nitrate data again have optimum AIC and MSPE when using the mixture model, but this time both tail-down models perform better than the tail-up model. This is a surprising result for two reasons. Firstly, nitrate is a chemical pollutant that would be expected to be carried by the river water downstream and thus more suited to the tail-up covariance structure. Secondly, Ver Hoef and Peterson (2010) suggest that the use of the tail-down structure on its own is of very limited use due to the fact that the covariance between flow-unconnected sites is higher than that of flow-connected sites, and so it is surprising to see the tail-down model on its own outperforming the tail-up structure.

As mentioned previously, Clement (2007) focuses on assessment of data observed at individual sampling locations rather than spatial prediction. The tail-up model is only mentioned as part of a literature review and is not used for either spatial prediction or as part of the spatiotemporal correlation structure used.

The issue to be addressed now is which model, or mixture of models, is best for modeling the River Tweed nitrate data. Part of the thinking behind this decision involved considering the processes at work on the river network in terms of a mass balance equation.

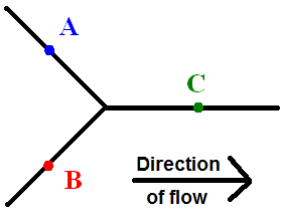


Figure 2.6. Simplified River Structure

Using Figure 2.6 to illustrate the point, consider the processes that would be at work on a simple river network such as this. At each point on the river, pollutants already in the 'system' could come from upstream or downstream while pollutants could be introduced to the system by seeping up through the river bed and whatever soil and rock make it up, as runoff from surrounding land or from the air by falling as rain, being blown by the wind etc. In geostatistics it is required to estimate the correlation (or covariance) between each pair of points on the river and so it is important to consider which of these elements might cause observations at one location to be correlated with observations at another.

First consider points A and C in Figure 2.6 and the context of nitrate data. Firstly the data at points A and C might be highly correlated simply because they are close to one another on the river, and this can be accounted for by using a covariance structure that uses stream distance as opposed to Euclidean

distance. The two locations might also be highly correlated if the composition of the bedrock is similar at both (and thus similar amounts of the determinand are being released at each), if the runoff from the surrounding land is similar or if the pollution being introduced from the air is similar. If sufficient covariate data were available then it is reasonable to think that, in theory at least, it could be possible to account for the bedrock, runoff and airborne elements to leave just the stream distance based process. However in practice it would be difficult to account for some or all of these, but it may be possible to account for them in slightly different ways. If points A and C were, for example, highly correlated due to the fact that the northern bank of that river was farmland and so runoff from the fields would be similar (with possible spikes when it rains after fields have been fertilised) then it could be said that this correlation is due to their proximity in Euclidean space. This might point to a Euclidean distance based process being useful as a surrogate for missing land use covariate data.

In terms of the comparison between the tail-up and tail-down models, the important relationship to consider is that between points A and B in Figure 2.6. The tail-up model would assign a covariance of zero to this pair of points, while the tail-down model would assign a non-zero covariance based on the distance from point A, down to the point of confluence of the two streams and back up to point B. The reality is that neither of these approaches on their own seem a realistic explanation of the processes that might be going on in the real world. One may question why points A and B would be correlated? If a stream distance based process were to make sense then the determinand in question would have to be able to move downstream and then back upstream. This would be very reasonable if the variable in question was the number of fish (or might be heavily influenced by the waste generated by fish moving between streams) but for a variable such as nitrate this does not seem like a particularly realistic scenario. If stations A and B had highly correlated nitrate levels, the most likely explanation

would seem to be that the bedrock in each stream was similar (and thus causing similar effects in each) or more likely that their close proximity in Euclidean space meant that the runoff from the surrounding land was introducing similar levels of the pollutant into each stream. This would again suggest that a Euclidean distance based process would be useful in the modeling of the covariance between two stations on a river network, possibly in order to account for a lack of covariate data. There appears to be more basis for the use of a tail-up structure, instead of tail-down. Combining this with Euclidean distance allows the model to account for a large proportion of the covariance in the data.

Therefore, the Tweed data will be analysed using a mixture of tail-up and Euclidean distance based covariance models. This seems a reasonable decision even when considering the models used in previous literature. Peterson and Ver Hoef (2010) used fish stock data in analysis and so it is not surprising that the tail-down structure improved the predictive accuracy of the model. Garreta et al. (2009) used nitrate data and yet found that the best model included the tail-down structure and the tail-down models on their own were more accurate than the tail-up model. However, there is much more mechanistic evidence to suggest that including a Euclidean distance based covariance structure in a mixture model would be better than using the tail-down structure instead. In Figure 2.6 it is unrealistic that the correlation between points A and B would be zero (as in the tail-up model) and so it would not be surprising if the tail-down model on its own outperformed the tail-up model. However, the tail-down model on its own would define there to be more correlation between points A and B than between A (or B) and C. Therefore it seems more intuitive to define the correlation between points A and B as being due to a Euclidean distance based process. It may then be that the Euclidean process will have quite a small range parameter to reflect the fact that it is accounting for localised land-use-type correlation.

2.2.1 Covariogram Modeling

Having decided on the form of the most appropriate model to use for the River Tweed data, it is now necessary to identify the parameters of the underlying autocovariance structure of the data. As has already been mentioned, this would usually be done using the variogram observed from the data but, due to the more complex structure of the tail-up model, it is the covariogram that will be used for the Tweed data. To understand why, consider the covariance structure defined in (1.10) and, as an example, the exponential structure defined in (1.11), which suggests that the observed covariances in the data are realisations of the function $w_{s,t}\theta_1 \exp(-\frac{h_{str}}{\theta_2})$ where $w_{s,t} = \Pi_{k \in B_{s_i,t_j}} \sqrt{\omega_k}$. This can be converted to the form of a variogram model as shown in (2.6). It is worth noting that $w_{s_i,s_i} = 1$ and so there is no weighting attached to the covariance at lag 0.

Expression (2.6) explains why it is not possible to separate the weighting structure (which is known for all pairs of locations on the river) from the observed semivariances in the data in order to estimate θ_0, θ_1 and θ_2. Using the relationship between the variogram and covariance as the starting point, the covariances defined by the tail-up model for a pair of flow connected locations are substituted in. This means that in the tail-up model structure, the observed semivariances will be assumed to be generated from the equation shown in the last line of (2.6). Therefore, 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. This is not the case if covariance is used. Dividing the observed covariance between two points by their weighting, as shown in (2.7), allows the estimation of the parameters of the model by use of a covariogram. Failure to adjust would seem to lead to bias in the estimation of θ_1 and θ_2 and result in poor descriptions of the underlying correlation structure.

$$\begin{aligned}\gamma(h) &= C(0) - C(h) \\ &= \theta_0 + \theta_1 - (w_{s_i,t_j}\theta_1 exp(-\frac{h}{\theta_2})) \\ &= \theta_0 + \theta_1(1 - w_{s_i,t_j} exp(-\frac{h}{\theta_2})) \end{aligned} \tag{2.6}$$

$$\begin{aligned} C(h) &= w_{s_i,t_j}\theta_1 exp(-\frac{h}{\theta_2}) \\ C(h)/w_{s_i,t_j} &= \theta_1 exp(-\frac{h}{\theta_2}) \end{aligned} \tag{2.7}$$

The standard way to form the covariogram, which can be used for the Euclidean distance model, is to plot the covariances against the distances (or 'lags') between each pair of stations. The covariance between stations s and t at time i is given by $C(s_i, t_i) = \sum_{i=1}^{N} \frac{(Z(s_i)-Z(\bar{s}_i))(Z(t_i)-Z(\bar{t}_i))}{N}$ (Cressie, 1991), where $Z(s_i)$ and $Z(t_i)$ are the values of the variable at stations s and t at that point in time, and N is the number of points of time where there is available data. The resulting plot is known as the covariogram cloud. The cloud of points is then 'binned' by averaging the covariances regularly over the lags to obtain the binned covariogram. The binning is usually performed by taking the mean values over regularly spaced distance ranges. However, for the Tweed data there were several occasions where a very small number of very large, negative covariances tended to drag the binned values down below zero. Therefore there were several instances where binning was performed by taking the median instead of the mean value at each of the ranges of distances.

When it comes to predicting at unsampled locations on the Tweed, both Euclidean and tail-up stream distance based covariance models will be used for

the covariance structure, separately, in order to see which gives the more accurate predictions. Prediction will only be performed using the yearly averages due to the lack of regular data over the network. It would be preferable to have a different covariance structure for each year to allow for changes over time, but there was far too much variability in the resulting covariograms, especially those using stream distance. This is understandable, since using the stream distance model reduces the number of pairs of stations available from 3403 to just 631, since the rest of those pairs are not flow connected. Therefore, just one set of covariance parameters was found for each of the models over the entire time period, using the overall average nitrate levels at each station.

Euclidean Distance Detrending

The Euclidean distance based covariogram will be calculated using the residuals from Euclidean distance based detrending in order to show how the estimation process works. The covariogram cloud is shown in Figure 2.7, while the mean and median binned covariograms are shown in Figure 2.8. The first two bins at non-zero lag are negative, which could suggest that the Euclidean distance based covariance structure is effectively non-existent for anything other than a lag of zero (or very close to it). However, the decision was made to base the covariance parameter estimation on the plot using median binning, as it seemed to follow a more well defined shape.

The binned version of the covariogram can now be used to estimate the underlying autocovariance structure of the data. The most commonly used method of obtaining estimates for the model parameters is to use weighted least squares techniques to fit a model to the binned values. There are various different weighting techniques that could be chosen for this purpose, the most basic of these being

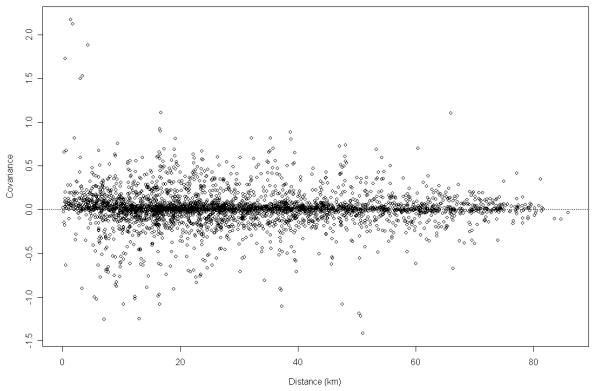
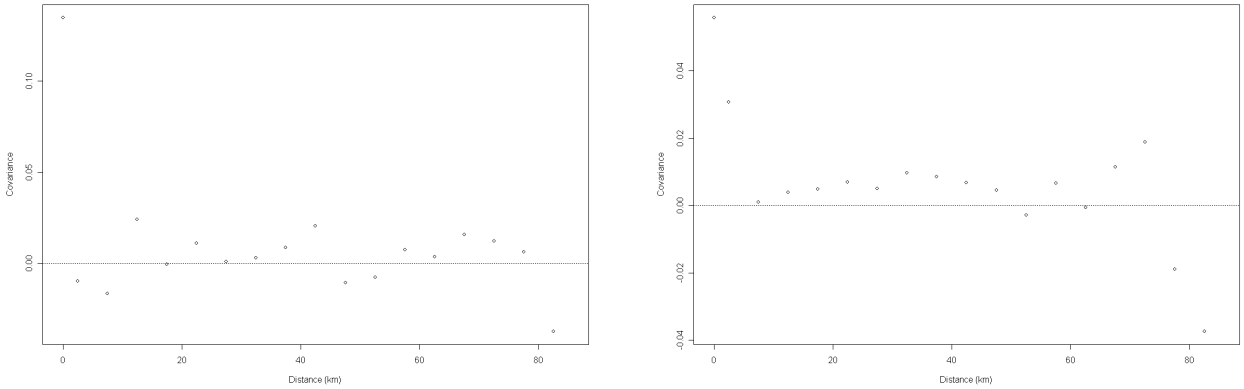


Figure 2.7. Covariogram cloud based on Euclidean distance



(a) Mean binning (b) Median binning

Figure 2.8. Binned Euclidean distance based covariograms, Euclidean distance detrending

to weight by the number of values that have been binned at each point. However, this approach tends to give less weight to the binned values at shorter lags than would be desirable, considering these are generally quite important given how quickly the covariance can tend to zero. Cressie (1985) proposes a weighting scheme that gives more weight to the shorter lags by weighting the bin b_j using $n(b_j)/\gamma(b_j)$ where $n(b_j)$ is the number of covariances binned for b_j and $\gamma(b_j)$ is the value of the variogram that the model predicts at this bin. Although the

modeling of the Tweed data has been carried out using covariances, the equivalence between variogram and covariogram, stated in equation (1.7), allows these weights to be used with the covariances too. This weighting method seems like a sensible way to proceed with fitting, as it gives more weight to lags where more points have been "binned". What makes this better than just a simple weighted fit is the inclusion of the variogram value, $\gamma(b_j)$, which gives added importance to the fit at the lowest lags. This is beneficial here as, given the nature of spatial data, there are sometimes more binned values at the "medium distance" (in this case around 20km) lags. This means that a standard weight would potentially fit quite poorly at the first few lags.

The shape of the exponential model fits the observed covariogram well and so it has been used both to demonstrate the covariogram fitting process and for some analysis of the Tweed data, though a study into other possible models is carried out in Section 3.2.2. Using the exponential model, as shown in (1.3), parameters were estimated using weighted least squares. The estimated parameters for the median binned covariogram were (the nugget) $\theta_0 = 0.010$, (the partial sill) $\theta_1 = 0.046$ and (the range parameter) $\theta_2 = 6.23$, and a plot of how this model fits the covariogram is shown in Figure 2.9. The very low range parameter indicates that the covariance between stations drops off considerably after around six kilometers, and looking at the plot of the model it seems that after just over fifteen kilometers the covariance is practically zero. This is probably a result of the detrending process having worked well enough that the trend explains most of the large scale variation in the system, and so that the random variation at one station will have only small scale influence on the random variation at another. The partial sill is also quite low and this is likely to be due both to the detrending process and the relative lack of variation in the values at the stations. Recall that it is yearly averages that are being used here and these are obviously going to be far less variable than if, for example, monthly averages had been used.

Consequently, when $(Z(s_i) - \bar{Z})$ is calculated for each station at each time point the difference between the average at year i, $Z(s_i)$, and the overall average $\bar{Z}$ reflects only random variation.

One of the problems with estimating the covariance structure of the data with a covariogram rather than the standard (and preferable, Cressie 1991) variogram is that the nugget is more difficult to estimate. Comparing the exponential form of the variogram (1.8) to the exponential form of the covariogram (1.11), it can be seen that the nugget forms part of the semivariance for all lags of the variogram, while it only features in the covariogram for lag zero. This means that in the least squares fitting of the variogram, the estimated nugget has an influence on all binned values, whereas it only has an influence at lag zero in the covariogram. This means that the 'best fitting' nugget in the covariogram tends to be the difference between the binned value at lag zero and the partial sill, meaning that lag zero is fitted perfectly. The only exception to this is if (as can be seen in Figure 2.12) the binned value at lag zero is less than that at lag one, in which case the nugget will tend to zero. This is a significant drawback of using a covariogram rather than a variogram.

This process of estimating the covariogram model parameters can be carried out in a very similar manner when stream distance is used instead of Euclidean distance. The observed covariances are calculated in the same way as when Euclidean distance was being used but this time the lags are calculated by using stream distance. However, the form of the stream distance based covariance model (1.10) suggests that the observed covariances are $\Pi_{k \in B_{s_i,t_j}} \sqrt{\omega_k} C_1(h_{riv})$ whereas in the Euclidean based models we were observing just $C_1(h_{euc})$. This means that to estimate the parameters of $C_1(h_{riv})$ we must divide the observed covariances by the weights $\Pi_{k \in B_{s_i,t_j}} \sqrt{\omega_k}$ associated with them. Finally, all pairs of stations that are not flow connected are excluded from the covariogram as these

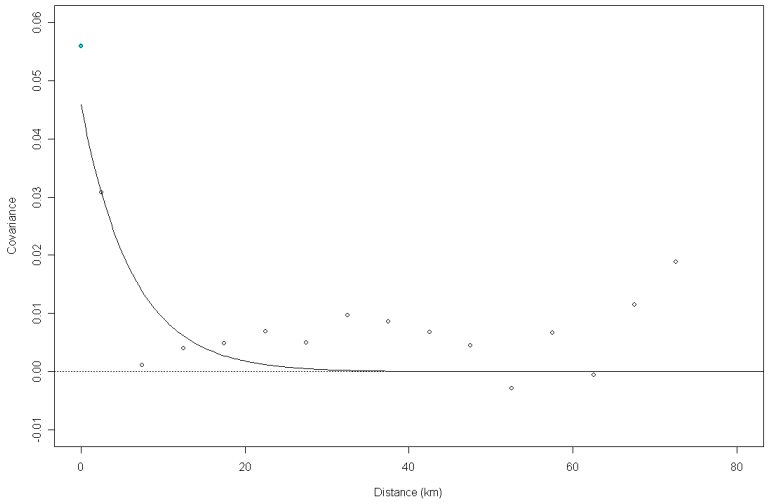


Figure 2.9. Fitted exponential model to Euclidean distance based covariogram

values are automatically set to zero under the model. The resulting covariogram cloud is shown in Figure 2.10.

Dividing the observed covariances by their weights introduces a further complication that can be seen in Figure 2.10. The covariances when using Euclidean distance, as shown in Figure 2.7, all lie roughly between -1 and 2. However, the stream distance based covariogram shows that the observed values lie in the range -60 to 10. The huge disparity is not necessarily an indication that the covariance is stronger when using this distance metric, rather it is indicative of several situations where the weighting between two stations is very small (due to a large distance between them and/or one of the stations lying on a relatively minor stretch of river) but the covariance is reasonably large. This appears to happen much more frequently when the covariance is negative, leading to a number of large negative covariances. The main problem with this arises when mean binning is used for the covariogram cloud. Figure 2.11(a) shows that the mean value in almost all of the bins is quite substantially (in comparison to the range of values in the Euclidean based covariogram cloud) below zero, and since valid

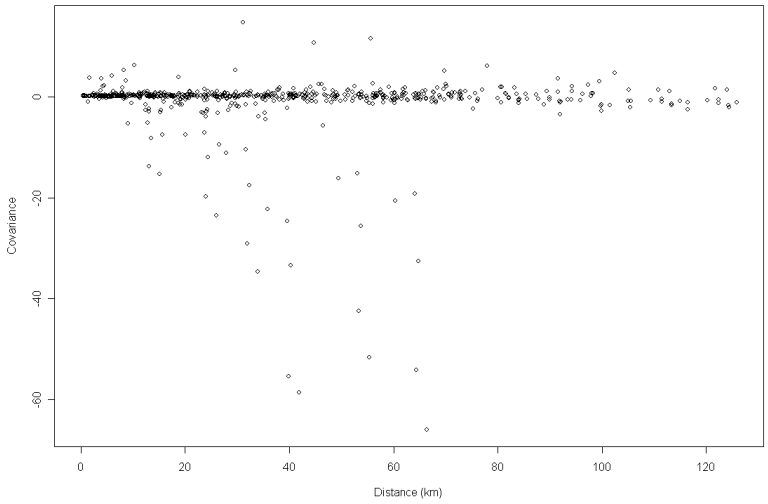
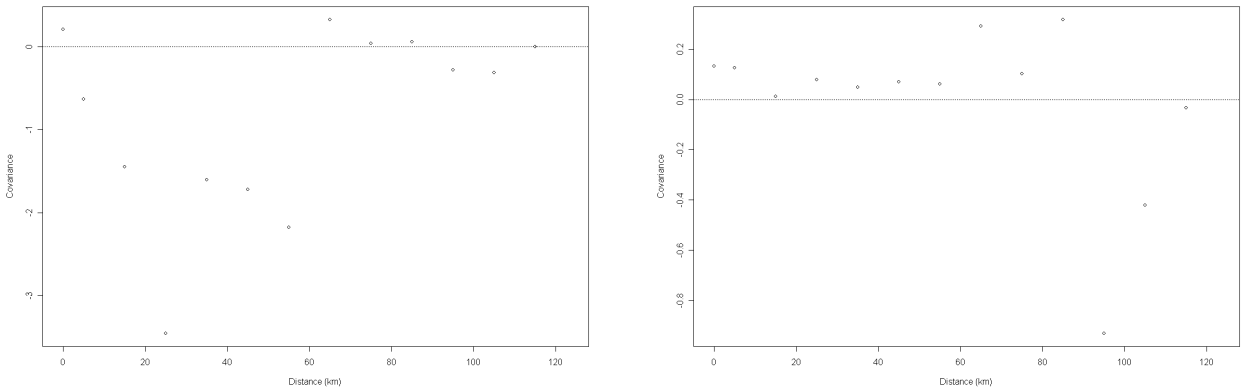


Figure 2.10. Covariogram cloud based on stream distance and Euclidean distance based detrending

covariogram models will not allow for a negative covariance this plot cannot be used to estimate a covariogram. The problem seems to stem from the few very low covariances pulling down the mean values.

Figure 2.11(b) shows the covariogram that has been binned using the median values. While the shape of the median binned plot is still not quite as smooth as would have been hoped, it is still much more regular than when mean binning was used. A model, shown in Figure 2.12, was fitted to the median binned covariogram using weighted least squares. This model takes the form detailed in equation (1.10) and has parameters $\theta_0 = 0.0001$ and $\theta_1 = 0.150$ and $\theta_2 = 32.6$. It is worth noting that the bins above sixty kilometers are formed with a maximum of twenty pairs of stations each (and usually much less than that). Fitting this covariogram is problematic, as using weighted least squares would fit much closer to the middle bins (between 20 and 60km), which are formed from around one hundred pairs each. This presents a problem, as it seems likely that the covariance will be tending to zero at these lags anyway and that any nonzero lags may just



(a) Mean binning (b) Median binning

Figure 2.11. Binned stream distance based covariograms, Euclidean distance detrending

be a result of random noise. In Figure 2.12, bins above 80km have been excluded as they are much less than zero and are made from only a small number of pairs of points each. It can be seen that the level of the fitted model seems to be raised, in order to better capture the four bins between 20 and 60km, reflecting the fact that there are more pairs of points at these distances apart. The fitted model does not fit as well as that for the Euclidean covariance. This is not surprising given the extra noise in the stream distance plot that is caused by far fewer pairs of locations (as locations must now be flow-connected) and adjusting for by the flow based weight.

Comparing the stream distance covariogram parameters to the Euclidean we can see that the partial sills are quite similar, with those for the stream distance being slightly higher, while the range parameter of 32.6km is far higher than in the Euclidean case (which was 6.23km). The partial sill is also much larger in the stream distance case (0.15) than the Euclidean (0.046). The increase in parameter values are not very surprising given that several pairs of (non-flow connected) stations have been excluded from this covariogram, and it is to be expected that these would have lower covariances especially at larger lags. However, such a huge

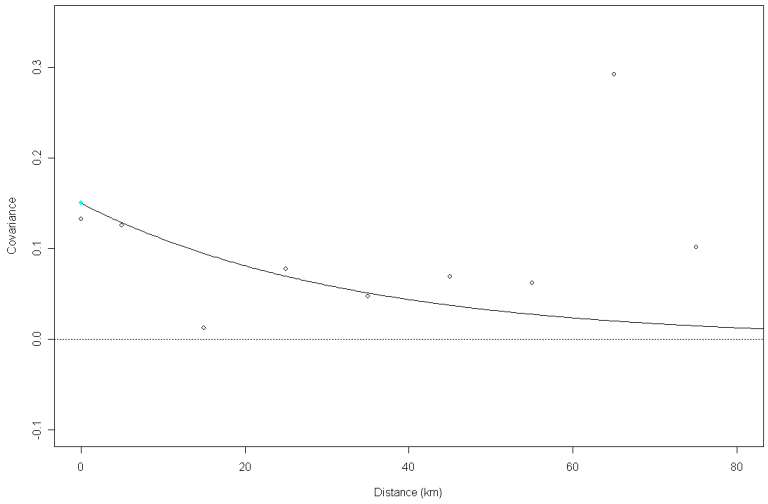


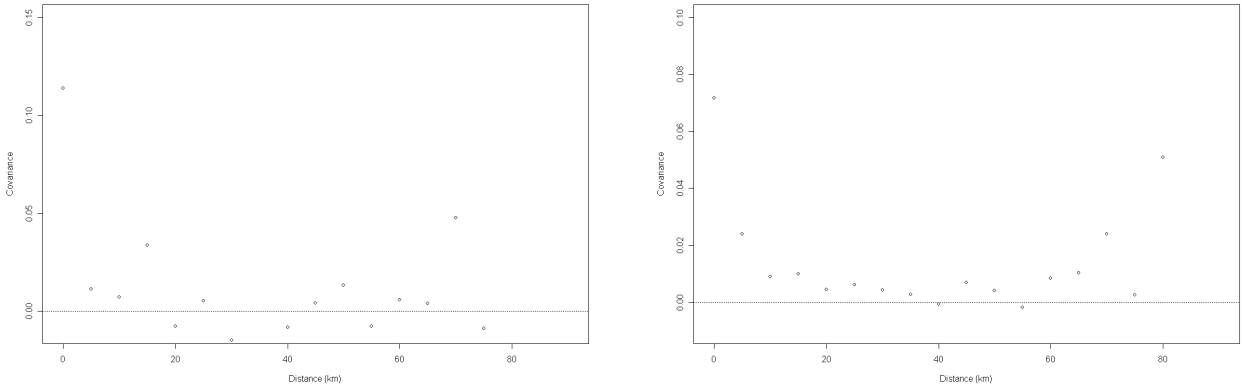
Figure 2.12. Fitted stream distance based covariogram model

jump in range parameter is slightly surprising and might highlight the difference in the success of the detrending process when used with different distance metrics. When the Euclidean detrending method was used in conjunction with the Euclidean distance based covariogram the range parameter was very low. This implied that the influence of one station on another was only over a very small Euclidean distance. Now, when used with the tail-up stream distance based covariogram, there appears to be more residual correlation structure left in the data as the covariance structure now has a much higher range. This means that a station has an influence over a much larger stream distance based range. What may be happening is that the Euclidean detrending is removing a Euclidean distance based process from the data leaving very little Euclidean based covariance, but perhaps leaving a more substantial residual covariance based on stream distance, reflected by the larger range parameter. Of course, stream distances tend to be larger than Euclidean by their very nature but the difference in magnitudes observed here suggests something more that this is occurring. This is an example

of the Euclidean and stream distance based processes present in the data seeming to balance one another out in the fitted trend and covariance models. This feature will be seen several times in Chapter 3. It would also have been possible to fit the covariogram using maximum likelihood, but this method was not for the Tweed data. Future work could compare the results from the least squares fit to a maximum likelihood fit to assess the impact of the covariogram estimation procedure on the outcome.

Stream Distance Detrending

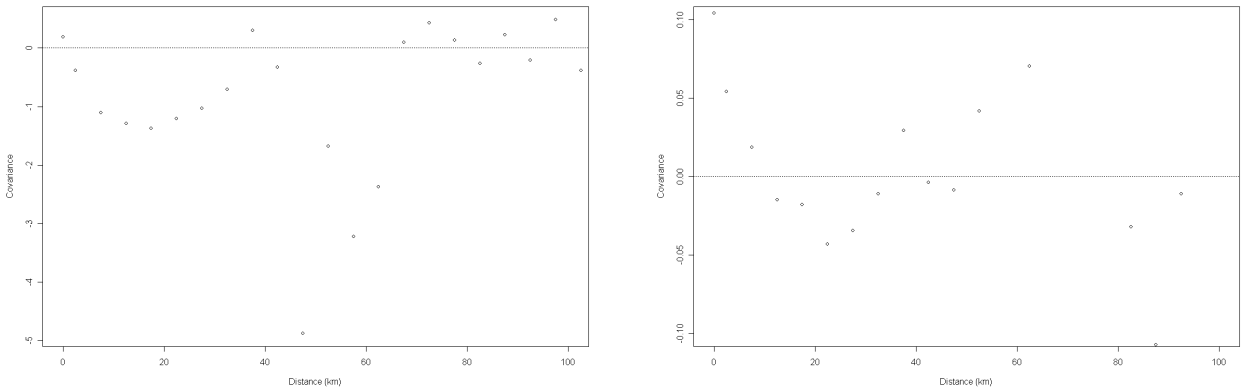
Having estimated covariance parameters using the residuals from the Euclidean distance detrending, covariance parameters will now be estimated for both the Euclidean and stream distance based covariance structures after detrending using stream distance. Figure 2.13 shows the mean and median binned covariograms using Euclidean distance based covariances, while figure 2.14 shows the mean and median binned covariograms using stream distance.



(a) Mean binning (b) Median binning

Figure 2.13. Binned Euclidean distance based covariograms, stream distance based detrending

The Euclidean distance based plots are quite similar for both mean and median binning, as was the case for the Euclidean distance detrended example. It

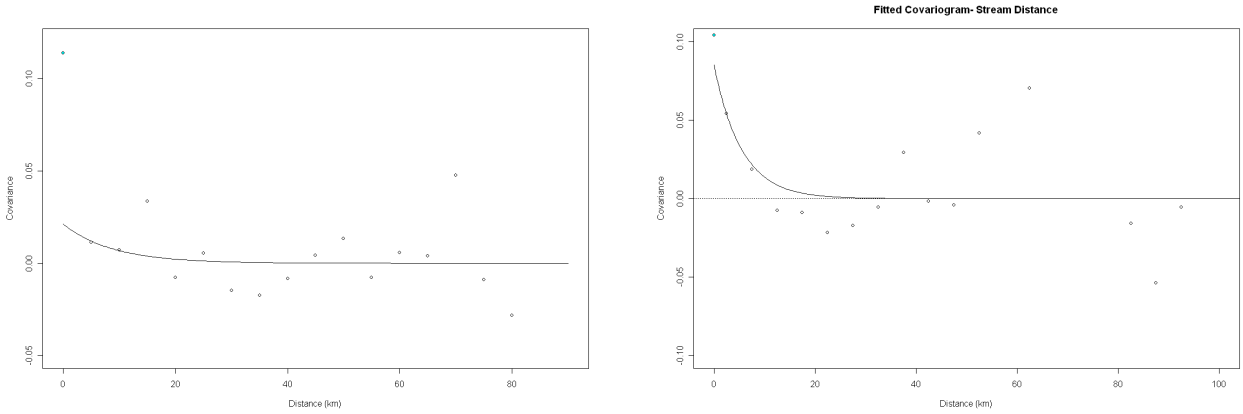


(a) Mean binning (b) Median binning

Figure 2.14. Binned stream distance based covariograms, stream distance based detrending

was felt that the 'default' binning method should always use means as opposed to medians. This was more tricky in the Euclidean detrended, Euclidean distance based covariogram, as the second and third bins were both negative. This meant that the estimated covariance structure tended to zero very quickly. However the mean binned covariogram was used for the stream distance detrended, Euclidean distance based covariogram as none of the binned values were negative until lags of 20km and above. There is obviously the potential for bias and error in this procedure and that will be discussed more in Section 2.2.2. For the stream distance covariogram the decision was much more clear cut as only the median binning left a sufficient shape from which to estimate a covariogram.

Figure 2.15 shows the covariogram models fitted to the binned covariograms. The exponential model was used for both. For the Euclidean distance covariogram, the estimated parameters were $\theta_0 = 0.093$ and $\theta_1 = 0.021$ and $\theta_2 = 8.9$, while for the stream distance covariogram they were $\theta_0 = 0.019$ and $\theta_1 = 0.085$ and $\theta_2 = 5.5$. The Euclidean parameters seem to reflect the fact that there is a large jump between the covariances at lag 0 and the first bin, with a very high nugget effect relative to the partial sill. All of the estimated covariance



(a) Euclidean distance based covariogram (b) Stream distance based covariogram

Figure 2.15. Fitted covariograms, stream distance detrending

Table 2.1. Estimated Covariance Parameters for River Tweed data

Detrending Method	Euclidean Nugget	Euclidean Partial Sill	Euclidean Range
Euclidean	0.010	0.046	6.2
Stream	0.093	0.021	8.9
None	0.045	0.759	5.6
'Connected' Grand Mean	0.035	0.082	13.8
Detrending Method	**Stream Nugget**	**Stream Partial Sill**	**Stream Range**
Euclidean	0.0001	0.150	32.6
Stream	0.019	0.085	5.5
None	0.659	0.422	1000
'Connected' Grand Mean	0.047	0.084	7.3

parameters are summarised in Table 2.1.

Comparing the two Euclidean detrended and the two stream distance detrended sets of covariance parameters, all seem to be roughly the same magnitude, but it is interesting to note that the sill (i.e. nugget + partial sill) is smaller than might have been expected. This is almost certainly due to the detrending process having removed some of the structure of the data beforehand. Indeed, the covariogram fitting process was repeated without having detrended first and the estimated parameters (also shown in Table 2.1) demonstrate the impact that

detrending has had. When the data were not detrended, the estimated partial sill for the Euclidean covariogram was 0.759, compared with 0.046 and 0.021 when they were detrended, and the estimated partial sill for the stream distance covariogram was 0.422, compared with 0.150 and 0.085 when it was detrended. Not having detrended has also led the range parameter for the stream distance covariance structure to be very large. This parameter was tending to infinity and 1000km was the largest value allowed by the weighted least squares fitting process (though increasing this value has very little effect on the sum of squares at higher values). As this parameter tends to infinity the curvature in the fitted model decreases to the point that the curve is almost straight. This might suggest that a linear with sill model may provide a more natural fit for the covariance. However, the use of exponential models with large range parameters is consistent with the literature. Ver Hoef et al. (2006), for example, estimate a range of 2217km for their data.

It is also interesting to compare the other estimated covariance parameters to those estimated in the literature. Garreta et al. (2009) fits a mixture of exponential tail-up and tail-down models to nitrate data. The tail-up component had estimated parameters $\theta_0 = 0.014$ and $\theta_1 = 0.085$ and $\theta_2 = \infty$. The nugget and sill parameters here are quite similar to those estimated for the detrended Tweed data, and it is worth noting that these are estimated by maximum likelihood rather than least squares. The river network used in this analysis was "more than 5000km long" and so significantly larger than the Tweed. This may have had an impact on the range parameter. Ver Hoef et al. (2006) concentrate solely on the tail-up model with an exponential structure and estimate $\theta_0 = 0.387$ and $\theta_1 = 3.55$ and $\theta_2 = 2217$. These estimates were based on SO_4 data from a stream network in Maryland USA, and so are perhaps not as easily compared to the results from nitrate data on the Tweed. The nugget and partial sill are much higher than those estimated for the Tweed data, even when no detrending had

been performed beforehand. The river network used in this example is smaller than the Tweed, and it is interesting to see such a high range parameter estimated for such a small network. Cressie et al. (2006) concentrate solely on the Euclidean covariance structure, spherical model, and estimate $\theta_0 = 0.881$ and $\theta_1 = 0.162$ and $\theta_2 = 6.07$. The river network used in this analysis is also smaller than the Tweed. In this paper, the data used were change in dissolved oxygen and so again may not be directly comparable to the Tweed data. In spite of this, the sill and range are roughly the same as those estimated for the Euclidean covariance structure with no detrending. In general, it is difficult to make direct comparisons with other estimated covariance parameters in the literature. However, it seems like those estimated for the Tweed data are reasonably consistent with those estimated for other datasets.

2.2.2 Criticism of the Covariogram Estimation Procedure

The process involved in estimating the covariance structure is open to quite a lot of criticism. The drawbacks of using covariograms in place of variograms have already been discussed, but it has also been shown that it is necessary to use covariograms when dealing with the weightings introduced by the tail-up covariance model. Peterson et al. (2006) state that "the fitted values of the autocorrelation parameters are dependent on the bin size selected. Thus, parameter estimates can vary from investigator to investigator as a function of bin size and this is very much the case for the parameters estimated for this analysis. The number of bins (and thus size of bins) was altered for each of the covariograms in order to produce a more regular, estimable shape. It is hoped that the estimated parameters from this section would not vary much if another investigator were to repeat the analysis. Small differences should not affect the outcome of kriging by very much, however the subjectivity of the parameter estimation process is a

form of bias.

Peterson et al. (2006) suggest log likelihood based methods as an alternative to weighted least squares covariogram fitting. Likelihood based methods will give more robust estimates of the covariance parameters when the errors follow a normal distribution (Jobson and Fuller, 1980), however if the assumption of normality does not hold the parameter estimates can be poor (Carroll and Ruppert, 1982). The assumption of normality for the covariances in the Tweed data does not seem to hold, at least for the stream distance based covariograms. Looking at Figures 2.11 and 2.14, the difference between using the mean and the median highlights the fact that symmetry does not hold for the stream distance based covariograms. The weighting in the tail-up structure appears to skew the lower tail of the distribution.

This demonstrates the problem of which method of estimation to use for the covariogram parameters. In the literature, for example, Ver Hoef et al. (2006) and Garreta et al. (2009) use a restricted maximum likelihood based approach to estimation, while Cressie et al. (2006) use variograms, as the use of a stream distance based component was ruled out using likelihood based methods. Gardner and McGlynn (2009) use variograms to estimate the tail-up model parameters, but do not adjust for the flow based weights. Ver Hoef et al. (2006) show a variogram with both stream and Euclidean distance based semivariances although state that "[as] there is no weighting for flow volume in the variogram, so it would not be appropriate to use this for estimating covariance parameters, as is often done in classical geostatistics". This means that either a covariogram, adjusted for stream weights as demonstrated in this chapter, or likelihood based methods must be used. Likelihood based methods were ultimately rejected for use with the Tweed data, but it would be interesting to see if this approach would produce significantly different results.

2.3 Conclusions on Estimating Trend and Covariance Structures

This chapter has discussed how trend surfaces and covariance models and their parameters were estimated for the River Tweed dataset.

Two trends were estimated for the nitrates data. The first was a standard nonparametric trend surface based on Euclidean distances. The second was a novel stream distance based approach that used the tail-up model as a starting point and adapted it to estimate a reasonably general trend by removing the flow weighting. Both trends were then used to detrend the data by subtracting the estimated trend from the observed data. This is the first time that either nonparametric trends or trends based on stream distance have been considered in the river network modelling literature.

There was then discussion as to which covariance models, or mixture of covariance models, would be most appropriate for use with nitrates data. It was decided that Euclidean and stream distance covariance models, along with a mixture of these, would be most appropriate. From here, the next stage was to estimate covariance parameters for each of these two models. This was performed for each detrending method, as there is no indication at this stage which method will be preferable for constructing the trend. Covariance parameters were estimated from the observed (binned) variogram using weighted least squares. This method has been employed in some previous literature in the area of stream distance based prediction but some others choose likelihood based methods. The estimated parameters for the Tweed data are quite similar to those predicted in some of the literature, but there is generally little agreement across different studies in terms of the estimated covariance parameters as a result of the different outcomes that are used in each.

Chapter 3

Assessing the Tweed Spatial Predictions

In this chapter, the different covariance and detrending structures that have been discussed will be used in an assessment on the River Tweed data. The predictive power of the models will be examined, separately for Euclidean and stream distance based detrending methods, by comparing their prediction accuracy using studies based on cross-validation. Four studies will be carried out in order to assess prediction accuracy and find out more about the facets of the trend models and covariance structures.

Study 1 This study takes the covariance parameters estimated in the previous chapter and tries to determine which mixture of the two covariance structures will minimise the root mean squared error.

Study 2 This study investigates the role that the covariance parameters take in determining the predictions. In particular, it looks at how an "optimum" set of covariance parameters could be produced by using those that minimise the RMSPE. This study will also examine the effect that choice of covariance

model has on the accuracy of predictions.

Study 3 This study focuses on how the fitted trend affects the predictions, and performs a sensitivity study to ascertain how much the predictive accuracy is altered by changing the bandwidth of the trend

Study 4 The final study simulates data using the River Tweed geography in order to assess the impact that different sampling schemes would have on the prediction errors.

3.1 Study 1– Assessing the Mixture Model for Covariance

The aim of this study is to find the best mixture of covariance models, when using the covariance parameters estimated in the previous chapter. It is vitally important to do so as subsequent work will require the use of a mixture of the two covariance models in order to predict at unsampled locations on the river. It is important to stress that this study aims to find the best mixture for this set of covariance parameters only, study two will focus on optimising the covariance parameters and mixture of them.

For each of the twenty-one years in turn, a subset constituting ten percent of the available stations for that year was excluded from the dataset and the remaining stations used to predict the nitrate values of these stations. This process was repeated until one hundred predictions had been made for each year of data. The covariance model used was a mixture model with Euclidean and tail-up stream distance based components (1.13), each using an exponential covariance model. The covariance parameters used with each covariance structure are those estimated in section 2.2.1, with a range of λ 'mixing parameters' (ranging from 0

to 1, in increments of 0.05). To ensure parity in the comparisons, the same subset of stations is dropped out for each mixing parameter. This will be repeated for both Euclidean and stream distance based detrending, as it is unclear what effect the detrending method will have on the mixing parameter.

As was discussed in the previous section, only one set of covariance parameters was used for all years of data. This seemed reasonable as there seemed to be little change in the yearly cycles and overall trend in any of the stations on the Tweed over time. Attempts were made to estimate different parameters for each year of data, however many of the covariograms proved too noisy to properly estimate parameters. It was felt that one set of covariance parameters, estimated using more available data, would be better than twenty one very poorly estimated covariograms.

3.1.1 Euclidean Distance Detrending

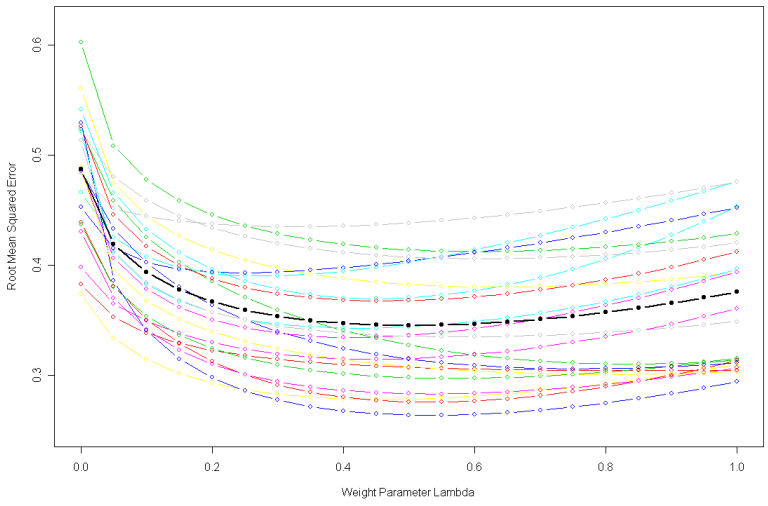


Figure 3.1. Estimated Prediction error by lambda for Euclidean detrended mixture model. The coloured lines denote different years of data, while the black line is the overall average

Firstly, the Euclidean distance based detrending method will be assessed using kriging to predict the values at the stations excluded in the cross-validation procedure. Figure 3.1 shows the root mean squared prediction error (RMSPE) plotted against the mixing parameter, λ, with each coloured line corresponding to a different year's data and the black line denoting the overall average for all 21 years of data.

Looking at the overall average in Figure 3.1, the optimum mixing parameter lambda, in terms of minimising the RMSPE, is somewhere around 0.5, meaning that the covariance structure is split 50/50 between the Euclidean and stream distance models. At $\lambda = 0.5$ there is an estimated error of 0.345 but, while 0.5 does give the lowest RMSPE, the relationship between the λ and the RMSPE is very flat for λ in the range 0.35 to 0.65. It is also worth noting that for λ equal to zero and one the covariance structure being fitted is the Euclidean and stream distance based models (respectively) on their own. Therefore, just using the Euclidean based covariance structure and the Euclidean based detrending method results in a RMSPE of 0.487, while just using the tail-up stream distance based covariance structure with the Euclidean detrending results in a RMSPE of 0.376. Consequently the 'best' mixture model is increasing the accuracy by 0.031, just less than 10%, over the stream distance model on its own. When compared to the RMSPE at the best mixture, the Euclidean covariance structure on its own ($\lambda = 0$) is significantly poorer than the stream distance covariance structure on its own ($\lambda = 1$), and this may be because Euclidean distance based detrending was used for this part of the study. This means that quite a lot of the Euclidean structure of the data may have already been accounted for in the trend and so there is not quite as much left to account for in the covariance structure.

Figure 3.2 will now be used in order to assess the bias of the predictions. This will allow us to see whether the predicted values tend to over or underestimate

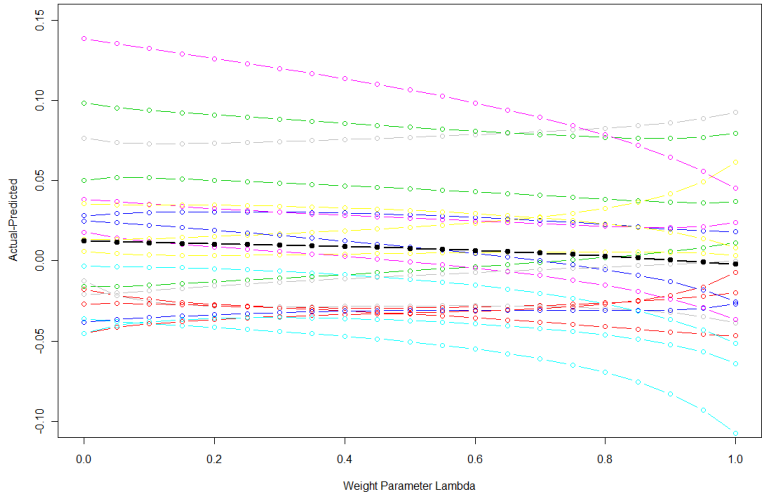


Figure 3.2. Assessing Bias in the Euclidean distance detrended results by looking at difference between actual value and predicted value

the true value, and see how this changes with the mixture of the covariance models. The figure shows that, on average, the predicted values tend to slightly overestimate the actual nitrate values. The highest bias of 0.013 is found at $\lambda = 0$, and it decreases steadily, with the lowest bias of 0.001 being found at $\lambda = 0.8$ and then continuing to decrease to -0.002 at $\lambda = 1$. It is worth noting that when looking at the bias in individual years, ten years fall below zero and eleven above. While this indicates an even spread above and below, there are a few years which have been relatively badly overestimated, and these drag the overall average slightly above zero. At the "best" mixture of 0.5, the bias is just 0.008. This indicates that on average, the mixture tends to slightly overestimate, but not by too much, but certain mixtures underestimate slightly.

3.1.2 Stream Distance Detrending

The stream distance based detrending was then assessed using the same cross-validation technique used for the Euclidean distance based detrending. Figure 3.3 shows the results from this, using the same range of λ mixing parameters that were used before and dropping out the same subsets of stations as before.

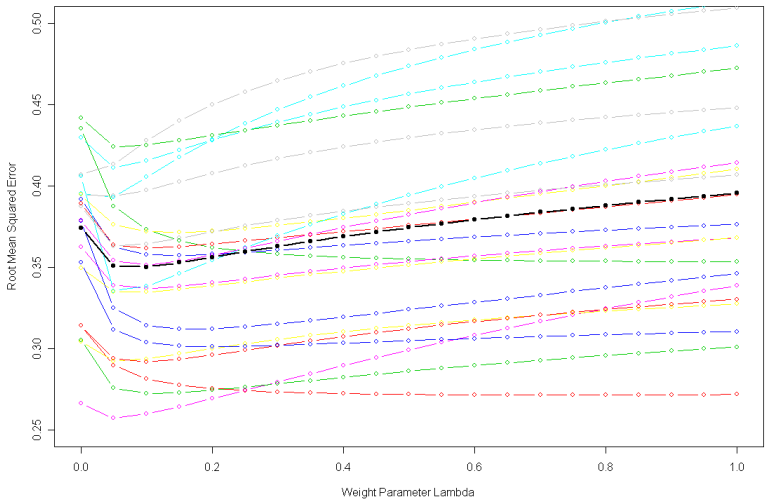


Figure 3.3. Estimated prediction error by lambda for stream distance detrended mixture model

Looking at Figure 3.3, there is not a huge difference in the RMSPE as λ changes. The lowest RMSPE observed was 0.350, which occurs at $\lambda = 0.10$, although it should be noted that λ values 0.05 to 0.25 all have RMSPEs within 0.01 of this value. This means that the 'optimum' model places 10% of the weight on the tail-up stream distance model and 90% on the Euclidean, meaning that using stream distance detrending has caused the optimum mixture to swing heavily in favour of the Euclidean model. Using just the Euclidean or tail-up stream distance covariance structures result in RMSPEs of 0.374 and 0.395 respectively. The magnitude of the results are quite surprising when viewed next to those from the Euclidean distance based detrending. In those results the optimum λ

produced an equal split between the two covariance structures and, possibly as a result of using Euclidean distance to detrend the data, the purely Euclidean distance based covariance structure produced a much poorer RMSPE than even the stream distance structure on its own (let alone the best mixtures). This might have suggested that as the stream distance detrending method is being used here, the stream distance based covariance structure on its own would be significantly worse than the best mixture or the Euclidean covariance structure on its own too. While still the worst possible covariance structure to use, the pure stream distance model is around 12% higher than the RMSPE of the best mixture, as opposed to the Euclidean structure being just over 40% higher in the previous example.

The best mixture model in the Euclidean detrended example has a slightly lower RMSPE than the best mixture in the stream distance detrended example (0.345 as opposed to 0.350), but on the other hand there is much less variation in the RMSPE values across the different λ values and across the different years. This means that it is unclear whether one of the detrending methods is 'better', at least in terms of which will lead to the lower RMSPE. One possible reason for having observed such similar results for both detrending methods may be that fitting a mixture covariance model 'fills in the gaps' left by the detrending process. In other words, the detrending method seems to have little impact on the lowest RMSPEs found because the mixture at which it is found will alter to account for the detrending distance metric used. So the best mixture using stream distance detrending will tend towards the Euclidean covariance structure and vice versa.

Figure 3.4 shows the bias in the stream distance detrended results. There is much similarity between these results and those for the Euclidean distance detrending (Figure 3.2), as the average bias for all years is highest at $\lambda = 0$

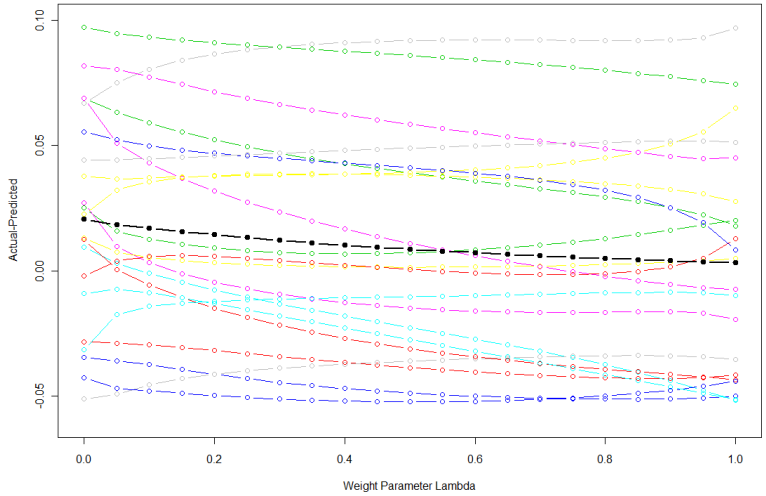


Figure 3.4. Assessing Bias in the stream distance distance detrended results by looking at difference between actual value and predicted value

(0.021) before falling steadily and having minimum value at $\lambda = 1$ (0.003). As with the Euclidean results this implies that, on average, the predicted values tend to slightly overestimate the nitrate levels. This suggests that neither detrending method is necessarily better than the other in terms of systematic bias. However, in the context of predicted values which (on the log scale) mostly fall between ± 4, overestimating on average by less than 0.03 does not seem too detrimental.

3.1.3 No Detrending

In order to further investigate the benefits of removing a trend prior to analysis, the cross-validation procedure will be used on the original data (with no trend having been removed). The results from this are shown in Figure 3.5. The covariance parameters used are shown in Table 2.1 with those estimated for the detrended data in Chapter 2. The parameters were estimated in the same way as is described in that Chapter.

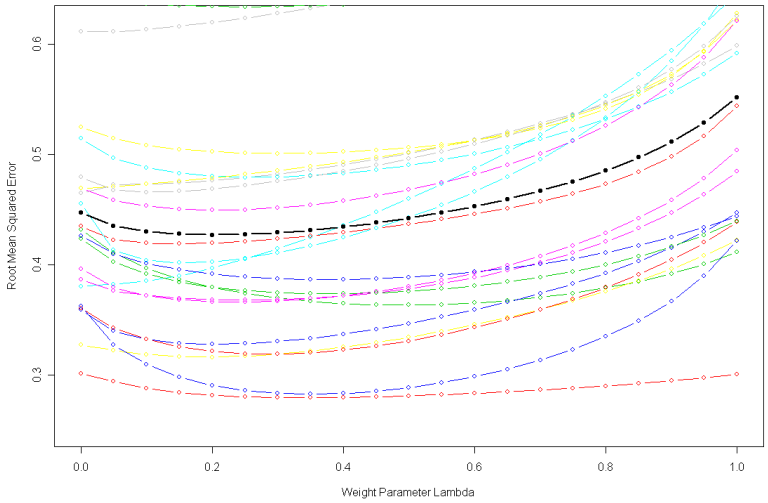


Figure 3.5. Estimated prediction error by lambda for original data

Figure 3.5 shows that when no detrending was carried out beforehand, the lowest RMSPE of 0.494 was found at $\lambda = 0.40$, while the pure Euclidean or stream distance based covariance structures give RMSPEs of 0.580 and 0.666 respectively. These results suggest that not removing a trend has led to a higher RMSPE compared to that obtained when using either detrending method. This was the expected result, as the detrended examples were not designed to re-calculate the trend for each iteration of the cross-validation. This means that, although the values at a subset of stations may be missing from the kriging part of the process, these stations are still used to calculate the overall trend. The detrended predictions are therefore based on both the estimated trend and any neighbouring observations in the same year but the non-detrended predictions are just based on the neighbouring observations and so it is not surprising that the detrended RMSPEs are lower. The only places that this is likely to be an inaccurate representation of the relative errors is in locations that are quite far from any observed values. In these locations the estimated trends would not be anywhere near as accurate as those used for this cross-validation.

This does not imply that the detrending was not necessary. The original motivation for detrending was to better satisfy the assumption of constant mean necessary for kriging rather than to increase the prediction accuracy. Lower root mean squared errors may be a consequence of this, as a trend surface calculated based on observed values is always going to be better than the overall mean fitted when using ordinary kriging on its own, but the validity of the kriging assumptions provide the real motivation for doing so.

3.1.4 A Different Approach to Study 1

The primary focus of Study 1 was the identification of the best mixture of the covariance models in terms of RMSPE, the results beg the question of whether the fixed covariance parameters have affected the outcome.

In order to see whether the conclusions on the use of a mixture model would change if covariance parameters were specifically chosen to suit each mixture of the two structures, a further simulation study was carried out. The simulation study was effectively exactly the same as those conducted earlier in Study 1, with the only difference being that for each λ from 0 to 1 in increments of 0.1 new covariance parameters were estimated simultaneously for the Euclidean and stream distance covariogram models.

Simultaneous estimation of the covariance parameters was not possible using covariograms for reasons similar to those described in Chapter 2. The use of a stream distance metric and the "tail-up" covariance structure would require the covariogram for stream distance to be divided through by the flow based weighting structure. As this is not done for the Euclidean covariograms, this would mean that two entirely different covariograms would be used and it is not possible to simultaneously fit these two entirely different covariograms while still accounting

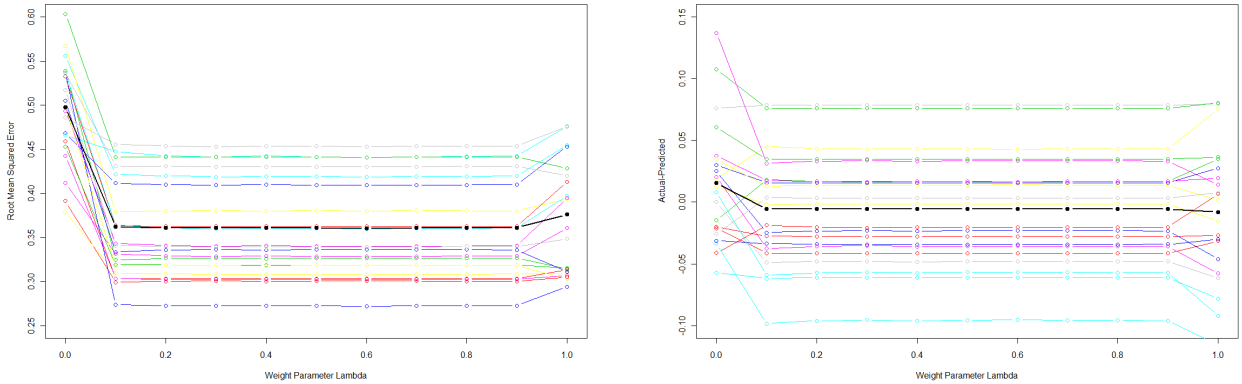
Table 3.1. Estimated covariance parameters, RMSPE and bias when covariance parameters are re-estimated for each λ

	Euclidean			Stream				
λ	Nug	Sill	Range	Nug	Sill	Range	RMSPE	Bias
0.0	0.002	0.031	23.4	-	-	-	0.4974	0.0154
0.1	0.001	0.020	52.2	0.003	0.474	14.1	0.3619	-0.0057
0.2	0.001	0.024	46.4	0.003	0.229	13.9	0.3610	-0.0054
0.3	0.003	0.028	44.7	0.003	0.151	13.9	0.3607	-0.0054
0.4	0.004	0.032	46.4	0.002	0.114	14.0	0.3608	-0.0055
0.5	0.004	0.039	45.1	0.002	0.091	13.9	0.3608	-0.0054
0.6	0.003	0.049	44.7	0.001	0.075	14.0	0.3605	-0.0054
0.7	0.005	0.065	45.1	0.001	0.065	13.9	0.3608	-0.0054
0.8	0.005	0.097	45.6	0.001	0.057	13.9	0.3609	-0.0054
0.9	0.006	0.192	46.4	0.001	0.051	13.9	0.3611	-0.0054
1.0	-	-	-	0.001	0.064	33.2	0.3760	-0.0082

for the portion of the variation accounted for in the other. This is almost certainly the reason that no prior literature has simultaneously estimated the parameters of the two structures using variograms or covariograms. As an alternate solution was required, it was decided to stop short of binning the covariograms and use the data from the covariogram cloud so as to find the least squares estimates of the six covariance parameters (for each λ), as determined by the non-binned data.

The estimated covariance parameters, along with the RMSPE an bias for each set are shown in Table 3.1, while the results are shown graphically in Figure 3.6. Due to the very computationally intensive nature of this study, only the results for the stream distance detrending method are shown.

Before discussion of the results themselves, it is very interesting to discuss the estimated covariance parameters, and how they change as the mixture moves from the pure Euclidean model, to a mixture of the two, to the pure stream distance model. It is interesting to see that the only huge changes in any of the parameters occur at $\lambda = 0$ and $\lambda = 1$, and perhaps at the mixtures above and below those values. For the Euclidean component of the mixture model, the



(a) RMSPE (b) Bias

Figure 3.6. Results of simulation study re-estimating covariance parameters for each mixing parameter

sill seems to increase as λ is increased, while the opposite is true for the stream distance based component. A plausible explanation for this is that as the mixture puts less weight on the covariance structure, its sill increases so that its overall contribution to the covariance structure as a whole remains similar. In contrast to this, the range parameters for both components do not change much at all, except for at $\lambda = 0$ and $\lambda = 1$. This would seem reasonable, as the sill is changing in order to increase the influence of that structure on the mixture, but the range after which there is little covariance between sites is unlikely to change.

Looking at the root mean squared errors in the table, and in Figure 3.6(a), it is clear that by allowing the covariance structure to change as λ changes, there is very little difference in the RMSPE (except for at $\lambda = 0$ and $\lambda = 1$). Again, this is to be expected, as the study is now optimising the covariance structure at each λ, essentially so that it fits exactly the same covariance structure at each. This is also reflected in the bias, which is around -0.0054 for all λ except 0 and 1.

As a different covariogram estimation technique was used in this part of the study to the one used in the earlier parts, there is perhaps limited merit in drawing

comparisons between them. However, it is interesting to note that both the root mean squared error and bias for this study were higher than in the previous parts of the study. This result is almost certainly down to the different covariogram estimation techniques that are being used.

The most interesting comparison that can be made between the two parts of study two comes from looking at the different approaches they take to modelling the covariance structure. The first part of the study takes fixed covariance parameters and is able to reduce the RMSPE by altering the mixing parameter. The second part of the study takes the opposite approach, by fixing the mixing parameter and adjusting the covariance parameters accordingly. The results from both suggest that the changeable parameters will effectively balance out the fixed parameters. The key to the predicted values in kriging, and thus the RMSPE, lies in the modelled covariance structure itself. For this reason, Study 2 will attempt to go deeper into the covariance structure and how it affects the predicted values and prediction error.

It is worth noting that the difference between detrending and not detrending that has been observed has not been quantified in terms of any significance test, and it is important to reiterate that the main motivation for removing a trend is to satisfy the assumptions of kriging. Given the lack of a statistical test to compare the two, it would not be correct to call the difference 'statistically significant', however the results imply that detrending is a significant step if one wishes to satisfy the assumptions of kriging.

3.1.5 Conclusions from Study 1

In general, it is clear that the use of a mix of covariance structures reduces the prediction error no matter which detrending method is used. This is consistent

with the findings of Peterson and Ver Hoef (2010) and Garreta et al. (2009), who also conclude that mixture models can lower root mean squared errors when compared to the individual distance metric covariance models.

The point of this analysis is not necessarily to show that removing a trend is a better approach than not doing so, but to show the differences in optimum mixing parameters between the methods. It seems a plausible explanation of the results is that there exists various processes at work on the river network; some based on the Euclidean distances between the locations; some based on stream distances, flows and connectivity; and potentially some others, possibly based on the tail-down model or on something unrelated to distance, such as rainfall.

The modelling procedure used for this study has the potential to deal with the Euclidean and stream distance related portions of this. When Euclidean distance is used to detrend, the best mixture model is a 50/50 split between the two distance metrics, meaning some of the residual Euclidean distance based process is explained along with the stream distance process left in the data. When stream distance is used to detrend, the best mixture model puts 90% of the weight on the Euclidean distance based covariance. This seems to imply that the trend accounted for quite a large part of the stream distance based process in the data, and so the remaining covariance structure was largely based on Euclidean distance. When no detrending was used, the mixture slightly tended towards the Euclidean model, which had 60% of the weight. Some part of this effect may be due to the estimated covariance parameters as, during the process of carrying out this study, several different sets of covariance parameters were used (not hugely different to the ones used in the example shown here). It was interesting to note that the lowest RMSPE barely changed at all, but the λ at which that RMSPE was found did change, seemingly to account for the differing covariance structures themselves putting more or less weight on each of the distance metrics.

This 'balancing' of the Euclidean and stream distance based elements of the underlying process is something that will be seen again when the Tweed network data is predicted at unsampled locations in Chapter 4.

A secondary study was then carried out in order to assess the effect of the mixing parameter when the covariance parameters were simultaneously estimated each time the mixing parameter was incremented. The 'balancing' of the Euclidean and stream distance based structures could again be seen, as all of the mixing parameters except for $\lambda = 0$ and $\lambda = 1$ produced almost exactly the same root mean squared error and bias. This is an interesting result and leads in to the work carried out in Study 2. Here, the role of the covariance parameters will be investigated in much greater depth and, rather than re-estimate the covariance parameters each time, optimum covariance parameters are chosen based on the minimisation of the RMSPE.

3.2 Study 2–Investigation of the Covariance Models and their Parameters

The second study into the covariance structure and the effect it has on spatial predictions will focus on the role of the covariance models and their estimated parameters and see what effect this has on the covariance mixture model. This will effectively allow the covariance parameters to be chosen by finding those which minimise the root mean squared error Specifically, the exponential, spherical and linear with sill models will be assessed to see if any will significantly lower the root mean squared error.

3.2.1 Investigating the role of covariance parameters on kriging predictions

It is important to accurately estimate all elements of the covariance structure. However, it is not immediately obvious from looking at the equations for covariograms or the kriging equations (1.19) that, once a covariance structure has been settled on, it is only the range parameter that has any influence on the predicted values. The nugget and sill only influence the estimated error, calculated from (1.22).

To see this, we must remember that kriging is a weighted average of the observations, with the weights being a function of the variogram (or covariogram). To illustrate, consider the exponential variogram function $\gamma(h) = c_0 + c_1 exp(-\frac{h}{c_2})$. Now, change the range parameter by having observations at 5 and 10km, and (firstly) fixing $c_0 = 0$, $c_1 = 1$, $c_2 = 10$. This means that $\gamma(5) = exp(-0.5) = 0.6065$ and $\gamma(10) = exp(-1) = 0.3679$ and thus the ratio of semivariances for these observations is very roughly 2:1. If the range parameter is increased to $c_2 = 20$ then $\gamma(5) = exp(-0.25) = 0.7789$ and $\gamma(10) = exp(-0.5) = 0.6065$. The ratio is now much closer to 1:1 than 2:1, so by increasing the range parameter there is more weight on the distant observation relative to the nearer one.

Now we fix the range parameter and vary the partial sill. As before assume $c_0 = 0$, $c_1 = 1$, $c_2 = 10$ giving $\gamma(5) = exp(-0.5) = 0.6065$ and $\gamma(10) = exp(-1) = 0.3679$. Now let $c_1 = 2$ so $\gamma(5) = 2exp(-0.5) = 2 \times 0.6065$ and $\gamma(10) = 2exp(-1) = 2 \times 0.3679$. This means that while the sizes of the semivariances are double what they were before, they are still in exactly the same ratio relative to one another and so therefore the partial sill has no effect on the predicted values in kriging (and only the kriging error is affected by it).

This can be seen theoretically by using the matrix formulation for ordinary

kriging, as shown in (3.1). In this equation, Σ is the matrix defined in (3.3), but this time λ is given by the first n rows of the matrix calculated in (3.1). In this new set of equations, Σ is formulated as shown in (3.3) and Γ is a single column matrix with entries $\gamma(x_1, x_0), ..., \gamma(x_n, x_0), 1$. Assume the matrix Σ can be broken down into components A, B, C and D as shown in (3.2). Component A is the $n \times n$ matrix with (i, j)th entry $\gamma(x_i, x_j)$, while B is then the $1 \times n$ column of 1's, C is the $n \times 1$ row of 1's and D is simply the 1×1 matrix containing 0. The inverse of Σ is therefore given by (3.4) (Boltz, 1923; Banachiewicz, 1937; Bernstein, 2005). As D is equal to zero, this can be simplified as shown in (3.4).

In the kriging equation (3.1), only the first n rows of $\lambda = \Sigma^{-1}b$ are used for the predictions and given (3.4), this is equivalent to $A^{-1}+A^{-1}B(-CA^{-1}B)^{-1}CA^{-1}b^*$ where b^* is the first n rows of b. Given this and the identity $(kA)^{-1} = k^{-1}A^{-1}$ for some scalar k (Horn and Johnson, 1985), and assuming the nugget is kept constant at zero, therefore a sill of k will mean that the weights λ_k in kriging will be given by (3.5). This can be simplified to show that it is equivalent to the weights λ_1 used with sill 1 (3.6). This shows that the weights used for kriging do not change as the sill is changed, if the nugget is set to zero and the range is constant. While this proof is for the case when the nugget is set to zero, simulations will later show that this is the case for a nonzero nugget effect too.

$$\lambda = \Sigma^{-1}b \tag{3.1}$$

$$\Sigma = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{3.2}$$

$$\Sigma = \begin{bmatrix} \gamma(z_1, z_1) & \cdots & \gamma(z_1, z_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(z_n, z_1) & \cdots & \gamma(z_n, z_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} b = \begin{pmatrix} \gamma(z_1, z_*) \\ \vdots \\ \gamma(z_n, z_*) \\ 1 \end{pmatrix} \tag{3.3}$$

$$\begin{aligned} \Sigma^{-1} &= \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA{-}1 & (D - CA^{-1}B)^{-1} \end{bmatrix} \\ &= \begin{bmatrix} A^{-1} + A^{-1}B(-CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(-CA^{-1}B)^{-1} \\ -(-CA^{-1}B)^{-1}CA{-}1 & (-CA^{-1}B)^{-1} \end{bmatrix} \end{aligned} \tag{3.4}$$

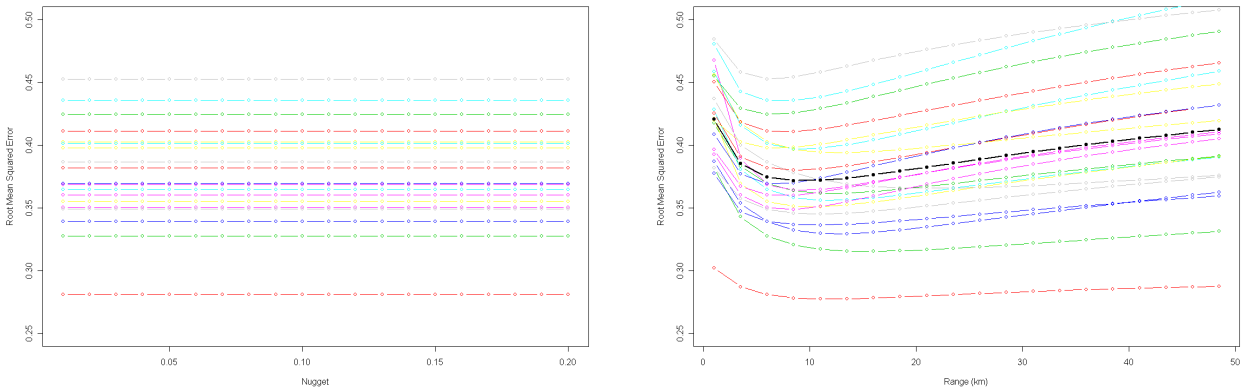
$$\begin{aligned} \lambda_k &= ((kA)^{-1} + (kA)^{-1}B(-C(kA)^{-1}B)^{-1}C(kA)^{-1})(kb)^* \qquad (3.5) \\ &= (k^{-1}A^{-1} + k^{-1}A^{-1}B(-Ck^{-1}A^{-1}B)^{-1}Ck^{-1}A^{-1})kb^* \\ &= (k^{-1}A^{-1} + k^{-1}A^{-1}B(-k^{-1}CA^{-1}B)^{-1}Ck^{-1}A^{-1})kb^* \\ &= (k^{-1}A^{-1} + k^{-1}A^{-1}B(k^{-1})^{-1}(-CA^{-1}B)^{-1}Ck^{-1}A^{-1})kb^* \\ &= (k^{-1}A^{-1} + k^{-1}A^{-1}Bk(-CA^{-1}B)^{-1}Ck^{-1}A^{-1})kb^* \\ &= (k^{-1}A^{-1} + k^{-1}A^{-1}B(-CA^{-1}B)^{-1}CA^{-1})kb^* \\ &= k^{-1}(A^{-1} + A^{-1}B(-CA^{-1}B)^{-1}CA^{-1})kb^* \\ &= (A^{-1} + A^{-1}B(-CA^{-1}B)^{-1}CA^{-1})b^* \\ &= \lambda_1. \qquad (3.6) \end{aligned}$$

The same is true of the nugget effect, but this is not as easily demonstrated as with the sill. In order to demonstrate the effect of changing the nugget while keeping the other parameters fixed (as well as demonstrating this for the partial

sill and range), the cross-validation procedure used in Section 3.1 was modified to run for a range of covariance parameters rather than mixing parameters. To simplify things, only stream distance based detrending was considered and the covariance structure used was the tail-up stream distance model on its own. As in Section 3.1, this was done for each of the 21 years worth of data separately and within each year subsets of stations were removed (10% of those available at a time) and estimated until there were 100 estimates for each set of covariance parameters. The covariance parameters were chosen to be of roughly the same magnitude as those estimated earlier, so the nugget ran between 0.01 and 0.2 in increments of 0.01, the partial sill ran from 0.01 to 0.3 in increments of 0.01 and the range ran from 1km to 48.5km in increments of 2.5km. This means that for each year, there were 100 root mean squared prediction errors for each of the 12000 combinations of covariance parameters.

Figure 3.7(a) shows the results obtained when the nugget was moved through the range 0.01 to 0.2 while the partial sill and range were fixed at 0.08 and 6km respectively. These fixed values were arbitrary choices based on the estimated parameters for the tail-up model with stream distance based detrending (table 2.1). The actual values of the RMSPE would change if a different range was chosen (but not if a different sill had been used). Despite this, no matter what range was chosen, varying the nugget has no effect on the value of the RMSPE. This is also true of the sill, though the graph of varying sill is not shown. Figure 3.7(b) shows the effect of varying the range between 1km and 48.5km while keeping the nugget and partial sill fixed at 0.02. This demonstrates that varying the range does affect the predicted values, and thus the accuracy of predictions. This will be investigated further in section 3.2.2

Showing theoretically that the nugget does not affect the kriging weights is more difficult than for the sill. It can be seen analytically that in Σ^{-1} (3.4) only



(a) Fixed partial sill and range, varying nugget

(b) Fixed nugget and partial sill, varying range

Figure 3.7. Demonstrating the effect of nugget and range

the lower right element $(-CA^{-1}B)^{-1}$ will change if $(A + k)$ for some constant nugget k is used in place of A. All other three elements remain constant. Any proof of this sort will require identity (3.7) (Miller, 1981), if A and $A + K$ are invertible, and K has rank 1 and g=trace(KA^{-1}). Setting K to be an $n \times n$ matrix with entries in all cells set to k will fulfill the requirement of K having rank 1. Proving that all but the lower right element of (3.4) remain constant is beyond the scope of the investigation being conducted here, but the simulation study results show, numerically at least, that the nugget does not affect the estimated values in kriging.

$$(A + K)^{-1} = A^{-1} - \frac{1}{1 + g} A^{-1} K A^{-1} \tag{3.7}$$

3.2.2 Choosing a Covariance Model Based on RMSPE

The nugget and partial sill of the (co)variogram do not affect the estimated values from kriging. It has been shown that varying the range parameter does affect the estimates. However it should also be noted that changing the covariance

structure also affects the estimates, and so this will be investigated further in this study. To see that different covariance structures produce different estimates, we recall the example from Section 3.2.1 where there were observations at 5 and 10km and we fixed the covariance parameters as $c_0 = 0, c_1 = 1, c_2 = 10$. This means that, according to the exponential variogram, $\gamma(1) = exp(-0.1) = 0.9048$ and $\gamma(5) = exp(-0.5) = 0.6065$. If the linear with sill model (defined for covariograms in (1.5), and expressed as a variogram using (1.7)) was used instead, we have $\gamma(1) = 1 - \frac{1}{10} = 0.9$ and $\gamma(5) = 1 - \frac{5}{10} = 0.5$. This means that the ratio of semivariances has changed very slightly, with more weight being put on the 5km observation in the linear with sill model than in the exponential model.

The choice of covariance structure is normally a choice made by the investigator, usually after looking at the general shape of the empirical covariogram. However, it is not always easy to select a covariogram model based on shape. For example, the Euclidean distance based covariogram (with Euclidean distance based detrending) shown in Figure 2.9 could be interpreted as having come from an exponential model or a linear with sill model because the covariances drop off so quickly. This makes it very difficult to tell whether a linear with sill or exponential structure will be most appropriate.

Peterson and Ver Hoef (2010) investigate the 'best possible' covariance structure by using cross-validation on data from a river network in South East Queensland, Australia. They use fish assemblage data sampled from 86 locations to generate PONSE (the "Proportion Of Native fish Species Expected") scores. These scores are then modelled using a two step procedure, first fixing the covariance structure and predicting the best set of covariates to use with each different mixture of covariance structures, using the Akaike Information Criterion to select variables. The next step involved fixing the covariates for each model and estimating the best covariance model and covariance parameters for each of the

components in the mixture of covariance structures. Every combination of tail-up, tail-down and/or Euclidean structures was considered, with the spherical, exponential, MARIAH (Ver Hoef et al., 2006), and linear-with-sill functions considered for the tail-up and tail-down structures and the spherical, exponential, Gaussian, and Cauchy functions considered for the Euclidean. A leave-one-out cross-validation procedure was then used on each of the 124 different combinations of covariates and covariance structures in order to identify those with the lowest RMSPEs.

The cross-validation procedure that will be used here is similar in style, but different in detail, to that used in Peterson and Ver Hoef (2010). The aim of cross-validation here is to determine whether a different covariance model with different covariance parameters for either or both the Euclidean and tail-up stream distance based covariance structures could reduce the root mean squared error, and assess how much it could be reduced by. In order to do this, three different covariance models will be used for each of the Euclidean and stream distance based elements of the mixture model, and a range of covariance parameters will be tested for each. Specifically, the exponential (as used in previous analysis), spherical and linear with sill models will each be used with range parameters running from 2km to 50km in steps of 2km when used for the Euclidean covariance and from 2km to 78km (again in steps of 2km) when used for the stream distance covariance. It has already been demonstrated that the nugget and sill do not affect the predicted values, and so these will not be considered here and will be fixed to 0.02 and 0.1 respectively. Additionally, mixing parameters, λ, from 0 to 1 in steps of 0.05 were used with each possible combination of model and range. For every year of data, subsets consisting of 10% of the available monitoring stations were removed from the dataset and had their values predicted using kriging (and each possible model/range/mixture combination) until predictions had been made at 300 locations. This means that there were $21 \times 300 = 6300$ prediction

Table 3.2. Lowest RMSPE for each combination of covariance models and the range parameter at which it was found

Euc Model	Euc Range	Stream Dist Model	Stream Dist Range	Mixture	RMSPE
Exp	8	Exp	36	0.45	0.3220
Exp	8	Lin	48	0.45	0.3193
Exp	8	Sph	62	0.45	0.3205
Lin	38	Lin	58	0.20	0.3133
Lin	38	Exp	60	0.20	0.3155
Lin	38	Sph	76	0.20	0.3147
Sph	14	Sph	58	0.45	0.3193
Sph	22	Exp	46	0.35	0.3208
Sph	14	Lin	46	0.40	0.3182

locations/times in total, and there were 3(Euclidean models) $\times$ 3(stream distance models) $\times$ 25(Euclidean ranges) $\times$ 39(stream distance ranges) $\times$ 21(different mixing parameters) = 184275 possible model combinations used to fit at each of them. The RMSPE was calculated for each one of these combinations, and the results presented in the table represent the lowest RMSPE found for each combination of models.

For simplicity, just the stream distance based detrended data will be assessed here, as opposed to both Euclidean and stream distance based methods. As analysis so far has shown that the difference between these detrending methods is not large, it was decided that just one would be sufficient for this study and the stream distance method was chosen. This was a rather arbitrary choice as there seems to be equal merit to using either one, from both a statistical and environmental viewpoint. However the stream distance method was chosen as the 'best' mixture model as the estimated covariance parameters in Section 2.2.1 was found at a mixture of $\lambda = 0.1$. This value was felt to be quite low, and so it was decided to use the stream distance detrended data to see if this was the mixture that produced the lowest RMSPE again.

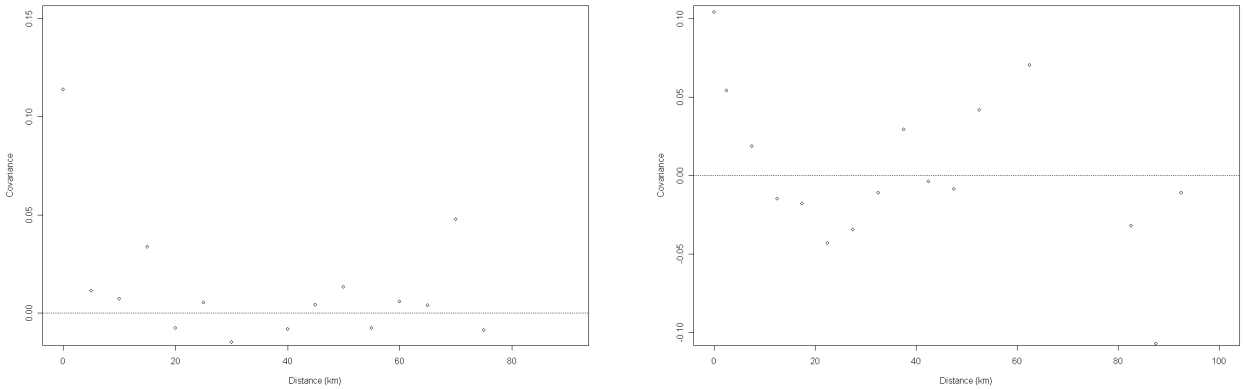
Table 3.2 shows the results from the cross-validation. Each line represents

a different combination of exponential, spherical and linear with sill covariance models for both the Euclidean and tail-up stream distance elements of the mixture model. Each line shows the minimum RMSPE found for each model and the range parameters and mixture of the covariance structures at which it was found. The RMSPE shown is the average RMSPE calculated over all 21 years worth of data.

The results show that there is not a large difference between the different combinations of models. The lowest RMSPE of 0.3133 is found by using the Linear model for both Euclidean and tail-up components, while the highest uses the exponential model for both models and has RMSPE of 0.3220. This means that the difference between highest and lowest is just 0.0087. The differences between the different models are so small that it is possible that fine-tuning the optimum ranges and mixing parameters, by finding the RMSPE at smaller intervals around these 'optimum values', could mean that a different combination of models was better than using linear with sill for both. While Peterson and Ver Hoef (2010) identify the value of the lowest RMSPE, and the covariance models used to achieve it, for each of the mixtures they consider there is no indication of the magnitude of improvement that can be expected by using one type of covariance model over another. The results in Table 3.2 suggest that any improvement is likely to be minimal.

An interesting thing to consider here is the range of shapes that each model can possibly take, and the impact this has on the prediction. Of the three models tested, the linear with sill model is by far the least malleable. If the true underlying covariance model were a linear with sill model then both the exponential and spherical models could take on an almost linear shape by increasing the range parameter towards infinity. However, if the opposite were true, and the data had come from either the spherical or exponential models, then the linear with

sill model would not be able to adapt its shape in any way to model the covariogram. Figures 3.8(a) and 3.8(b) show again the covariograms constructed using the River Tweed data with stream distance based detrending. The shape of the underlying covariance model is not instantly obvious in either but is especially vague in 2.13(a), where the linear model seems reasonable but the short range means that any curvature present in the empirical covariogram is very difficult to see with the binning at larger intervals such as this. However, if we had seen a clear 'exponential' shape then it is likely that the linear model would not fit well at all, and therefore would not have come out as the best model, as observed in Table 3.2.



(a) Euclidean distance covariogram

(b) Stream distance covariogram

Figure 3.8. Binned covariograms, stream distance detrending

The stream distance based detrending method was used here as it was felt that it could be interesting to examine the previously estimated 'optimal' mixing parameter of $\lambda = 0.1$, found using the exponential model for both Euclidean and tail-up covariance structures, and assess whether the optimum mixture would be found at such a low λ value. The lowest RMSPEs for each of the models, shown in Table 3.2, is found using a mixture somewhere between 0.2 and 0.45, suggesting that the previously estimated figure seems rather low. One reason for this might be that the range parameters, which were previously fixed, are now

allowed to take a range of values. This means that, while the best mixture using the previously estimated Euclidean and stream distance ranges may well have been at 0.1, if all three parameters were optimised, the best mixture parameter would likely be slightly higher.

Now consider the range parameters that provide the lowest RMSPEs for each of the model combinations. The range parameters estimated for the empirical covariogram were 8.9km and 5.5km for the Euclidean and stream distance based components respectively. The range for stream distance based covariance seems far too low in the context of the kind of ranges providing the lowest RMSPEs in Table 3.2. The estimated parameters for the exponential model provide a more direct comparison than the spherical and linear models, and in Table 3.2 the stream distance structures that use the exponential model are estimated to have ranges of 36, 60 and 46km. It is appropriate to question why the empirical estimate is so much smaller than the optimum values estimated here? There seem to be two possible explanations for this. The first is that the noise in the empirical covariogram has seriously affected the ability to accurately predict the covariance parameters. The other is that separate rather than joint estimation of the covariance parameters in the mixture model has had an effect. Recall that both elements of the mixture model have previously had their covariance parameters estimated separately, so it is entirely possible that the difference between the estimates and 'optimum values' is a result of the fact that this analysis is considering the effect of changing all three elements – Euclidean range, stream distance range and mixing parameter λ – simultaneously.

Moving on to look at the optimum range parameters for the Euclidean distance element of the covariance mixture model, it is interesting to see that the optimum range parameter is almost always constant for each covariance model shown in Table 3.2. This is rather surprising, as the same is not true for the stream

distance element. All three combinations of models that include an exponential Euclidean covariance model have an optimum range of 8km, while all three linear with sill models have optimum range 38km. While two of the three spherical Euclidean covariance models in Table 3.2 also have the same range of 14km the third breaks the pattern by having the minimum RMSPE at 22km. There also seems to be a high level of similarity in the best mixtures. For all three models that have an exponential Euclidean covariance structure, the optimal λ is 0.45 and all three that have the linear with sill model have the optimal λ as 0.2. The three combinations with a spherical model for the Euclidean covariance all have different optimal λ values: 0.35, 0.40 and 0.45. As the range of λ values used in the cross-validation were in intervals of 0.05, it is plausible that the difference seen here may be due to rounding and the results may suggest that the optimal λ is the same for all combinations of mixtures using the spherical Euclidean covariance model too.

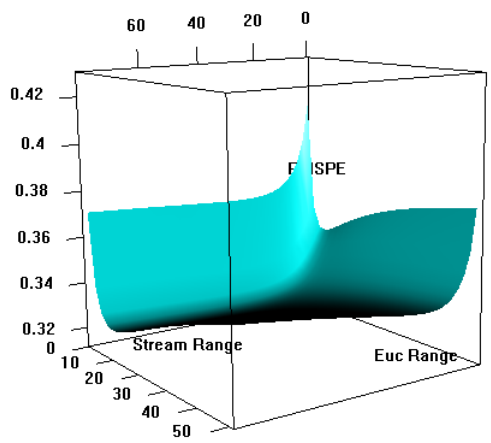
One may question the reason for such strong similarities in range and mixture parameters when using a particular Euclidean covariance structure, and not for particular stream distance structures. It would be convenient to use this as evidence that it is possible to estimate the two components of the mixture model separately, given that it suggests that no matter which model, or range parameter, has been used for the stream distance based element of the mixture model, the range and λ parameters will remain constant. This would seem likely if the opposite was also true and the range parameter for the stream distance element of the mixture model remained constant for each covariance model used, but this is definitely not the case here. This may be another example of the covariance structure balancing out the detrending. Only stream distance detrending has been considered for this example, and so it is possible that this is removing a large part of the stream distance based process in the data and leaving a fairly well defined Euclidean process. This could then lead to more agreement between

the Euclidean ranges estimated for the different models, while the stream distance element of the covariance has a less well defined structure left in the data leading to quite different range estimates.

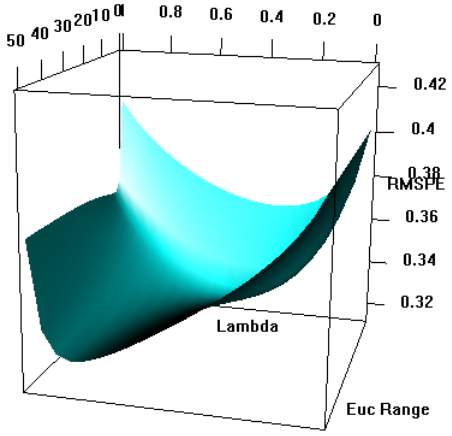
In general, there is even more evidence that the covariance structures, ranges and mixing parameters balance one another out to create a reasonably consistent root mean squared error no matter which combination of models is used. This is similar to what was seen when comparing the Euclidean and stream distance based detrending methods in the previous study. In Table 3.2, the difference between the smallest RMSPEs for each combination of models is minimal, less than 5% in all cases. This implies that the range and mixing parameters at which these lowest values are found are changing to suit the different models that are being used for each. This is evidence that the model chosen for each distance metric is not as important as the range and mixture parameters chosen with it.

3.2.3 Further Investigation into the Different Covariance Models

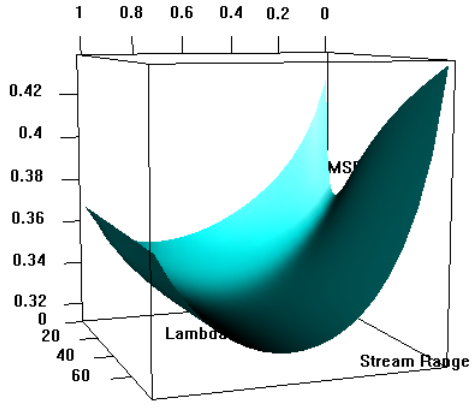
To understand the results of this study in more detail, plots were constructed showing the change in RMSPE over the range of values taken by each of the varying parameters (Euclidean range, stream distance range and λ) in the cross-validation. This is a four dimensional problem and so to help visualise it, each plot was constructed by fixing one of the varying parameters at its 'optimal' predicted value. Figure 3.9 shows the effect that changing the parameters of the exponential Euclidean, exponential stream distance covariance mixture model has on the RMSPE in the cross-validation. Figures 3.9(b) and 3.9(c) respectively fix the stream distance and Euclidean range parameters and show that quite significant decreases in RMSPE are observed as λ moves between 0 and 1, confirming earlier



(a) Fixed $\lambda = 0.45$



(b) Fixed stream range = 36km



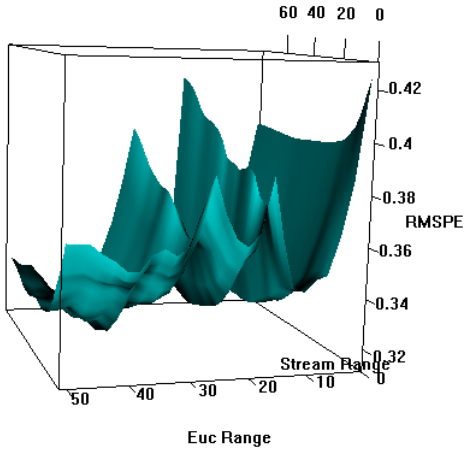
(c) Fixed Euc range = 8km

Figure 3.9. Demonstrating the effect on RMSPE of changing the parameters of the exponential Euclidean, exponential stream distance mixture model

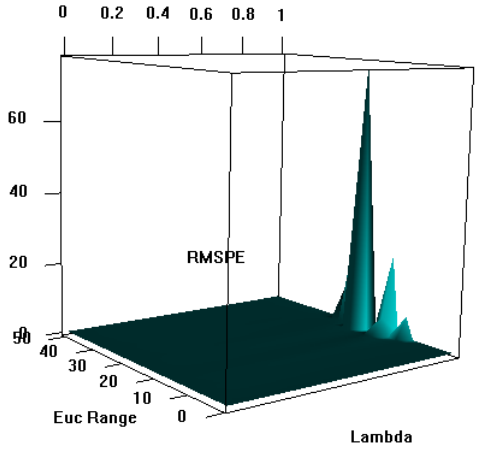
findings. However, there are potentially more interesting features that can be observed when λ is fixed instead.

Figure 3.9(a) shows the RMSPE as the Euclidean and stream distance ranges are varied when λ is fixed at 0.45. From this plot it can be seen that altering the Euclidean distance has far more effect on the RMSPE than altering the stream distance; a feature that is the same for all combinations of models that were considered. This may be a consequence of using stream distance based detrending in this study, meaning that there is less to be gained by making small adjustments to the range of the stream distance based element of the mixture model and more of a Euclidean structure left in the data that can be better accounted for by slight changes in the Euclidean element. This may be the reason that the optimum range parameters for the Euclidean distance appear to be very similar across all the Euclidean elements of the mixture model that use the same covariance model, and the reason why the same is not also true of the range parameters for the stream distance elements.

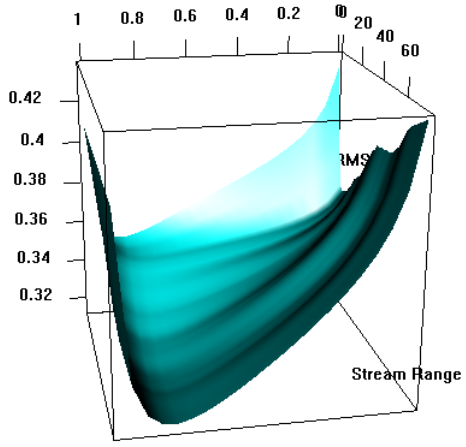
Figure 3.10 shows some anomalies in the results. These anomalies occur whenever the linear with sill model is used for the Euclidean element of the mixture model. Figure 3.10 shows what happens to the RMSPE as the optimum λ, stream distance range and Euclidean range are fixed. Figure 3.10(a) is very jagged along the 'Euclidean Range' axis compared to the very smooth plot for the exponential-exponential model shown in Figure 3.9(a). This seems to be at its worst in two large peaks around a Euclidean range of 16km and 26km. Figure 3.10(b) shows how bad this becomes when the optimum λ is not fixed, as the RMSPE is almost as high as 80 (as opposed to a maximum of around 0.42) around $\lambda = 1$ and a Euclidean range between about 14 and 28km (though the largest peaks are at 16km and 26km). Finally, Figure 3.10(c) confirms that the problem lies in the Euclidean distance element of the mixture model as fixing the



(a) Fixed $\lambda = 0.45$



(b) Fixed stream range = 36km



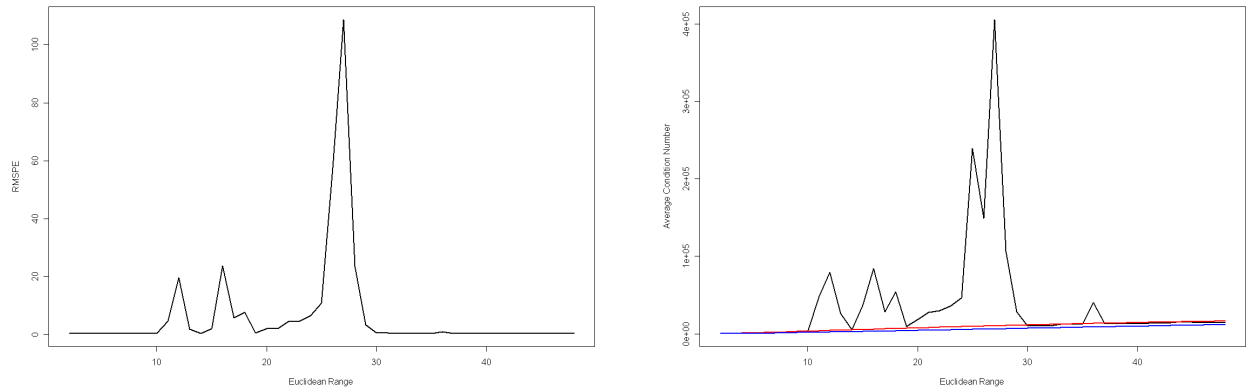
(c) Fixed Euc range = 8km

Figure 3.10. Demonstrating the effect on RMSPE of changing the parameters of the linear with sill Euclidean, linear with sill stream distance mixture model

Euclidean range at 38km gives a plot much more in keeping with those for the exponential-exponential mixture model (though the plot is still slightly irregular and not particularly smooth).

Why is the model that is by far the most simplistic giving such large errors? To investigate further, the problem was altered slightly to suit more closely the conditions under which such large RMSPEs are occurring. The anomaly seems to be due to the Euclidean distance element of the mixture model and so this time instead of a range of λ values, λ was fixed at 1. This is essentially just the Euclidean model on its own with no tail-up element in the mixture. This time a larger range of Euclidean range parameters was considered, running from 2km to 48km in 1km increments. The same cross-validation as before was run on this modified problem but this time the condition number of the matrices formed for kriging (matrix Σ in (1.21)) were recorded for each of the different runs. The condition number of a matrix gives a measure of how close the matrix is to being singular (Cheney and Kincaid, 2007). A large condition number indicates that the matrix may be ill-conditioned, and thus close to being singular, while a small condition number indicates that the matrix is well conditioned. The magnitude at which a matrix is ill-conditioned will increase as the size of the matrix itself increases, and so rather than compare the condition numbers to a test statistic we will compare them to the condition numbers of the matrices formed using different covariance models.

Figure 3.11(a) shows the changing RMSPE of the linear with sill model as the Euclidean range parameter is increased. The exponential and spherical models are not shown as the RMSPE is below 0.42 at all times in each and so barely show up on this scale. Figure 3.11(b) shows the mean condition number of the matrices used to obtain these results (the matrices, and thus the condition number, change for each one of the 300 runs as a different subset of data is excluded/predicted



(a) RMSPE of linear with sill model as Euclidean range increases

(b) Condition number of all three model combinations as Euclidean range increases

Figure 3.11. Investigating the condition number of kriging matrices

in each and thus the average of all the condition numbers is shown). The black line shows the condition number when using the linear with sill model, while the red and blue lines show the condition numbers for the exponential and spherical models respectively. The peaks in condition number seem to match up with the peaks in RMSPE in Figure 3.11(a). It can also be seen that the condition numbers from the other two models are much lower than those taken by the linear with sill model at these peaks, implying that the large RMSPEs are due to the fact that the matrices that produce these predictions are ill-conditioned. It is worth noting that the linear with sill model is not always ill-conditioned, as it has similar magnitude to the other models' when the range parameter is below 10km or above 30km (with the exception of a slight fluctuation at around 37km.) and, at these ranges, the RMSPE shown in Figure 3.11(a) seems much more plausible too.

It is unclear why the linear with sill model should create an ill-conditioned matrix. It is a valid variogram/covariogram model (Cressie, 1991; Webster and Oliver, 2001) when used with Euclidean data, as is the case here, and so should

not produce singular matrices. The matrices being produced here are not singular but just close enough that the predictions are adversely affected. Closer inspection of the results from each year in turn suggest that certain combinations of available monitoring station locations, dropped out stations and Euclidean range parameters lead to high condition numbers. The jagged nature of the plot recalls Ver Hoef et al. (2006, Figure 1), which demonstrates the need for the tail-up model by showing the eigenvalues from the covariance matrices generated for Euclidean models with stream distance. For the linear with sill model, the line is very 'bumpy', especially when compared to the smooth exponential and spherical lines. It seems possible that something similar is occurring in the observed condition numbers for study 2. It is surprising that this would be observed when using Euclidean distance, as the model should produce a valid covariogram, but the nuances of this particular dataset might be causing the matrices to become ill-conditioned when certain combinations of range/mixing parameters are used. There are a few pairs of stations that are located very close to each other on the river network, sometimes with a (Euclidean) distance of just forty metres between them. It may be that the added curvature of the exponential and spherical models gives more of a difference between the rows/columns corresponding to pairs of locations like these, meaning that it is just the linear with sill model that is struggling under these conditions.

3.2.4 Conclusions from Study 2

Study 2 has demonstrated the role the range parameter in the covariance structure plays in determining the estimated value in kriging. It has also shown that neither the nugget nor the sill affect kriging estimates, and only impact upon the standard error calculation. Section 4.3.2 will return to this concept.

Having shown the role of the range parameter, cross-validation was constructed to test three different covariance models and establish how they affect the root mean square error. Three models, exponential, spherical and linear with sill, were used in both the Euclidean and stream distance components of the mixture model in each of the nine possible combinations. These models were used with a variety of mixing and range parameters in order to determine the best combination for each of the nine models.

The differences between the lowest RMSPEs for each model combination were modest, suggesting that the type of covariance model does not have a significant effect on the prediction error. The lowest RMSPE was found by using the linear with sill model in both the Euclidean and stream distance components of the mixture. Despite this, further investigation showed that the linear with sill model was very volatile when used for the Euclidean covariance structure. The condition numbers of the covariance matrices produced in this way were, under certain conditions, very large. This suggests that the matrix was close to singular. Consequently, care should be taken if the linear with sill model was to be used as part of the Euclidean covariance structure.

The range and mixing parameters at which the lowest RMSPEs were found also gave further insight into the mixture model. The optimum mixtures were all between 0.2 and 0.45, suggesting that the mixing parameter estimated in Section 2.2.1 might have been slightly too low. For the range parameters, the Euclidean covariance model was very consistent in the ranges at which the lowest RMSPEs were found. All three exponential models had optimum ranges of 8km, all three linear with sill models had optimum range 38km and two of the three spherical models had range 14km. However, the range parameters for the stream distance model were not as uniformly consistent. This, combined with the differences in lowest RMSPE being very small across the model combinations, may suggest that

we are again seeing the optimum models balancing themselves out to explain as much of the stream and Euclidean distance based processes present in the data as possible. Therefore, it seems that the choice of covariance model is not hugely important if it is possible to determine the best range and mixing parameters in a study such as this.

3.3 Study 3– A Sensitivity Study for Trend Bandwidth

Section 2.1 outlined two methods for smoothing the data in order to produce a nonparametric trend surface using Euclidean and stream distance based smooths respectively. The chosen bandwidth has a large impact on the way these surfaces are estimated, with a large bandwidth being much smoother and more general while a small bandwidth gives an estimate that is more responsive to individual observations. It is important to try to choose a bandwidth that will balance out generality with responsiveness, but it is also vitally important to ascertain that this choice will not have a dramatic effect on later results. In Study 3, the effect of changing the bandwidth of the trend will be assessed using a sensitivity study.

A cross-validation procedure was used in order to test the sensitivity of the bandwidth. For bandwidths ranging from 10km to 100km in 10km increments, data from 8 stations (corresponding to 10% of the total) were excluded at random and a trend surface predicted using smoothing on the overall averages from the remaining stations (i.e. the averages over all 21 years worth of data). Kriging was then carried out on the yearly average data for all 21 years with a range of covariance parameters. This process was repeated 100 times for each of the bandwidths with different stations excluded each time, so that in total there

Table 3.3. Results of Sensitivity Study for Euclidean Distance Detrended Data into Trend

Bandwidth (km)	Euclidean Range (km)	Stream Range (km)	RMSPE
18	1	26	0.3343
20	47	20	0.3072
30	63	31	0.3062
40	73	28	0.3049
50	80	28	0.3033
60	84	22	0.3020
70	85	20	0.3011
80	90	18	0.3004
90	89	15	0.2999
100	93	16	0.2996

were 10 bandwidths, with 100 runs each, with 21 different years worth of data analysed in each run over a range of covariance parameters. For each bandwidth, the Euclidean and stream distance range parameters were allowed to vary between 1 and 100, and those that are shown in the results were the ones that gave the lowest RMSPE over the entire set of 100 runs. Only the range parameter was allowed to vary, as Study 2 showed that the nugget and sill have absolutely no effect on the predicted values, and thus the RMSPE.

This technique was performed for both Euclidean and stream distance based trends. In both of these cases, the covariance model used is a mixture model with Euclidean and tail-up stream distance based elements (1.13). The weighting between these two elements is fixed at $\lambda = 0.50$ for the stream distance trend and $\lambda = 0.10$ for the Euclidean trend. These are the same mixing parameters that were estimated as part of Study 1.

Tables 3.3 and 3.4 show the results of this analysis for Euclidean distance and stream distance detrending respectively. The results appear to suggest that the lowest RMSPE is found at the highest possible bandwidth. Only a subset of possible bandwidths were used, but it is important to remember that, in an area the size of the Tweed, a bandwidth of 100km produces a very smooth surface. The

Table 3.4. Results of Sensitivity Study for Stream Distance Detrended Data into Trend

Bandwidth (km)	Euclidean Range (km)	Stream Range (km)	RMSPE
10	1	5	0.2896
20	50	10	0.2870
30	14	23	0.2818
40	11	26	0.2766
50	9	30	0.2734
60	7	36	0.2716
70	6	45	0.2704
80	5	53	0.2697
90	5	78	0.2692
100	4	91	0.2689

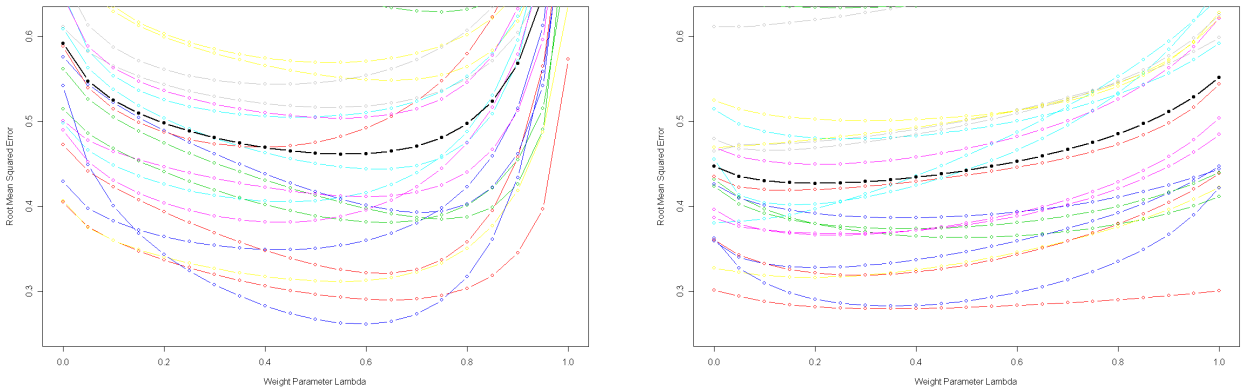
decrease in RMSPE seems to be negligible after around 20km for the Euclidean trend, but is still dropping up to 100km. If this trend continued on after 100km, the best possible RMSPE should be found by removing the overall mean value from the data. It is difficult to say for certain at which point there is negligible change in RMSPE for the stream distance trend, as there is little change as the bandwidth increases. The analysis was repeated using several fixed sets of covariance parameters, but this made no difference to the overall conclusion that the lowest RMSPE is found at the highest bandwidth. The Euclidean distance based detrending seems to be more affected by the change in bandwidth than the stream distance based detrending, but this may be simply as a result of starting with a higher RMSPE at the lowest bandwidth than the stream distance detrending method. It is worth noting that the improvements in RMSPE found by increasing the bandwidth are very slight and so it seems that the bandwidth has very little impact on the accuracy of future predictions. So therefore it can be argued that the bandwidth should be chosen so that systematic trend in the system is removed but while ensuring that the trend does not fit the data too closely.

Taken to its logical conclusion, the results from this study would imply that

as high a bandwidth as possible would lead to the lowest RMSPE. As the bandwidth tends to infinity, the predicted trend tends to an overall average of all observations. However this does not seem to make sense given the evidence for detrending that has already been demonstrated. The reason for this result almost certainly lies in the construction of the sensitivity study. The trend in this study was re-estimated for each set of stations that were 'dropped out', but by adjusting the study so that the trend was only estimated once for each bandwidth (no matter what stations were excluded) the results did not suggest that a higher bandwidth was better. However, not excluding these stations from the trend calculation would bias the results, and so therefore it was decided to keep this study in its current form. The results from this study do suggest that the effect of altering the bandwidth of the trend is minimal, and so analysis will continue using the trends calculated in the previous chapter.

To double check the effect that a really high bandwidth has on the prediction accuracy, the cross-validation used in Section 3.1 was carried out again, but having removed a 'grand mean' trend i.e. a trend with an 'infinite' bandwidth. For the Euclidean distance detrending this is fairly straightforward, with the overall mean of all stations being used as the trend. For stream distance based detrending, equation (2.5) with infinite bandwidth means that the predicted trend at a particular station is the mean of all stations that are flow connected to it.

As different trends were removed, new covariance parameters had to be estimated for each distance metric using the same method described in Section 2.2. For the Euclidean distance grand mean the parameters are exactly the same as those estimated when no detrending has been performed. This can be seen since $Cov(X,Y) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})(y_i-\bar{y})}{N}$ when no detrending has been performed, but if the grand mean $\mu = \bar{x} = \bar{y}$ has been subtracted beforehand then the top line of the fraction becomes $((x_{*i} - \mu) - \bar{x_{*.}})((y_{*i} - \mu) - \bar{y_{*.}})$. However, since the data



(a) Euclidean distance based grand mean detrending (b) Stream distance based grand mean detrending

Figure 3.12. Assessing the RMSPE for different mixture models using grand mean detrending

have been detrended, $\bar{x_{*.}} = \bar{y_{*.}} = 0$ and so the covariance is exactly the same as it would have been if no detrending had been performed at all. This means that the estimated RMSPEs are also exactly the same as those shown in Section 3.1, where no detrending had been performed beforehand. This is because ordinary kriging has been used to predict at the excluded locations, which removes a grand mean. Consequently, there is no difference between removing a grand mean before analysis, and accounting for it in ordinary kriging.

The connectivity data in the stream distance based grand means mean that the covariances are not the same, and so the estimated covariance parameters for that method were nugget = 0.035, sill = 0.082 and range = 13.8km for the Euclidean distance based covariance and nugget = 0.047, sill = 0.084 and range = 7.27km for the stream distance based covariance. Figure 3.12 shows the RMSPEs for both the Euclidean and stream distance based 'Grand Mean detrending' (the Euclidean one being exactly the same as that shown in the non-detrended analysis in Section 3.1). The prediction errors are significantly higher than those obtained after removing a 'proper' trend (i.e. with finite bandwidth). The lowest RMSPEs are 0.4614 and 0.4270, which are found at $\lambda = 0.55$ and

$\lambda = 0.2$ for the Euclidean and stream distance based grand means respectively. There is also a much wider variability of values among the individual years' RMSPEs. This confirms that removing a trend with very large bandwidth is not the best course of action, and so the bandwidth sensitivity study results are not useful to show the optimum bandwidth, rather they show that altering the bandwidth does not have a large impact on the predicted values.

3.3.1 Conclusions from Study 3

Study 3 has shown that the effect of bandwidth of trend on the root mean squared error is modest. The results of the study suggest that a trend with a very large bandwidth would produce the lowest possible RMSPE. However this has been shown to be the incorrect conclusion by means of a cross-validation similar to that in Study 1. This suggests that the design of the study itself was not able to separate out the effects of trend and

This result is likely to be a consequence of the slightly inflexible nature of the sensitivity study conducted in Study 3. Previous analyses have allowed mixing parameters and even range parameters to change in order to find the lowest RMSPE, but these remain fixed in Study 3. It seems likely that the design of the sensitivity study itself is leading to the results seen here, rather than the bandwidths. Alternative approaches were tested but none could eliminate this problem. It was not possible to separate the trend bandwidth effect from the other parameters that either had to be fixed or allowed to vary in the design of the study.

The difference between lowest and highest RMSPE in this study is only around 10% and it is likely that this difference may be due to the fixed covariance parameters. Other studies in this chapter have shown evidence that the trend and

covariance balance one another out when the best covariance parameters are found by their RMSPE, and this may also provide some explanation as to why it was not possible to separate the trend bandwidth effect from the effect of the other varying parameters. The aim of this study was to investigate the bandwidth of the trend and what effect it has on the RMSPEs. It can be concluded that there definitely is an effect, but that it is minimal and likely to be even less when other parameters are allowed to vary too. Therefore, there can be some confidence that the trend bandwidths selected in Section 2.1 will not have much influence on the results in Studies 1 and 2, and that the flexible nature of the covariance structures used in these studies would adapt to allow for whatever bandwidth was used.

3.4 Study 4– Simulation Study

In Study 4, a simulation study will be carried out in order to investigate the properties that are exhibited by the tail-up models and the potential prediction accuracy obtained from different locations of sampling stations.

The basis of the simulation was the geography of the River Tweed, including the flow data, which is used to determine the weighting structure used in the tail-up model, and data simulated at various locations. Initially, attempts were made to simulate a river network structure as well as data. These simulated networks were very artificial, much more closely resembling a 'rooted tree' than a river network. This approach was rejected quite early on, as the straight 'rivers' biased results heavily toward Euclidean distance-based models. Simulating the river network geography is not impossible given access to GIS software. For example, Peterson et al. (2007) simulate the geography of a small river network, however they use it to assess the validity of different covariance structures rather

than for a simulation study such as the one conducted in this chapter. There may be future potential for simulation studies using a simulated river network using GIS software and a freely available toolbox called the Functional Linkage of Watersheds and Streams (FLoWS) (Theobald et al., 2005). However, access to such software was limited and it was felt that focusing too much attention on the river geography would detract from the more interesting questions. Therefore the River Tweed structure was used for simulations. There is no reason to believe that the Tweed is an atypical example of a river network, and so it would be hoped that results obtained would be representative of those that could be expected for different river network geographies.

3.4.1 Study Specification

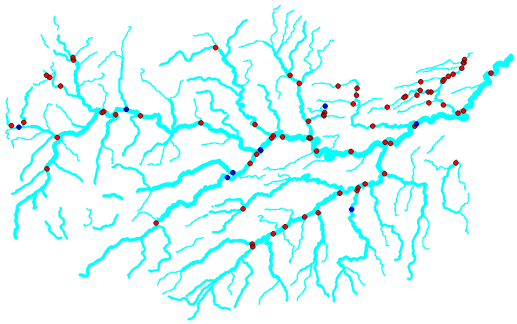
Using the Tweed network, data were simulated at a variety of locations. The responses were drawn from a multivariate Normal distribution with mean zero and covariance matrix Σ, corresponding to the covariance matrix created by the mixture model (1.13) over a range of λ mixing parameters. Since the simulations are generated with mean zero, there is no systematic trend present in the generated data. This means that there is no need to detrend the data before analysis. The parameters used in the mixture model are those that were estimated after the stream distance based detrending of the River Tweed data, shown in Table 2.1. These parameters are used as they are assumed to be representative of those that would be found on any river network. The covariance parameters estimated using the Euclidean distance based detrending could also have been used but a decision was made to concentrate on the stream distance based detrending method's parameters.

Table 3.5. Numbers of simulations carried out for each sampling scenario

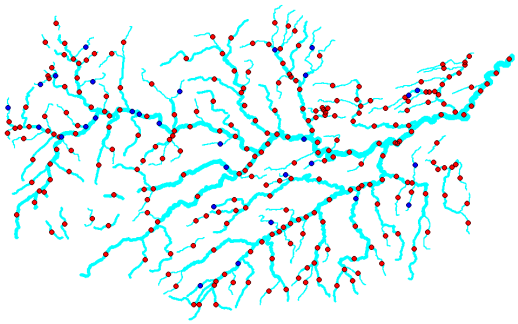
Scenario	*Sampling Sites*	*Prediction Locations*	*Runs*	*Total Simulations*
Tweed Sample	77	8	540	4260
Full Sample	268	30	144	4260
Other Sample	85	213	20	4260

As demonstrated several times in Chapter 2, a weighted average of the Euclidean and stream distance based covariance structures has the potential to provide a more accurate means of prediction than either model on its own. In order to further investigate the mixture model, simulations will be generated using underlying covariance structures corresponding to the mixture model over a variety of λ mixing parameters (1.13). Three separate sampling scenarios were examined in the simulation study, and the numbers of sites and simulated runs are detailed in Table 3.5, while representations of these structures are shown for illustration in Figure 3.13. In these maps, the red circles correspond to simulated sample locations while the blue circles are simulated prediction locations to be predicted using the sample locations.

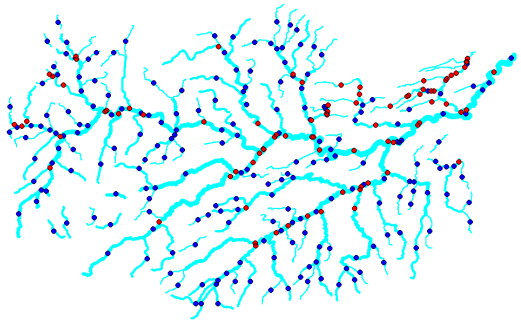
The first sampling scheme is referred to as the "Tweed Sample" and simulates at the 85 sampling locations monitored on the Tweed by SEPA. Eight of these locations were then removed and their values predicted using the remaining 77 locations. This was repeated until just over 4000 simulations were obtained. This entire process was then repeated so that results were obtained for every λ between 0 and 1 in increments of 0.05 in the underlying covariance structure. It should be noted that, in order to keep the results comparable, the seed on the computer was reset for each true λ to ensure that the same simulated values and subset of dropped out stations were used for each. Figure 3.13(a) illustrates this sampling scheme but it should be remembered that the blue 'prediction locations' are a random subset of all 85 locations, and the subset shown are just for illustration.



(a) "Tweed Sample"



(b) "Full Sample"



(c) "Other Sample"

Figure 3.13. Examples of simulation sampling structures. Red stations denote simulated 'monitoring sites' while blue stations denote 'prediction sites'.

The second sampling scenario, referred to as the "Full Sample", is more of a 'best case' sampling scenario. This time, one simulation was generated on each of the 298 stretches of water on the Tweed. A random sample of 30, representing around 10% of the total number of simulated locations, was dropped out and their values predicted using the remaining 268 locations. This was repeated until the number of simulations matched that of the Tweed Sample, and then the whole process was repeated for each true λ mixing parameter. Figure 3.13(b) illustrates this sampling scheme and, again, the blue 'prediction locations' are a random sample from all locations and will change for each simulated run.

The final scenario, referred to as the "Other Sample", was designed by again generating simulations on each of the 298 stretches of river. Then, the 85 sampling locations actually monitored by SEPA were used to predict at the remaining 213 locations, and this process was repeated until the number of simulations matched the number in the other two sampling schemes. This sampling scheme is something of a worst case scenario as the preferentially sampled (Diggle et al., 2010) nature of the Tweed locations means they are likely to cluster and many of the remaining 213 locations are likely to be further downstream. Figure 3.13(c) illustrates this sampling scheme, but here the prediction locations and sampling locations remain fixed over all simulations.

In each of the sampling scenarios, the covariance parameters used in predictions, apart from the λ in the mixture models, are the same as those used in the underlying model that generated the simulations. This allows the effect of λ to be separated from the process of parameter estimation. When comparing the results here with those obtained from the actual Tweed nitrate data, it is important to remember that it is assumed that the detrending process completely removed any trend in the data, as a zero mean has been specified for all the simulated values. Despite the detrending that was performed on the Tweed nitrate data it is unrealistic to think that every monitoring location has mean zero afterwards.

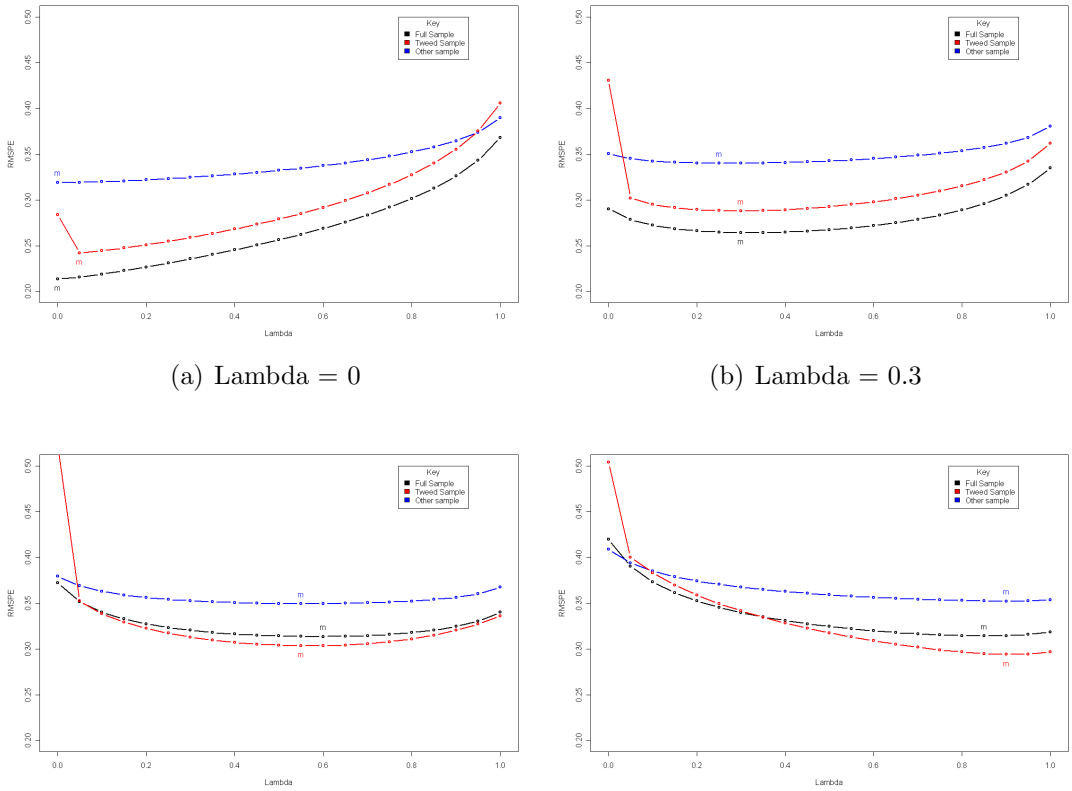
3.4.2 Analysis of the Mixture Model

The aim of each simulation was to find the λ at which the lowest RMSPE was found, and examine this in relation to the specified true λ used to generate the data. This would hopefully give insight as to whether the lowest RMSPE would be found at or near the mixture of covariance structures from which it was generated, and whether the different sampling schemes would lead to a systematic

deviation from this value.

Twenty-one graphs, one for each true λ value, with curves for each of the three sampling methods, were created and four of these plots are shown in Figure 3.14. The λ value which gives the lowest RMSPE in each plot is indicated by the letter "m" just above or below the line. As before, λ values of zero and one correspond to the Euclidean and tail-up stream distance covariance models (respectively) on their own. These plots show that the λ value that gives the lowest RMSPE is almost always the true λ value, or very close to it. In fact, out of all the λ values for all simulation types, there are only three occasions where the lowest RMSPE of any of the three sets of simulations is found more than 0.05 away from the underlying λ. The differences in RMSPE in each plot are very small over quite wide ranges of λ but it is very encouraging that the optimal performance is always at values of λ very close to the true one. This indicates that the mixture model with a λ value other than zero or one should provide more accurate predictions than either the Euclidean or tail-up stream distance based models on their own, if the underlying covariance structure is itself a mixture. This means that as long as reasonably accurate covariance parameters have been estimated, the fact that the lowest RMSPE is found at a certain λ mixing parameter would imply that the underlying covariance structure is the mixture model with a mixing parameter close to the observed optimal λ.

The relative positions of the three lines on the graphs gives some insight into potential bias that may be observed due to the placement of monitoring sites on the Tweed. As expected, the Other Sample (the worst case scenario) gives far poorer results than the other two sets of simulations and, although the difference does seem to decrease as the true λ increases, this is the case for all true λ values. This is not surprising, as it is to be expected that the locations which lie further away from clusters of monitoring stations will be predicted with less accuracy.



(a) Lambda = 0

(b) Lambda = 0.3

(c) Lambda = 0.6

(d) Lambda = 0.9

Figure 3.14. Simulation study results showing RMSPE bahaviour for three sampling schemes under four different true λ values

It is encouraging to see that the best prediction accuracy for these locations is generally found with an estimated mixing parameter close to the true one despite this relative lack of neighbouring stations.

The behaviour exhibited in the other two sets of simulations seems far more interesting. The Full Sample starts off having more accurate predictions than the Tweed Sample, but as the true value of λ increases this seems to change so that the Tweed Sample produces the more accurate predicted values more often. This is slightly surprising, especially given that the Full Sample has been referred to as something of a best case scenario. In reality though, it is a best case scenario in the context of having very regular sampling and so will produce something more like a 'benchmark' with which to compare the other sets of simulations rather

than always producing the best set of predictions. The reason for the Tweed Sample being either better or worse than this benchmark is almost certainly due to the use of two different distance metrics. The Euclidean model, and mixtures close to it, lead to lower RMSPEs in the more tightly clustered sampling of the Tweed Sample, while the stream distance model, and mixtures closer to it, lead to lower RMSPEs in the more diffuse sampling structure of the Full Sample. This may be due to the extra river network information compensating for the sparsity of neighbouring data.

3.4.3 Preferential Sampling

In this particular context, it may well be that the sampling locations are indeed neither random nor systematic, and recent developments in sampling theory have addressed "preferential sampling" (Diggle et al., 2010). The choice of location to sample may depend on the expected value of the measurement at that location, meaning that there is a stochastic dependence between the sampling locations and the nitrate value. This means that if monitoring locations are chosen on the Tweed because it is thought they will have high (or low) nitrate values, as opposed to choosing sites at random, then this would comprise preferential sampling. The use of such a sampling scheme would lead to bias in the covariogram estimation and the final predicted values. A sampling scheme with heavier monitoring around potentially high value nitrate areas will have the effect of over-estimating the nitrate levels over the entire area, while heavier monitoring around low value areas would produce under-estimates.

Given the layout of monitoring stations on the Tweed and conversations with staff at SEPA, the locations monitored on the Tweed are almost certainly located in a preferential manner. Accounting for a preferential sampling scheme

in analysis is still an area under investigation in the literature. Currently proposed methods (Diggle et al., 2010) to counter the bias inherent in these sampling schemes are very complex and computationally intensive and, given that analysis is already very computationally intensive, will not be adopted here. However, the simulation study suggests that the root mean squared errors produced in the three studies performed so far may be slightly biased due to the preferentially sampled nature of the river network. However, the simulation study also suggests that the conclusions of these studies will not be affected by this, as the preferential sampling does not lead the lowest root mean squared error to occur at mixing parameters different to the true value.

3.4.4 Conclusions from Study 4

The simulation study carried out in Study 4 has provided some valuable insight into the implications of the sampling scheme used for the River Tweed data as well as the results already obtained in Study 1.

Study 1 concluded that the best mixture of covariance structures for the Euclidean detrending method was a 50/50 split between Euclidean and stream distance based structures, and for stream distance detrending it would be a 90/10 split in favour of the Euclidean structure. The results from Study 4 have suggested that this implies that the underlying process at work on the Tweed will be very close to the mixture model with these mixing parameters.

Study 4 has also shown how different sampling schemes on the Tweed can affect the RMSPE estimated in cross-validation studies. The sampling scheme used for the River Tweed appears to be preferential in nature, with a higher frequency of sampling stations in areas with high nitrate levels. The simulation study suggests that the effect that this has on the RMSPE differs depending

on the underlying covariance structure. For true λ less than around 0.5, the prediction error for the Tweed sampling scheme is lower than a non-preferential sampling scheme. After sround $\lambda = 0.5$ the opposite is the case. This shows not only that the RMSPE will be affected by the sampling scheme, but that the way it is affected will alter depending on the nature of the underlying process from which the data were generated.

3.5 Conclusions from the four studies

This chapter has consisted of four studies designed to assess the different trends and covariance structures that could be used with river network data.

The first study concluded that using a mixture of the tail-up and Euclidean covariance structures lowered the prediction error when compared to either structures on their own. This result was the same for both stream and Euclidean distance based detrending, though the best mixture parameters were different for each method. The results from the final study suggest that this could imply that the underlying processes at work on the nitrate data will resemble the covariance mixture models with these mixture parameters.

A variation on study one, that re-estimated the covariance parameters for each different increment of the mixture model was also carried out. The results were virtually identical for all mixtures except $\lambda = 0$ and $\lambda = 1$ (the pure Euclidean and stream distance models), seeming to suggest that the way that the estimated covariance parameters changed was effectively "balancing out" the difference made by altering the mixing parameter. This was not surprising, given that the covariance structure that the mixture model was seeking to model was the same for each λ value, meaning that a "good" estimate of this structure should

be quite similar no matter what mixing parameter is being used. As the mixing parameter changed, the sills in the two components of it changed with it, but the ranges remained very similar. In the context of the subsequent results found in the second study, this explains why the root mean squared error is almost identical for each of the mixture models.

The second study examined the role that the estimated nugget sill and range of the (co)variogram play in calculating the predicted values in kriging. The nugget and sill were shown to have no effect, and are used only in determining the standard error attached to the predicted values. The range parameter was shown to be the only one to affect the predictions. With this result, the next part of the study looked at three different covariance models and tried to assess whether one would produce more accurate predictions of the River Tweed data. There was very little difference between any of the nine possible combinations of covariance models used. The lowest prediction error was seen when using the linear with sill model for both elements of the mixture model, but it was then demonstrated that this covariance model was very volatile when used with the Euclidean covariance structure. Therefore, it does not seem to make much difference what covariance models are used but further analysis should avoid using the linear with sill model for the Euclidean covariance structure, for this dataset at least.

The third study examined the role that the bandwidth of the fitted trend plays in the prediction error. It concluded that there is very little difference between the prediction errors as the bandwidth is increased and therefore that the bandwidths used in the trends in the previous chapter were justified. The lowest prediction error was seen at the highest possible bandwidth, but this seemed to be due to the way the sensitivity study had been conducted rather than this being the best bandwidth to use.

The final study looked at what impact the sampling structure used to collect

the data on the River Tweed is likely to have had on results. It suggested that the preferential nature of the sampling scheme will lead to bias in the prediction errors, but also that this bias will change depending on the nature of the process that generated the data.

The results from the four studies provide insight into how predictions can be made for the River Tweed data. Study 3 implies that the trend bandwidths will not have much impact on the accuracy of predictions and so the trends that were estimated in Chapter 2 should be sufficient. Study 1 implies that a mixture of covariance structures will provide the lowest prediction errors, and these will be found at a mixture of $\lambda = 0.5$ for the Euclidean distance based detrending and $\lambda = 0.1$ for the stream distance based detrending. These parameters will be used in the analysis of the Tweed data. Finally, as study 2 concluded that the choice of covariance model will not have a significant effect on the prediction errors, the exponential model will be used for both Euclidean and stream distance based covariance structures in subsequent work.

Chapter 4

Predicted Nitrate on the River Tweed

The nitrate values across the River Tweed will now be modelled and values will be estimated at unsampled locations. Results are presented separately for Euclidean and stream distance detrending, though many features are similar for each.

4.1 Euclidean Distance Based Detrending

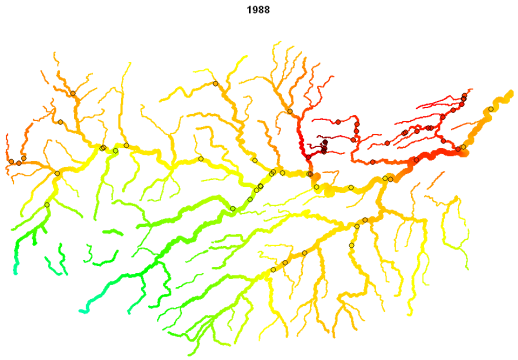
4.1.1 Comparing the Covariance Structures

After Euclidean distance based detrending, each of the locations on the river that make up the network had twenty-one years of yearly average nitrate levels predicted using kriging. Use of yearly averaged data or single snapshots in time has been standard in the literature. In general, the vast majority of locations on the Tweed are monitored around 9-12 times per year, with only a very small

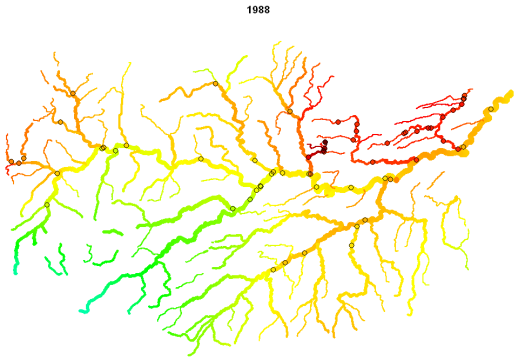
handful monitored less frequently and one station monitored more often. With differing numbers of observations available at different locations, there would be issues with the assumption of constant variance. However, it was felt that this assumption was valid given that the vast majority of the 83 monitoring sites had roughly the same sampling frequency.

In order to show the differences between the two distance metrics, the Euclidean and tail-up structures were initially used on their own in the covariance model (i.e. the mixture model with weights 0 and 1), as well as the best mixture of the two, $\lambda = 0.5$ (calculated in Section 3.1). It is worth noticing that, while the nitrate values displayed are on their original scale in the plots of predicted values, the colour scheme was created on the log scale, meaning that the colour scheme is more responsive to slight change at the lower end of the scale. Therefore, even a slight change in the shade of red denotes quite a large change in nitrate level.

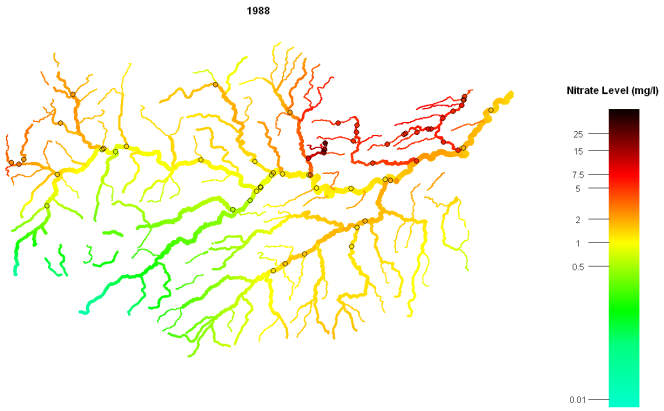
Figure 4.1 shows the the predicted values over the entire network for the year 1988, using the Euclidean, tail-up and mixture covariance models. At first glance, all three covariance model produce similar sets of predictions. The differences between them are subtle but are key to the argument of how best to represent a river network. The trend component dominates the most obvious features of the plots and in that respect but this does not imply that the covariance model does not matter, rather it shows that the large scale differences are mainly accounted for by the trend while the smaller nuances are provided by the covariance structure. As the distance between prediction points and observed locations decreases, so the difference between the covariance structures tends to zero. Therefore it is only certain areas in which there is a big difference between the covariance structures, but the differences between the covariance models demonstrate the arguments for using one, other or both of the distance metrics. These differences will now be discussed in more detail.



(a) Euclidean based covariance



(b) Tail-up covariance structure



(c) $\lambda = 0.5$ mixture model

Figure 4.1. Kriged Network plots after Euclidean distance based detrending

Area "b" (as marked on the key in Figure 4.2) demonstrates very well the differences between using only the Euclidean structure and only the tail-up structure. The Euclidean trend that was removed before analysis designated this area a very low nitrate value (green in colour) due to quite a low overall mean at the station, but before 1990 the level here is generally slightly higher (orange in colour). This leads to higher values predicted for the surrounding streams including, in the Euclidean covariance structure, the streams directly to the north and south-east. As these streams are not flow connected to the station though, the stream distance covariance model does not allow the value at the monitoring

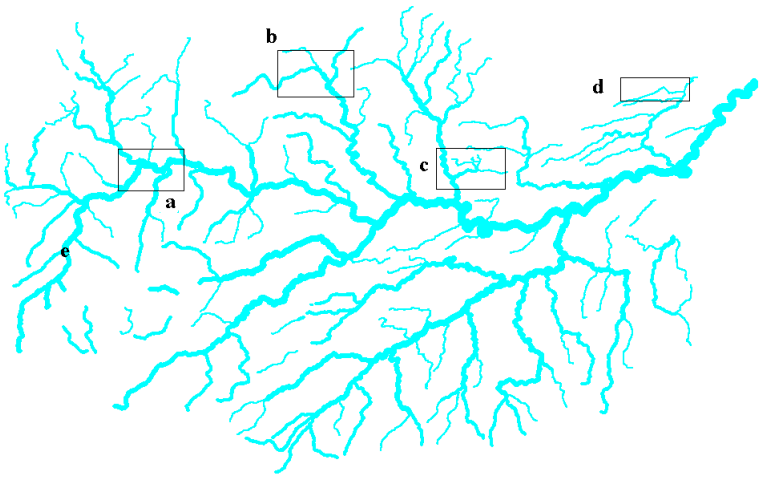


Figure 4.2. Location reference key

station to affect these unconnected streams in any way, meaning that lower values are predicted at these streams. In fact, the station with the most influence on the values at these streams under this covariance structure is the one very near the point where this network of streams meets the main river. Even this station has little effect due to the reasonably large distance between it and the streams. This means that, in the stream distance covariance model, the nitrate level is tending towards the estimated trend level in these streams, as very little additional information is being added in the kriging process. This demonstrates that the stream distance model assigns a covariance of zero to pairs of points that are not flow connected.

Whether or not this property is beneficial is a much more difficult matter to address, and depends on whether it is believed the higher values are due to diffuse or point source pollution. If there is a belief that the higher value is as a result of, for example, a slightly higher amount of fertiliser being used on a

field that borders both the stream with the station and the stream to the north, then the Euclidean model based prediction seems much more acceptable. This is because the higher nitrate level one would expect would be seen in both streams, though not for very much of the more northern stream due to the low range parameter of the Euclidean covariogram. In this case, one argument against the Euclidean model is that the value at that station is also affecting the stream to the south-east which may not border the field that has been fertilised. The alternative scenario is that the higher nitrate level was being caused by a sewage plant discharging effluent. Under these circumstances it seems unreasonable to allow the values on that stream to affect other unconnected streams and therefore the stream distance model for covariance seems more appropriate.

This is where the mixture model is likely to offer a compromise by allowing a hybrid of the two concepts. It is very interesting to observe how the estimated range parameters seem to work in tandem with the estimated λ in order to create the best possible mixture. To see this, recall that the best λ estimate was $\lambda = 0.5$, an even split between the Euclidean and tail-up covariance structures. Also recall that the estimated range parameters were 6.2km for the Euclidean covariance model and 32.6km for the tail-up stream distance covariance model. This means that while there is an even split between the two, the influence of a monitoring station extends much further for the tail-up model. Area "b" in Figure 4.1(c) demonstrates this, as using the mixture model has not resulted in estimated values halfway between those in Figures 4.1(a) and 4.1(b). Rather, the fitted values seem much more akin to those using the tail-up structure in Figure 4.1(b). This demonstrates that the estimated best λ often changes to compensate for the different estimated covariance parameters, a feature that was noted in the studies conducted in Chapter 3. Therefore it seems likely that if the estimated range parameters for the two different distance metrics had both been equal, then it is possible that the optimum estimate of λ in the mixture model

would be higher, in order to give more weight to the tail-up model. With the parameters as they are, the weight is much more evenly split to reflect the fact that the Euclidean distance covariance has a very small range compared to the tail-up model.

One very interesting difference between the two covariance structures is the impact that the weighting structure in the tail-up model has on the predicted values. Looking at area "a" on both Figures 4.1(a) and 4.1(b), the benefits of the weighting become evident. In the figure using Euclidean based covariances, Figure 4.1(a), all stations in the area are weighted solely based on the Euclidean distance between them and the location at which the nitrate level is to be predicted. It is worth noting that only one of the four stations in this area actually lies on the main body of the River Tweed (the station furthest west that is partially obscured in the figure by another station) with the rest lying very close to the location where a tributary feeds into the Tweed. It is because of the lack of weighting that the only location that is actually on the Tweed is given far less importance in the prediction at the surrounding points in Figure 4.1(a) than perhaps it should. Looking at the influence of the station in the centre of the area (on the tributary that flows from the south) also shows the benefits of including both flow-connectedness and a weight based on the size of the rivers. This tributary is quite large in terms of the ratio between its flow volume and the flow volume of the river it is flowing into, suggesting that it will have a substantial effect on the nitrate levels of the Tweed after the point of confluence. The Euclidean model does not reflect this, with its predictions suggesting that after (and even before - another argument against this covariance structure) the point of confluence the nitrate level will drop quite suddenly and then rise very quickly again. On the other hand, the tail-up model takes account of the fact that the tributary that this station lies on is relatively large and that the western-most point should be given more weight as it lies on the River Tweed. This results in

the main River Tweed having predicted nitrate levels that progress much more smoothly from yellow/green to green to yellow/orange at the next station along the Tweed (outwith area "a"). The nitrate levels at the two northern tributaries are all but ignored as the relative size of those tributaries make them too small for the stream distance model to attach much weight to their observations. This seems a more intuitive way to predict the nitrate levels than that of the Euclidean model.

In this circumstance, it is encouraging to see that the mixture model produces fitted values that are a good compromise between the two extremes of the individual covariance structures. The most influential points in area "a" are still those located on the River Tweed (as with the tail-up structure) but the station located on the tributary in the North-East of the region is allowed to have more of an effect on subsequent estimates. This means that the transition from low to slightly higher values located downstream (green $\rightarrow$ yellow) is quicker than in the tail-up stream distance model, and seems more natural in relation to the predicted values of surrounding streams. Of course, no predictions are available at these locations and so any preference is based on speculation of what is likely to be occurring rather than data. In this context it is good to see that the mixture model appears to give sensible predictions, in addition to having provided the lowest prediction error in Section 3.1.

4.1.2 Assessing the Change in Nitrate on the Tweed

Figure 4.3 shows plots of the estimated yearly average nitrate level, predicted using the mixture covariance model with $\lambda = 0.5$, for a set of years ranging from 1986 to 2006 (the range of available data). This should give some insight into how the nitrate levels may be changing over time at both observed and unobserved

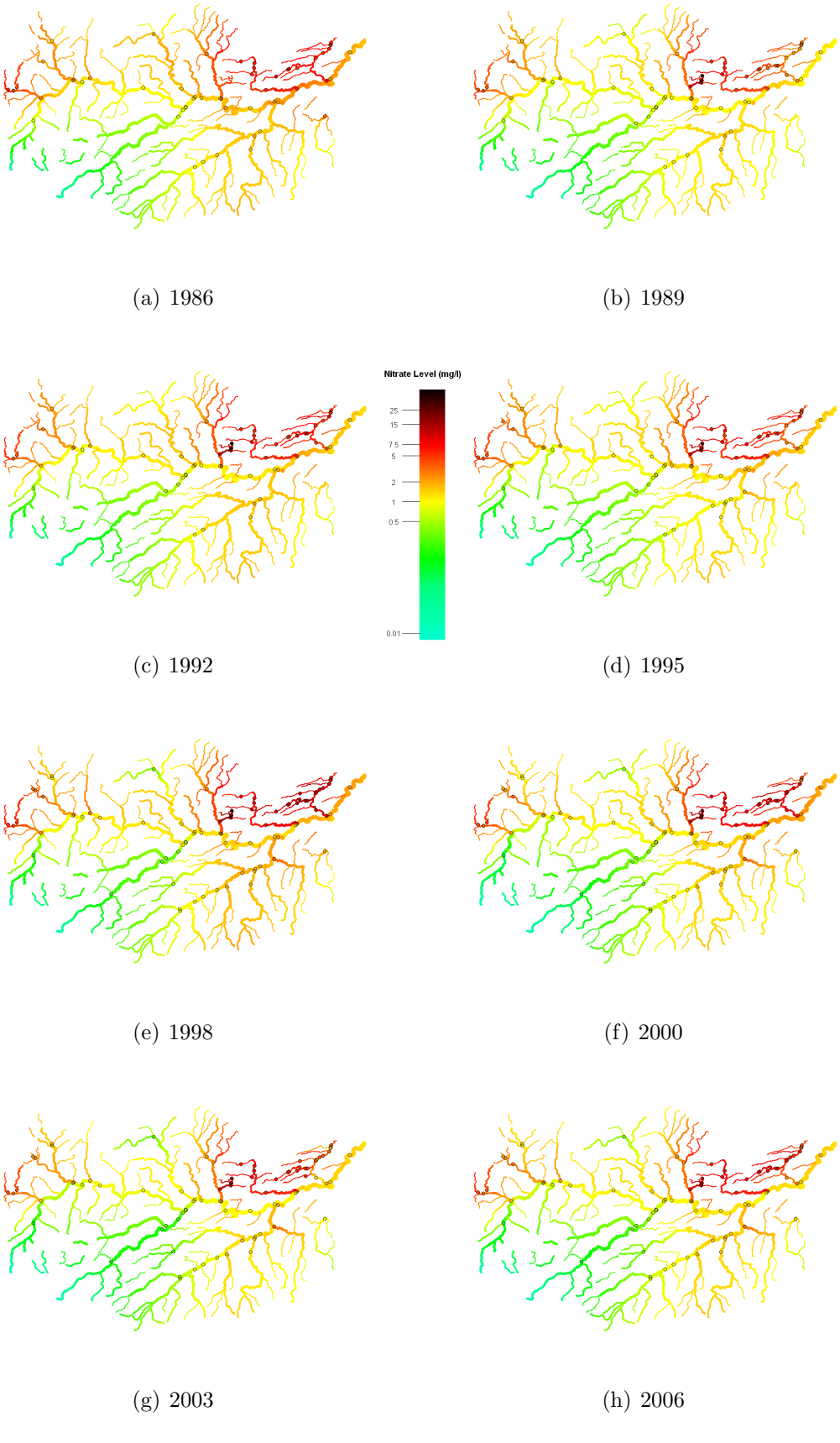
locations. Again, reference will be made to the locations marked in the key shown in figure 4.2.

Changes at Turfford Burn

Looking at the fitted values over time, there do not appear to be any substantial changes over time. This was to be expected given that preliminary analysis (see Section 1.2.1) suggested that any trends at individual stations were very slight. Turfford Burn and the streams surrounding it are shown as area "c" on Figure 4.2. These are the streams with by far the highest nitrate levels throughout the time period, making it one of the areas in which it would be most desirable to see a reduction. There were no observations there (or in the immediate vicinity) in 1986 and so the predicted values in that year are lower than would be expected and are really only being influenced by the fitted trend. Comparing the figures for 1989 and 2006, there are some interesting features to notice at Turfford Burn. In 1989 there are two small tributaries with a reasonably low red colour (with average predicted nitrate values on these tributaries 5.37 mg/l and 5.22mg/l) , one with a very high almost black colour (average predicted nitrate 29.1 mg/l), while the rest of the tributaries in this area, influenced mainly by two stations, are a maroon colour somewhere in between the two (with nitrate values between 11.7 mg/l and 19.8 mg/l). This is the first year of regular monitoring in these locations and so, if 1989 is regarded as the baseline year for this region, this means that all but two of the streams are over the upper limit of 11.3 mg/l as specified by the Nitrates Directive (European Parliament, 1991) as monitoring commences. This means that there were unacceptably high levels of nitrate in this region and possibly explains why monitoring commences in this previously unsampled region.

The Nitrates Directive was introduced in 1991, and so a reduction in nitrate

levels would be expected some time after this. Despite this legislation nitrate levels actually rose in most of the tributaries around Turfford Burn between 1989 and 1998. By 1998, the nitrate levels on the two streams with lowest values had risen to an average of 7.6 and 9.4 mg/l and the worst area from 1989 had fallen slightly to 26.1 mg/l. However the rest of the streams ranged from 16.5 to 26.7 mg/l, the worst affected of which rose by around 7 mg/l in this period. This year is not unusual, as there is a steady rise in values through the 90's. After 2000 this trend seems to reverse, as noted in the exploratory analysis in Section 1.2.1. By 2005, the last year with full January-December data, the estimated nitrate levels have decreased significantly, with all streams having an estimated average between 6.4 and 14.6 mg/l. The two lower nitrate streams now have slightly higher average values than before, but as the worst affected area has dropped by around 12 mg/l this does not give cause for concern. It is worth noting that 1998 has generally higher average nitrate levels than other years, even when compared to 1997 and 1999. Jarvie et al. (2002) note this spike in nitrate levels on the River Tweed around 1998, and attribute it to "high summer HER" where HER is defined to be "rainfall which penetrates the ground after allowing for evapotranspiration and interception losses". Even taking this spike into account, there is still a general pattern of rising average nitrate values throughout the 90s and then declining values from around the turn of the century onwards.



(a) 1986 (b) 1989

(c) 1992 (d) 1995

(e) 1998 (f) 2000

(g) 2003 (h) 2006

Figure 4.3. Predicted values over the entire river network, calculated by kriging using Euclidean detrending and a mixture covariance model with $\lambda = 0.5$

Changes at Other Locations

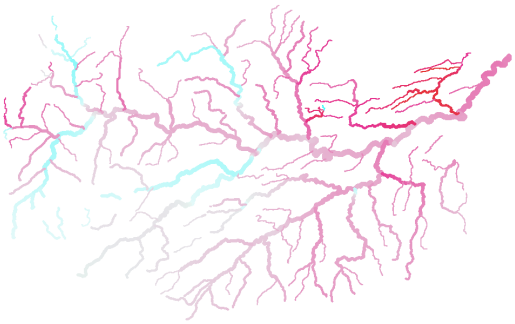
The Leet area, a network to the north of Coldstream consisting of the streams to the South of and including area "d", shows a similar rise before 1998, followed by a decline after 1998, though to a slightly lesser extent. However, there are few other large changes in nitrate in any other locations in the network. Some slight differences can be seen in and around area "a" on the key, but these are as a consequence of the custom palette used to plot the predicted values. The palette was created on the log scale and transformed back to give the nitrate values on their correct scale. Creating this using the logged values means that the differences between lower values are much more apparent than the differences between higher values. A different palette on the original data scale was considered but, when tested, the entire river network was almost completely the same colour, and only the Coldstream/Turfford Burn areas stood out. The custom palette used for plotting here allows the different estimated values to stand out from each other much more. In all years, the vast majority of the network is coloured somewhere between turquoise and orange, meaning that the predicted values are mostly less than 3-4 mg/l. Even a change in colour from green to yellow only signifies an increase in value of around 0.5 mg/l.

The South-West of the region is worth mentioning in more detail. It can be seen that there is very little change in the green colour of the predicted values over time here, and so it is estimated that this region will have very low nitrate values. However there is much less frequent monitoring around here (especially compared to the North-East region) and so the uncertainty here is much greater. Uncertainties will be covered in more detail in Section 4.3, but the lack of observations means that the fairly constant values over time in this region is not surprising as there is less opportunity to pick up small changes in individual smaller streams.

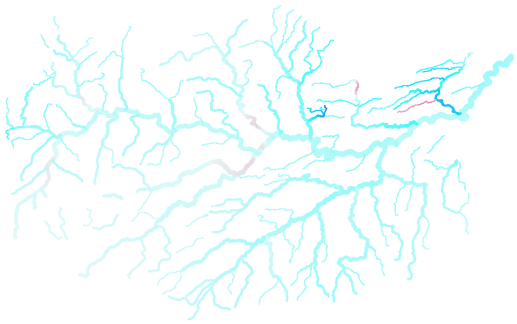
General Changes Over Time

The slight rise in nitrate between the start of the monitoring period, and the late 1990's, then a steady decline until the end of the monitoring period is a feature that is found all around the river network. Figure 4.4 shows the differences in nitrate levels over the years, firstly between 1989 and 1998 (Figure 4.4(a)), then between 1998 and 2005 (Figure 4.4(b)) and finally over the whole time period (figure 4.4(c)). The years 1989 and 2005 were chosen as the start and end points here as prior to 1989 there are several monitoring sites that are yet to be monitored, while the data for 2006 ends in October. In these figures, blue areas represent an improvement/decrease in predicted nitrate levels over the time while red represents poorer/increased nitrate levels and grey areas represent little/no change.

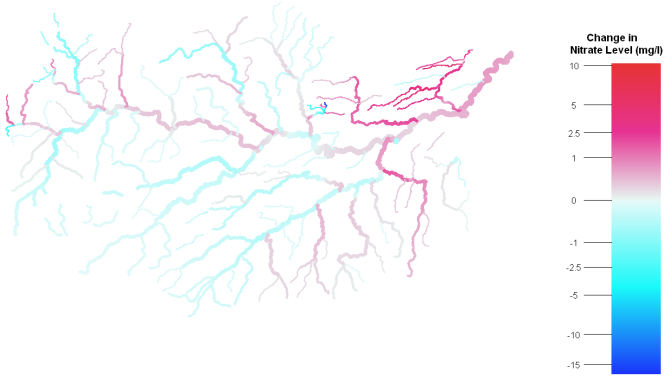
Figure 4.4(a) shows that, in general, the predictions suggest that there was a slight increase in nitrate across the majority of the river network between 1989 and 1998. These increases seem to be at their largest at Turfford Burn (area "c" as discussed before), the Leet and its network of tributaries meeting the Tweed at Coldstream (the network of streams to the South of and including area "d") and the streams located in between these two mini-networks, just North of Kelso. Excluding these locations, all but a handful of the remaining locations showing an increase in nitrate of less than 1 mg/l, with the few remaining locations rising by between 1 and 2 mg/l. Figure 4.4(b) shows how this trend is reversed between 1998 and 2005. All but a handful of locations show a decrease in estimated nitrate levels, with the maximum increase being just 0.76 mg/l. The vast majority of the network shows a decrease of less than 1 mg/l. The areas showing the large rises between 1989 and 1998 have by far the largest decrease in estimated values between 1998 and 2005. It has already been mentioned that 1998 was a particularly poor year for nitrate levels, and so a slightly different picture would



(a) Difference 1998 - 1989



(b) Difference 2005 - 1998



(c) Difference 2005 - 1989

Figure 4.4. Change in predicted average nitrate level over different time intervals

emerge if a different year was used for the middle time point. However, the general picture would remain very similar as 1997-2000 was the peak period for nitrate levels in the dataset and so using 1998 provides a reasonable, and in some locations dramatic, year with which to compare the starting and ending time points.

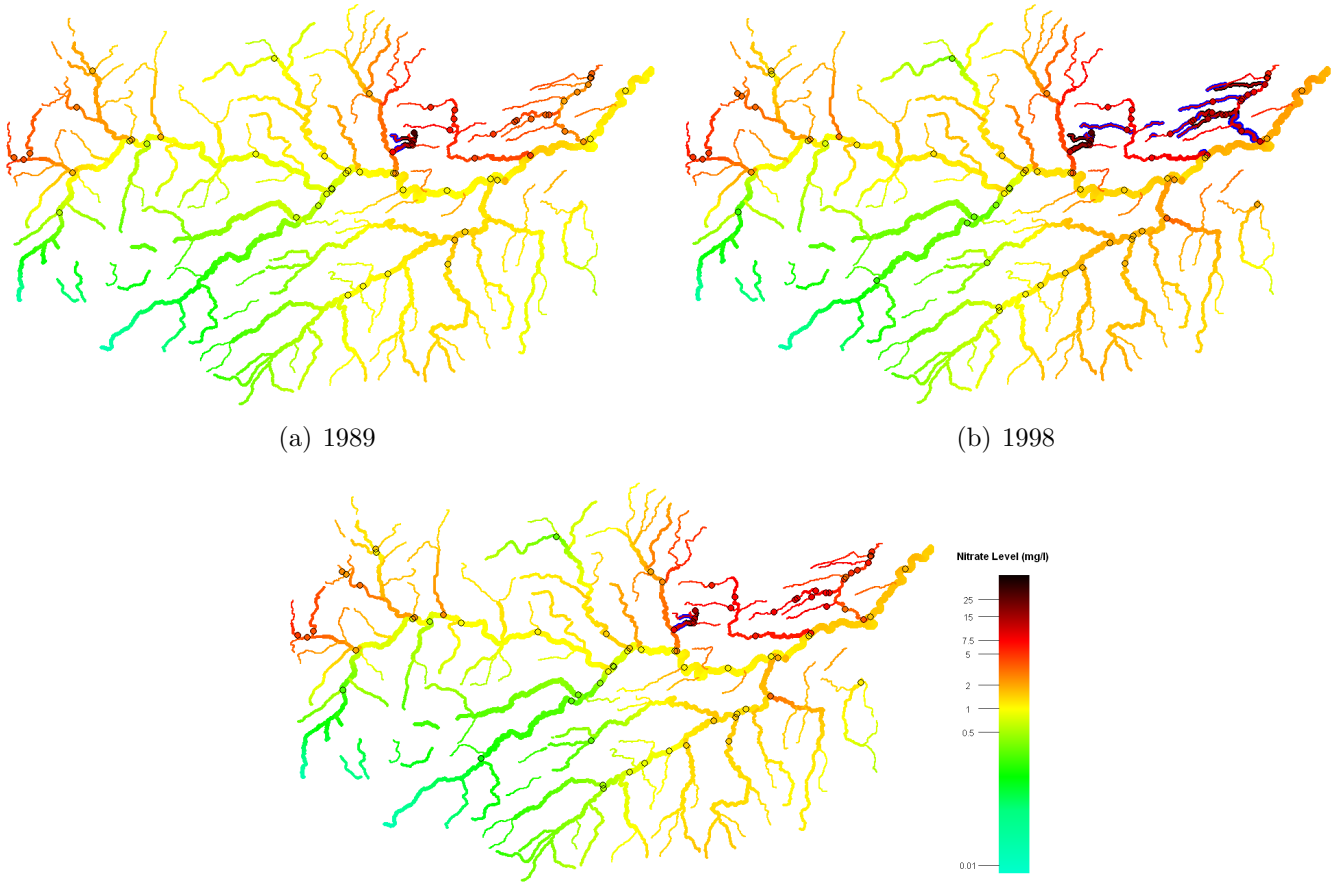
Overall, there is not much of a change between the predicted nitrate levels in 1989 and 2005 (Figure 4.4(c)). There is a rise in nitrate levels predicted in the tributaries to the North East of the region (around Coldstream, as already noted) and some other isolated streams around the region. Most of these rises are less than 2 mg/l although some of the stream segments around Coldstream

are above 3 mg/l. The rest of the network exhibits little change, almost entirely within ± 0.5 mg/l, with the only large decreases being seen in the Turfford Burn area. The problem with looking at just the start and end points of the data like this is that it ignores the fact that vast improvements have been made since 1998, which somewhat 'cancel out' the increase in nitrate levels seen before 1998.

From a legislative point of view, we can also consider the impact that the changing nitrate level has had with respect to the Nitrates Directive. Figure 4.5 shows which areas of the Tweed network are above the upper and lower Nitrates Directive limits at 1989, 1998 and 2005. Those areas above the lower limit (9.04 mg/l) are shown with a blue background, while those above the upper limit (11.3 mg/l) are shown with a black background. In 1989 the only areas above either limit are located in the Turfford Burn area. The central part of this area is worst affected, being above the upper limit, while two more westerly segments are predicted to be above the lower limit. According to the Nitrates Directive which, it should be remembered, did not come into effect until two years later, action should have been taken on all segments above the upper limit and the segments that were above the lower limit if there was evidence of an upward trend. As no data are available anywhere before 1986 it is difficult to tell whether there is evidence of an upward trend before 1989, but given all the analysis so far on the changes between 1989 and 1998 there is evidence of an upward trend after 1989.

Looking at Figure 4.5(b), it can be seen that by 1989 almost the entire Turfford Burn area is above the upper limit, as is the lower stream segment marked as area "d" on the key. Many of the system of tributaries north of Coldstream, and a couple of segments in between Coldstream and Turfford Burn, are also above the lower limit. It seems likely that action would (or at least should) have been taken to reduce the nitrate load at all of these locations, as there is clear evidence of an upward trend. It does seem likely that action was taken, given

that by 2005 (Figure 4.5(c)) the only locations above either limit are at Turfford Burn again. This time, the vast majority of Turfford Burn is exceeding the lower limit, and only a very small segment has predicted values above the upper limit of 11.3 mg/l. This is very encouraging to see as not only has the nitrate level been significantly reduced between 1998 and 2005 but it is very likely that the reduction has been due to the action taken by SEPA to reduce nitrate levels.



(a) 1989

(b) 1998

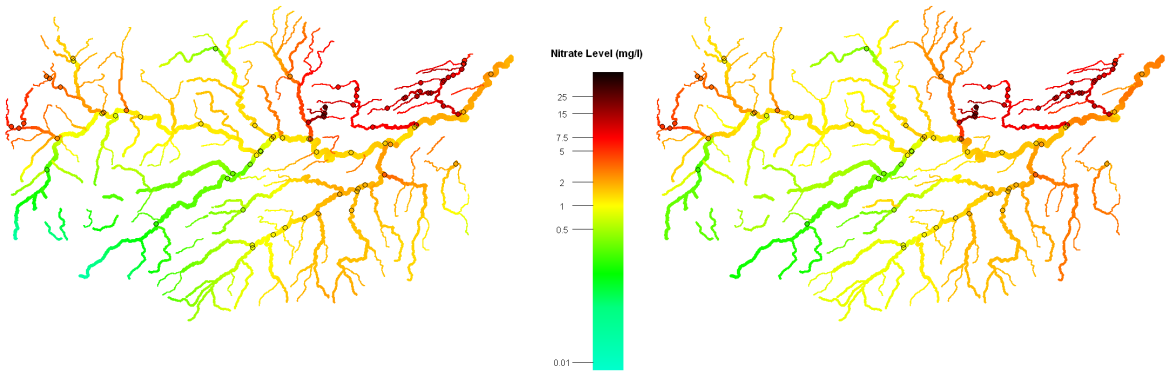
(c) 2005

Figure 4.5. Areas above nitrates directive limits. Blue background denotes exceedence of lower limit, black background the upper

4.2 Stream Distance Based Detrending

4.2.1 Comparison to Euclidean Detrending

Predictions can also be made across the entire river network using the stream distance based detrending method. In previous sections there has been some evidence suggesting that there is little difference between the two detrending methods when used with a mixture of Euclidean and tail-up covariance structures. This seems to be because the lowest RMSPE is found at a λ value that allows more of the covariance structure based on the distance metric not found in the trend– so a stream distance based trend gives a mixture heavily in favour of the Euclidean covariance structure, while the Euclidean trend is weighted slightly more towards the stream distance (tail-up) covariance structure.

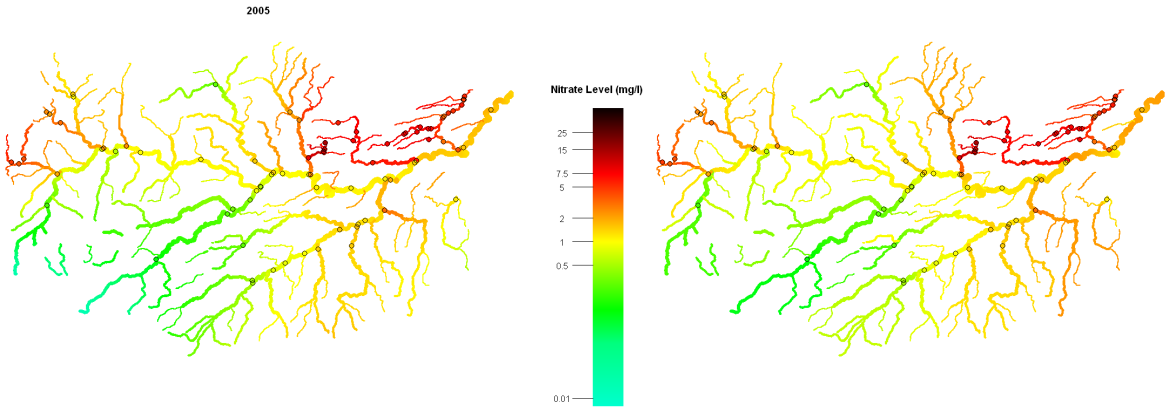


(a) Euclidean detrending (b) Stream distance detrending

Figure 4.6. Comparison of predicted nitrate in 1998 for the different detrending methods

In general, the difference between Figures 4.6 and 4.7 are subtle but very significant in terms of which detrending method is most suitable for use on a river network. Differences are mainly found in areas with more sparse monitoring. For example, in area "b" for both years, the Euclidean detrending predicts slightly lower values in the more northern and eastern streams in this area as the trend

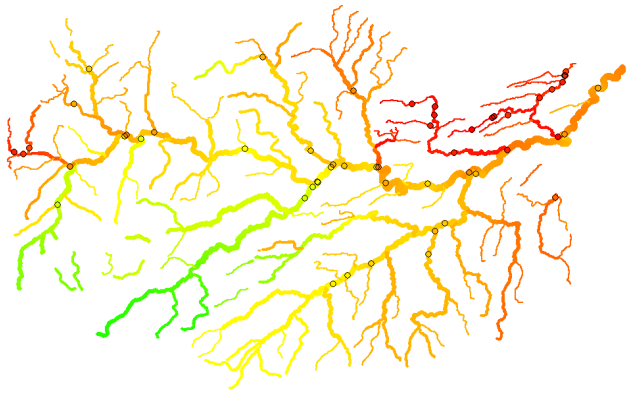
here takes account of the average value at the (flow-unconnected) station to the south. The stream distance detrending does not use this monitoring station at all to calculate the trend at these two streams and thus a higher value is predicted



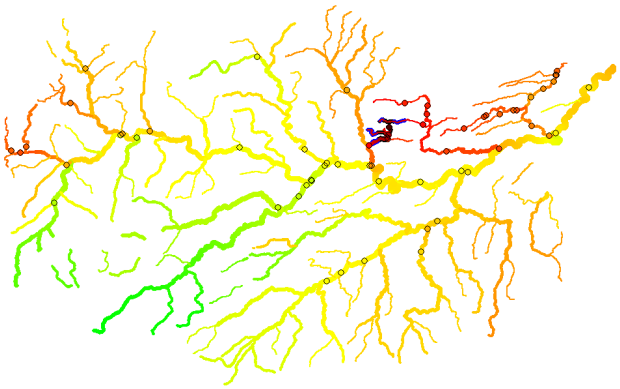
(a) Euclidean detrending (b) Stream distance detrending

Figure 4.7. Comparison of predicted nitrate in 2005 for the different detrending methods

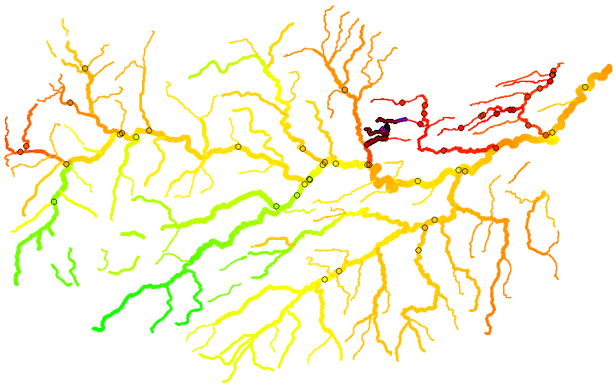
Looking at area "a" and the area of the River Tweed just to the south east, there is a slight change in colour between the two detrending methods indicating that the stream distance detrending is predicting slightly higher values here than the Euclidean. This is the case for many such streams, where there are very few surrounding, flow-connected monitoring stations causing the stream distance model's prediction to tend towards the overall mean, something which is even more likely to happen when the stream in question has a relatively low flow volume. In these circumstances the Euclidean model tends towards a mean of the values in the immediate surrounding area, regardless of whether they are flow-connected.



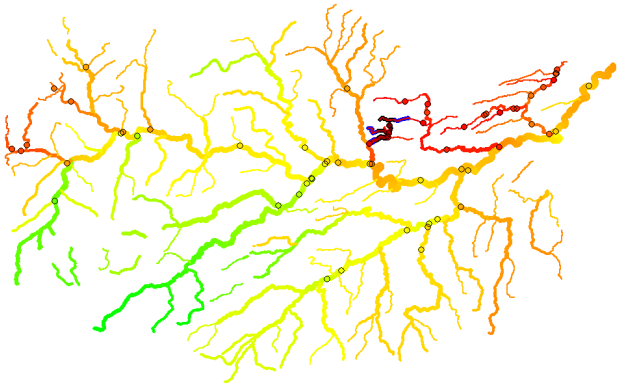
(a) 1986



(b) 1989

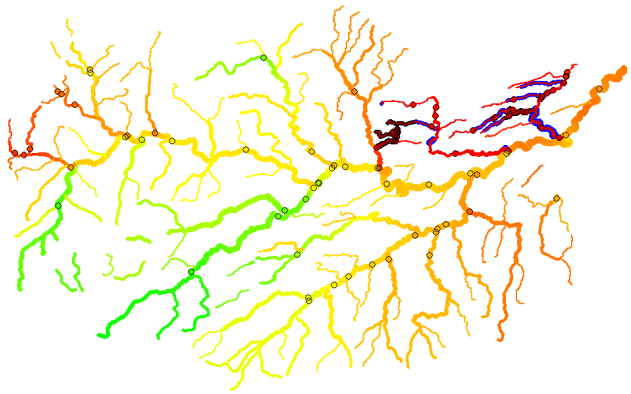


(c) 1992

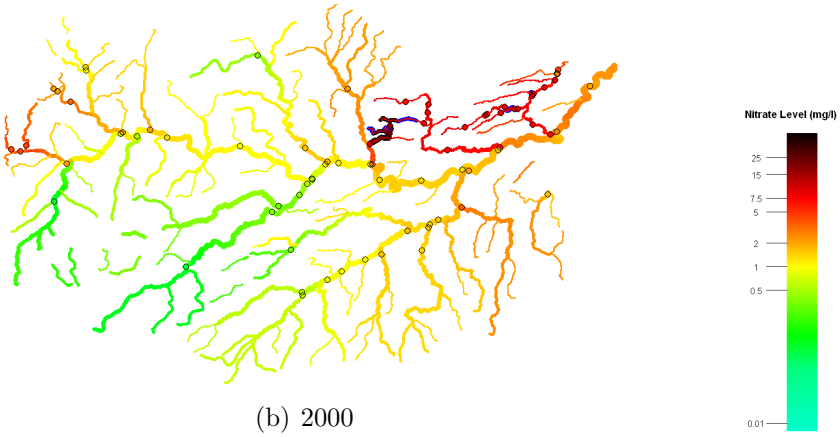


(d) 1995

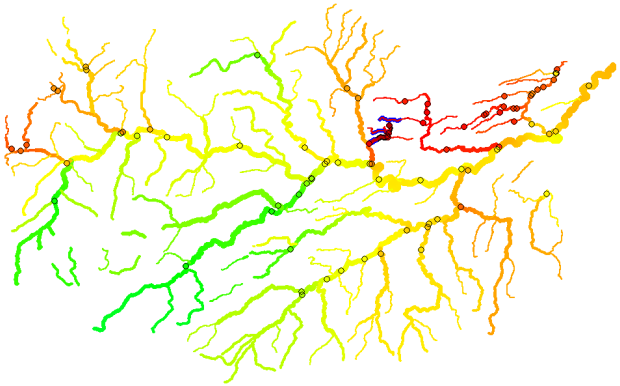
Figure 4.8. Predicted values over the entire river network, calculated by kriging using stream distance detrending and mixture covariance model with $\lambda = 0.9$



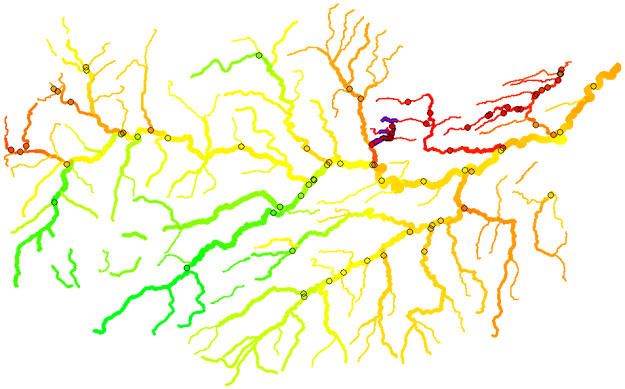
(a) 1998



(b) 2000



(c) 2003



(d) 2006

Figure 4.9. Predicted values over the entire river network, calculated by kriging using stream distance detrending and mixture covariance model with $\lambda = 0.9$

Again, it is important to stress that the differences here are subtle in the context of such a large network but are crucial to deciding the most appropriate detrending metric. The locations where there are larger differences are in areas with few surrounding monitoring stations, meaning that a choice between the two detrending methods should be based on personal belief about which method best represents a river network trend. In general though, Figures 4.6 and 4.7 show that the differences between the detrending methods are minor and, when combined with the optimum mixture models, the predicted values produced by each are largely indistinguishable.

This is the first time that nonparametric trends, or any trend based on stream distance has been considered in the literature. There is little evidence to suggest that a stream distance based trend is preferable to Euclidean, but there is also no evidence to suggest it is a poorer option. This allows a degree of flexibility in approach. Since the choice between the two does not need to be made on the grounds of prediction errors, the distance metric used for trend can be chosen based on suitability for the purpose. Consequently, the choice of trend distance metric can be made on a biological or ecological basis.

4.2.2 Estimated Yearly Average Nitrate Levels

Moving away from comparisons between the detrending methods, the stream distance based detrending with the covariance mixture of $\lambda = 0.9$ can be assessed on its own by looking at the predicted values for the river network over various time points, as shown in Figure 4.8. The maps shown in this figure are supplemented with blue and black outlines, signifying predicted values over the upper and lower limits specified by the Nitrates Directive respectively (as was done for the Euclidean detrending in Figure 4.5).

The predicted values are generally very similar to those obtained using Euclidean detrending and so will not be discussed again. However, this is a chance to see how the areas above the two limits are changing over time. These areas are almost exactly the same for both detrending methods as all lie in areas with quite dense monitoring and thus have very similar predicted values. It can be seen that in 1986 no areas are crossing either threshold due to the fact that monitoring had not begun at the Turfford Burn area at this point in time. This is why 1986 was not used as the baseline year for the comparisons made in the previous section. The predicted values at Turfford Burn in 1986 are mainly composed of the estimated (stream distance based) trend, as there are few connected monitoring stations in the nearby vicinity. As the bandwidth used in the trend was relatively large in order to keep the results quite general, this means that the fitted values here are slightly lower than would be expected if data were present, as several stations are being smoothed to obtain the trend in this area. By 1989, monitoring of this area has begun and most of the streams in this area are predicted to have nitrate levels over either the lower or upper limit, with about half over the lower limit and half over the upper limit. The effect of using Euclidean distance in the covariance structure can be seen by the fact that the unconnected stream to the North is also highlighted as being over the threshold. Moving through the 90's, the affected areas tend to stay the same but an increasing number of them move above the upper limit. 1998 is the worst year included in the study and some streams around Coldstream are slightly over the lower limit. While 1998 generally had more exceedances than a typical year, there are still some of the areas around Coldstream that cross the lower limit in 1997 and 1999 too. By 2000 the average nitrate levels are back to roughly the same as they were in 1995, with many areas crossing the upper limit but all located in and around the Turfford Burn area. Things improve steadily between then and 2006, by which time the size of the area previously estimated to cross the lower limit has reduced slightly,

and most of the streams previously over the upper limit are now just exceeding the lower limit.

4.3 Uncertainties in Kriging Predictions

So far, the errors associated with the kriging predictions have not been mentioned. There is also error associated with the trend prediction, and it is necessary to combine the two in order to obtain the error for each of the predicted values on the River Tweed. If we assume that the final predicted value at point i on the river is the sum of the estimated trend t_i and the kriging estimate at that location z_i then the variance at point i will be given by (4.1).

$$Var(t_i + z_i) = Var(t_i) + Var(z_i) + 2Cov(t_i, z_i) \tag{4.1}$$

In (4.1), $Var(z_i)$ is given by (1.22) during the kriging process. Remember that the the trend is calculated using (2.4). The predicted trend at a point x can be rewritten as $\hat{m}(x) = P\mathbf{y}$ where P is the first column (or the only column when using the local mean method of smoothing, as is done for the stream distance trend) of the matrix created from $(X^TWX)^{-1}X^TW$. The resulting estimates of the variance of the trend will be spatially correlated. The equation for the variance of the trend is shown in (4.2). In this equation, P represents the smoothing matrix used to create the trend. The notation used is to ensure consistency with the matrices produced in the additive modelling in Chapter 5. The smoothing matrix is adjusted to take account of potential spatial correlation using the covariance matrix Σ. Σ is estimated using a mixture model with $\lambda = 0.1$ when using stream distance detrending and $\lambda = 0.5$ when using Euclidean distance detrending, with exponential models for both elements of the mixture

and covariance parameters as shown in Table 2.1.

$$Var(\hat{m}) = P\hat{\Sigma}P^T \tag{4.2}$$

An estimate of the covariance between the trend and the prediction $Cov(t_i, z_i)$ is required to complete the variance formula in (4.1). This was estimated separately for both detrending methods using the data. This was done by taking the estimates of the trend and predicted value at each location throughout time and estimating the covariance using (4.3). This was also done separately for each year of data but these were similar enough that the overall figures were used. The covariances were estimated to be $Cov_{euc}(t_i, z_i) = 0.0437$ for Euclidean distance detrending and $Cov_{str}(t_i, z_i) = 0.0594$ for stream distance detrending.

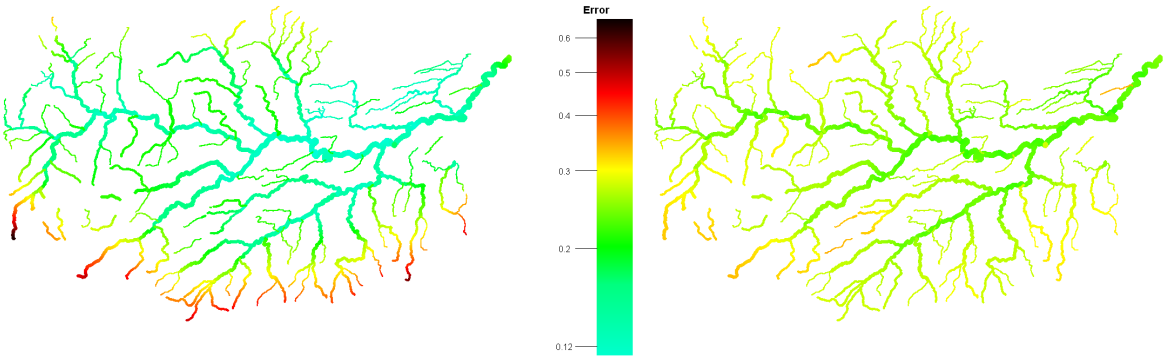
$$Cov(t_i, z_i) = \frac{1}{N}\sum_{i=1}^{N}(t_i - \bar{t})(z_i - \bar{z}) \tag{4.3}$$

4.3.1 Graphical Representation of Error

Figure 4.10 shows the overall estimated standard error $\sqrt{Var(t_i + z_i)}$, comprising the estimated error for trend and prediction added to twice the covariance (4.1), obtained from the Euclidean and stream distance based detrending methods in 1998. It is worth noting here that the errors shown are on the log scale. As the trend surface does not change for each of the 21 years, the component of the overall error coming from the trend does not change. Also, the covariance parameters are fixed over time and after around 1989 there is a reasonably consistent set of observed monitoring stations, and these factors together mean that the component of the overall error coming from the kriging prediction will not change much between the years either. This means that it is only necessary

to look at the errors for one year, as the conclusions are virtually identical no matter which year is chosen.

It may initially seem surprising that the uncertainty seems uniformly low in the middle of the networks, but this is not the case. The uncertainty increases as the predicted location moves further away from the data point, but not enough that it is picked up too much by the colour scale. A very close look shows slight differences in light blue/turquoise/light green points in the middle of the Euclidean detrended network and green/light yellow points in the stream distance detrended version. It is only towards the extreme edges of the network, where the distance to the nearest observed point gets very large, where there is a big increase in uncertainty.

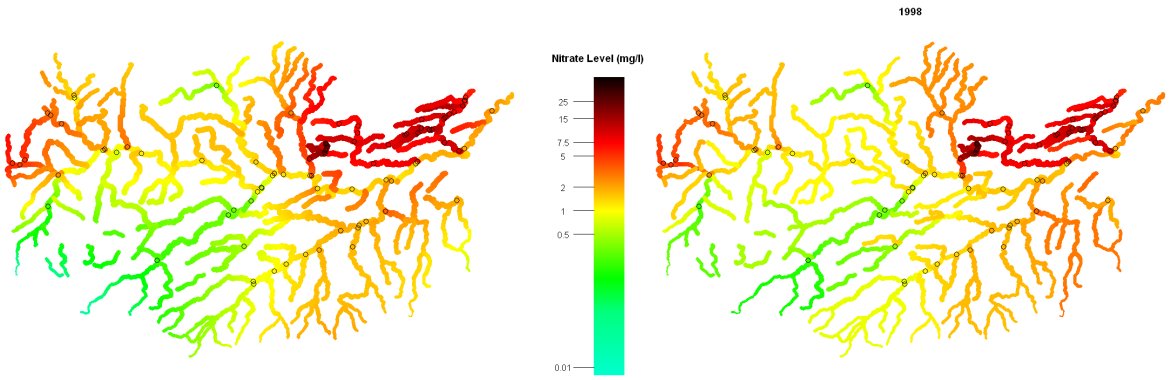


(a) Euclidean detrending (b) Stream distance detrending

Figure 4.10. Comparison of kriging error in 1998 for different detrending methods

Literature such as Cressie et al. (2006) and Ver Hoef et al. (2006) has tried to present the errors alongside the predicted values. Cressie et al. (2006) achieved this by varying the colour of the points according to the predicted value while varying the size of the point according to the 'inverse' of the error. This meant that the points with the largest error were represented as the smallest points, while those with the smallest error appear larger. Figure 4.11 shows something

similar applied to the River Tweed predictions. In these plots, the size of the river is no longer proportional to the flow on that stretch of river and is now inversely proportional to the error. The errors are exactly the same as those shown in Figure 4.10, however the colour scale used in that figure was on the log scale in order to show more of a difference between areas of the network. This is not the case in Figure 4.11. The scale here is changing in a 'linear' manner, so that the difference in the size of points with errors 0.2 and 0.3 is equal to the difference in the size of points with errors 0.3 and 0.4. This means that, for both detrending methods, there is not much variation in size across the plot. This is consistent with findings in the literature, such as in Cressie et al. (2006), for example, who concluded that the range of standard errors in their river network was not large, and the larger errors tended only to occur near the edges of the map.



(a) Euclidean detrending (b) Stream distance detrending

Figure 4.11. Combining both predicted value and prediction error on the same plot, 1998 data

4.3.2 The Relationship between the Error and the Covariance Parameters

Comparison between the errors produced by the two different detrending methods is possible but seems of limited value given how dependent the errors are

on the estimated covariance parameters. For instance, from Figure 4.10 it seems that the central (densely sampled) area of the network has lower error when using Euclidean detrending; however the extreme edges of the network have much higher errors. This is not necessarily a feature of the detrending methods, instead it reflects the role that the covariance parameters estimated for kriging take in the estimated error for the predictions. In Section 3.1, cross-validation showed that the root mean squared prediction error when using Euclidean distance detrending along with a mixture model with $\lambda = 0.5$ was 0.345. Using the same mixture and covariance parameters, the average kriging error for the entire network is 0.226. For the stream distance detrending the RMSPE was 0.350 while the average kriging error was 0.273. The differences between these numbers suggest that the kriging error is underestimating.

The nugget and sill parameters estimated from the variogram or covariogram have the most impact on the predicted error. Poor estimation of these parameters will lead to under or over estimating the errors, in the same way that altering the range parameter was shown to produce better RMSPEs in Section 3.2. One alternative would be to adapt a cross-validation approach to generate errors in Section 3.1 so that a different RMSPE could be calculated for every monitoring station and a more accurate representation of the potential error could be constructed. Obviously this could not be done for unobserved locations, but it may be possible to combine the results with a simulation study, similar to one of those performed in chapter 3 to work out errors relative to one another.

4.4 Conclusions on the River Tweed Predicted Nitrate Levels

This chapter has examined the yearly average nitrate values predicted at unsampled locations using kriging with a variety of detrending methods and covariance structures.

For the first time in the literature, a trend surface based on stream distance has been considered. There were only small differences between the predicted values generated using Euclidean and stream distance based detrending with their optimum mixture model (as determined in Section 3.1). The locations where there were anything more than minor differences between the detrending methods tended to be in locations on the periphery of the network with few nearby flow-connected monitoring stations. This provides more evidence that the distance metric used in the detrending process is balanced out by the use of a covariance mixture model and consequently that the choice of detrending metric is not crucial if the optimum mixing parameter has been estimated sensibly. Consequently, the choice of distance metric for the trend estimate can be made on the basis of whichever is thought to best represent the biological or ecological processes that are likely to be driving the trend.

The estimated values themselves give insight into how the yearly average nitrate level is changing between 1986 and 2006. In general, nitrate levels tend to exhibit a slight rise between 1986 and 1998, before decreasing slightly after that. The period 1997-1999, and in particular 1998, has been noted in previous chapters and other literature Jarvie et al. (2002) as having quite high nitrate levels due to unusual environmental conditions, and this is confirmed by the estimated averages in these years. The Turfford Burn area has, by some way, the highest levels of nitrate on the entire River Tweed network, especially in 1998 where

almost all streams had estimated nitrate levels over the Nitrates Directive upper threshold. However, these areas subsequently experienced a substantial decrease in yearly average nitrate level so that by 2005 most are only in exceedance of the lower, rather than upper threshold.

This chapter has also introduced a novel way to visualise the estimated values. Previous literature has generally estimated at a very small subset of locations and plotted the predicted values as isolated points on the larger network. The plotting style introduced here estimates at more locations on the network in order to provide a more coherent picture of the network as a whole. This aids understanding of the estimated values and the tail-up model as a whole, as the features of this covariance structure, such as the sharp changes at points of confluence are more easily identified.

Chapter 5

Additive Modelling in Space and Time

The investigation into modelling of nitrate levels on river networks has thus far looked at individual spatial snapshots in time, and has not considered the effects of space and time. There is a growing number of examples of the use of tail-up models in the modelling of river network data in space, but examples of modelling through space and time using this structure are still rare in the literature.

Examples of spatio-temporal models can be seen in areas such as air pollution (Shaddick and Wakefield (2002), Guttorp et al. (1994) etc.), rainfall (Brown et al., 2002), snow water (Huang and Cressie, 1996), variation in fish in a lake (Reyjol et al., 2005) and many more. There are also many examples of river data being modelled through space and time, for example Cressie and Majure (1997), Clement et al. (2006), Akita et al. (2007) and Thorp et al. (2006), however these examples do not consider the tail-up model when modelling spatial correlation. There are only a very small number of papers that consider the tail-up structure

for the modelling of river network data through space and time.

Money et al. (2009); Gardner and McGlynn (2009) are the only examples found in the literature of the tail-up structure being used in any sort of space-time analysis. Money et al. (2009) consider the use of the tail-up structure in their space-time models, but express concerns that their data set has very few pairs of flow-connected stations, which could make the flow-connectedness of the tail-up structure a hindrance. They go on to fit space/time models using stream distance but not in the form of the tail-up model. They use a slight generalisation of the tail-up model for the spatial element of the spatio-temporal model. This model was then used in a cross-validated study and was shown to reduce mean squared error by around 10% when compared to a model using just Euclidean distance for the spatial element. Gardner and McGlynn (2009) do use the tail-up model but, their analysis does not account for temporal correlation. They fit parametric models to nitrates data from the Rocky Mountains using the tail-up model to account for the spatial correlation. They include terms to allow for seasonality in the data but the analysis assumes that the residuals from this model are temporally uncorrelated. Gardner and McGlynn (2009) provides the closest approximation to what would like to be achieved with the River Tweed data, but it would seem more desirable to be able to model this using nonparametric techniques, given the highly complex nature of environmental data. It would also be desirable to adapt the tail-up model to produce a trend surface as part of such a model, rather than just using it to define the residual spatial correlation.

Despite the lack of literature exploring space-time models that incorporate the tail-up model structure, it is the obvious extension of the analysis that has been carried out so far. All of the environmental analysis in chapter 4 comes from yearly averages, but there is clearly much information being lost by aggregating to yearly data. Therefore the next step of analysis will be to model the

Tweed data through space and time, and it was decided to do this using additive models. Additive modelling is a nonparametric method that is very well suited to the modelling of environmental data and have been used many times for such purposes; see Bowman and Azzalini (1997); Ferguson (2007).

The aim of this Chapter is to use additive modelling techniques to model the nitrate levels on the Tweed through space and time. Part of this process will involve adapting the tail-up model to produce a suitable smoothing function for the spatial trend, so that existing additive modelling techniques can be applied to the river network setting for the first time.

5.1 Additive Models

Additive modelling is a nonparametric form of regression. In the environmental context, a nonparametric approach allows for greater flexibility in the slopes for the trends and seasonality that are fitted to the data. Additive models do not rely on assumptions of linearity in the way that generalised linear models do and so if an assumption of linearity does not seem reasonable, an additive model can estimate terms as nonlinear functions. Section 2.1 discussed at great length the reasons for choosing to model the spatial trend using nonparametric smoothing, and an additive model is a natural extension of the functions used to generate the spatial trends there.

$$Y_i = \mu + \sum_{j=1}^{k} m_j(x_{ij}) + \varepsilon_i \qquad \begin{matrix} i = 1, ..., N \\ j = 1, ..., k \end{matrix} \tag{5.1}$$

Additive models take the general form shown in (5.1), where the m_j are

smooth functions estimated nonparametrically, with one for each of the k predictor variables in the model (Hastie and Tibshirani, 1990). The m_j are constrained such that $\sum_{i=1}^{N} m_j(x_{ij}) = 0$ for $j = 1, ..., k$. The errors ε are assumed to be correlated with mean 0 and variance $\Sigma\sigma^2$, where Σ is the correlation matrix. The model is fitted by first obtaining smooth functions $\hat{m}_j(x_{ij}) = S_j y$ for each of the model components, before using the backfitting algorithm, as detailed in Hastie and Tibshirani (1990), to smooth between the components of the model in order to obtain the final estimates for each component. One of the possible additive models that will be used for the River Tweed data is shown in (5.2). This model uses three different smooth functions, m_1, m_2 and m_3 to model the spatial trend, the trend across the years and the trend within each year respectively. This model will be used to illustrate the principles of additive modelling.

$$log(nitrate) = \mu + m_1(spatial) + m_2(year) + m_3(day) + \varepsilon \qquad (5.2)$$

5.1.1 Smooth Functions

A novel method will be developed to enable spatial smoothing using stream distance. This will be formulated so that it incorporates the flow-connectivity and flow weighting developed in the tail-up covariance model.

There are several methods that can be used for the smoothing function m_j, such as smoothing splines and kernel methods, and examples of these can be found in Hastie and Tibshirani (1990) and Bowman and Azzalini (1997). Two kernel methods, local mean and local linear with normal kernel functions, were used in the fitting of spatial trends in Section 2.1 and so these methods will be used again in this chapter. The fitted values generated from local mean or local linear smooth functions are unlikely to be much different to those generated by

other methods such as splines, and so these methods were used in analysis mainly due to their use in earlier work. However, other methods such as splines have advantages, and these will be mentioned in Section 6.3. Local smoothing has the advantage of depending only on distance measures, while spline bases require more complex modifications to adapt to the river network setting. In all, three different smoothing functions will be used: a local mean to produce the spatial smooth, a local linear to smooth the spatial trend across the years and a circular smoother for the seasonal effect across each year. These will be illustrated using the three elements of (5.2).

Spatial Smoothing Function

Spatial smoothing, which uses the local mean method, is similar to the spatial smoothing function that was created in Section 2.1.1, but this time is combined with an indicator function, $\delta_i(x)$, based on flow weighting. The spatial smooth function $m_1(spatial)$ in (5.2) is taken as the least squares estimator $\hat{\alpha}$ which arises from (5.3), where w is a weight function based on the normal kernel density with bandwidth h, $\delta_i(x)$ is a weighted indicator function based on the tail-up correlation model and $x_i - x$ is the stream distance between point x and point x_i. The weight function used to produce the least squares estimator (5.4) is just a normal kernel density function but the indicator function $\delta_i(x)$, shown in (5.5), is a novel way of using the flow weighting and flow-connectedness proposed in the tail-up model in a nonparametric setting. In this equation, B_{x,x_i} is the set of all streams between stations x and x_i including the stream upon which the furthest upstream of x and x_i lies. The ω_k are based on the estimated flow volumes in each stream, as they were in Section 1.3.2, but again could be based on stream order or some other surrogate. This means that the weighting in the local mean is

based on both flow connectedness and the relative size of streams on the network.

$$\min_{\alpha} \sum_{i=1}^{n} \{y_i - \alpha\}^2 w(x_i - x; h)\delta_i(x) \tag{5.3}$$

$$w(x_i - x; h) = exp(-\frac{1}{2}\frac{(x_i - x)^2}{h^2}) \tag{5.4}$$

$$\delta_i(x) = \begin{cases} \Pi_{n \in B_{x,x_i}} \sqrt{\omega_k} & \text{if } x \text{ and } x_i \text{ are flow connected} \\ 0 & \text{if } x \text{ and } x_i \text{ are not flow connected} \end{cases} \tag{5.5}$$

Equation (5.3) can be solved by expressing the model in the form $Y = X\theta + \varepsilon$ and solving for $\hat{\theta}$, i.e. $\hat{\theta} = (X^T W X)^{-1} X^T W \mathbf{y}$. In this form, X denotes a vector of 1's, while W is a diagonal matrix with elements $w(x_i - x; h)\delta_i(x)$ on its diagonal.

Temporal Trend Smoothing Function

For the non-seasonal component of the temporal trend, denoted $m_2(year)$ in (5.2), a local linear kernel smooth will be used. A local linear would be preferable to local mean for the spatial component but when used with the flow connected information requires at least two monitoring stations within two bandwidths distance of the location at which the prediction is to be made. This did not allow prediction across the entire network and so a local mean was used instead, but the local linear approach can be used to smooth the long term temporal trend.

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \{y_i - \alpha - \beta(t_i - t)\}^2 w(t_i - t; h) \tag{5.6}$$

$$w(t_i - t; h) = exp(-\frac{1}{2}\frac{(t_i - t)^2}{h^2}) \tag{5.7}$$

For the local linear smoothing method, the smooth function $m_2(year)$ for some year t is taken as the least squares estimator $\hat{\alpha}$ arising from (5.6). Converting this to vector-matrix form as before allows this to be estimated using $\hat{\theta} = (T^T W T)^{-1} T^T W \mathbf{y}$. This time T denotes a matrix with a column of 1's and a column with elements $(t_i - t)$ and matrix W is a diagonal matrix with elements $w(t_i - t; h)\delta_i(t)$ on it's diagonal.

Seasonal Smoothing Function

A third smoothing function is required to fit the seasonal component of the proposed additive model, $m_3(day)$, though it is really just a variation on the local mean estimator. This term quite clearly behaves in a 'cyclical' manner, with day 365 of year 1 being directly followed by day 1 of year 2 and so circular smoothing is needed for this component. The circular smoothing function can be obtained using the local mean smoothing shown in (5.8) along with a Von Mises weight function, as shown in (5.9). In this equation, $d_i - d$ represents the number of days between d_i and d and $r = 365$. This weight function produces an estimate on a cyclic scale, and has been used in the context of additive modelling for other analyses of environmental data; see for example Ferguson (2007). Alternative approaches are available; for example Loader (1999) details a circular weight function that uses a periodic distance function to achieve similar results. In the Tweed data, there was no observations collected on the 29th February during leap years. However, if predictions were required on this date then the equation

allows for this in the same way that any other date could be calculated.

$$\min_{\alpha} \sum_{i=1}^{n} \{y_i - \alpha\}^2 w(d_i - d; h) \tag{5.8}$$

$$w(d_i - d; h) = exp(\frac{1}{h} cos(2\pi \frac{d_i - d}{r})) \tag{5.9}$$

5.1.2 The Backfitting Algorithm

The backfitting algorithm is used to fit additive models by iteratively smoothing with respect to each covariate, taking the residuals based on the other components of the model as the response each time (Hastie and Tibshirani, 1990). The mean of the response is subtracted prior to fitting the model and additional adjustment is made so that $\sum_{i=1}^{n} \hat{m}_i(x_{ij}) = 0$.

$$\hat{m}_j^l = S_j(y - \textstyle\sum_{q<j} \hat{m}_q^l - \sum_{q>j} \hat{m}_q^{l-1}), \quad j = 1, ..., k. \tag{5.10}$$

The backfitting algorithm is shown in (5.10) and shows the smoothing function for component j of the additive model at each iteration l. In this equation, S_j represents the smoothing matrix used in the estimation of each of the components of the model, $m_j(x_j)$ before the backfitting algorithm is used, i.e. $\hat{m}_j(x) = S_j y$. This is multiplied by the 'residuals' from the current iteration of the model, where the current iteration of the model is the sum of the values of the $j-1$ components already smoothed in iteration l, $\sum_{q<j} \hat{m}_q^l$, and the values after the previous ($l-1$th) iteration of the remaining $k-j$ components, $\sum_{q>j} \hat{m}_q^{l-1}$. After convergence, the $\hat{m}_j^l$ from the final iteration are the smoothed values of each of

the k components.

$$P_j^l = (I - P_0)S_j(I - \sum_{q<j} P_q^l - \sum_{q>j} P_q^{l-1}), \quad j = 1, ..., k. \tag{5.11}$$

Giannitrapani et al. (2005) demonstrate that this can be adapted to obtain the projection matrices P_j which create the final estimates for each of the components after convergence. If $\hat{m}_j^l = P_j^l y$ then (5.11) demonstrates how to obtain the P_j. Here, P_0 denotes an $n \times n$ matrix with elements $1/n$, while I is the n^{th} order identity matrix. The projection matrices can then be used to derive the standard errors associated with each component, as shown in (5.12). In this equation, $\sigma^2 = RSS/n - tr(2S_j - S_j S_j^T)$ as detailed in Ferguson (2007).

$$\begin{aligned} Var\{\hat{m}_j(x)\} &= var\{P_j y\} \\ &= P_j var\{y\} P_j^T \\ &\simeq P_j P_j^T \hat{\sigma}^2 \end{aligned} \tag{5.12}$$

5.1.3 Standard Errors and Model Selection

(5.12) shows the equation to derive the standard errors in the fitted additive model when the errors are assumed to be uncorrelated. The residuals from models of the Tweed data are likely to be spatiotemporally correlated, and so adjustments must be made to the formulae. McMullan (2004), Ferguson (2007) and Bowman et al. (2009) detail how this can be achieved, either by first binning the data to reduce dimensionality or by using the correlation matrix Σ.

Giannitrapani et al. (2005) show that the residual sums of squares and approximate degrees of freedom for error, given projection matrices P_j and correlation matrix Σ, can be adjusted to account for residual correlation using (5.13) and (5.14).

$$RSS_j = y^T(I - P_j)^T\Sigma^{-1}(I - P_j)y \tag{5.13}$$

$$\begin{aligned} df_{err_j} &= tr\{(I - P_j)^T\Sigma^{-1}(I - P_j)\Sigma\} \\ &= n - tr(P_j^T + \Sigma^{-1}P_j\Sigma - P_j^T\Sigma^{-1}P_j\Sigma) \end{aligned} \tag{5.14}$$

From these equations, σ^2 can be estimated from RSS/df_{err}, and the standard errors for each of the components m_j are are taken to be the square root of the diagonal entries of $cov\{\hat{m}_j\} \simeq P_j\Sigma P_j^T\hat{\sigma}^2$. df_{err} can be calculated in the same way as the df_{err_j} in (5.14) by using the overall projection matrix, $P = \sum_i P_j$ in place of the P_j.

It is also possible to assess nested models using the residual sums of squares and degrees of freedom produced by these formulae by means of an approximate F-test (Hastie and Tibshirani, 1990). The F-statistic (5.15) is compared to an F-distribution with $(df_2 - df_1)$ and df_1 degrees of freedom. Here, model 1 (and thus RSS_1 and df_1) correspond to the full model while model 2 (RSS_2 and df_2) denotes the reduced model. The F-test tests the hypothesis of 'no effect versus effect' between model 1 and model 2, and so an F-statistic greater than the critical

value from the F-distribution favours the full model over the reduced model.

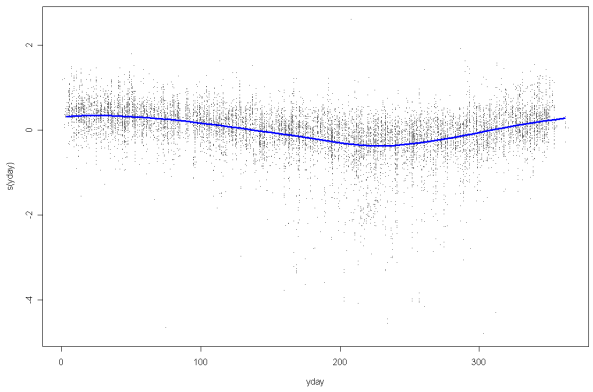
$$F = \frac{(RSS_2 - RSS_1)/(df_2 - df_1)}{RSS_1/df_1} \tag{5.15}$$

5.2 Additive Modelling on the River Tweed

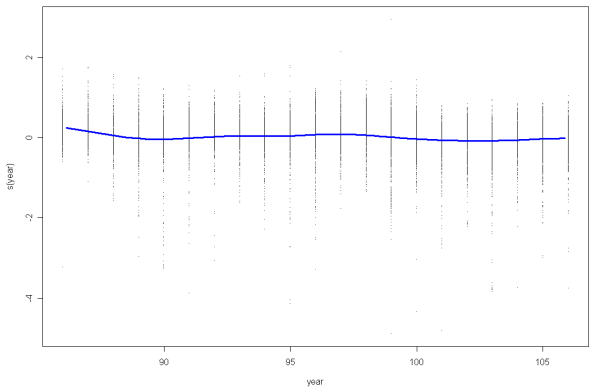
The first model that will be considered for the Tweed data will be the model shown in (5.2), and restated below in (5.16), containing a spatial term, a seasonal term and a term for the overall trend. Before fitting this model, it is required to set the bandwidth for the smoothing functions used with each of the components of the model. A bandwidth of 15km was used for the spatial component, in keeping with the bandwidth used to form the trend in the spatial analysis carried out in previous chapters. For the seasonal component, the bandwidth equated to around two months (60 days), while the bandwidth for the long term temporal trend was set at 2 years.

$$log(nitrate) = \mu + m_1(spatial) + m_2(year) + m_3(day) + \varepsilon \tag{5.16}$$

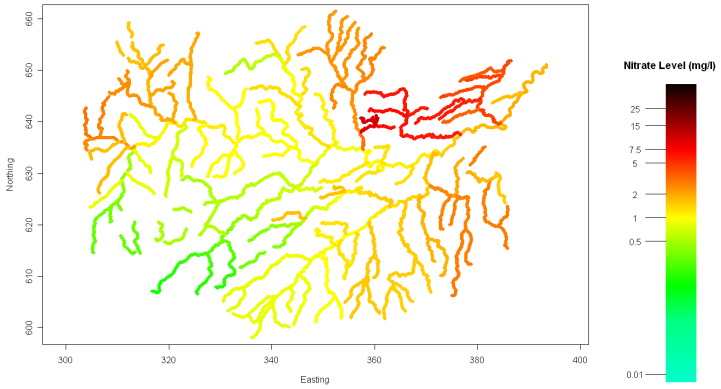
Figure 5.1 shows plots of the seasonal, temporal and spatial trends in the logged nitrate value fitted in model (5.2). The seasonal component peaks in value around the end of January and is at its lowest in July. As would be expected given the analysis in previous chapters, the temporal trend is not particularly strong, dropping very slightly between 1986 and 1989, increasing slightly until 1998 before decreasing slightly thereafter. The fitted spatial trend will not be discussed further, as it is very similar to the stream distance based trends discussed in previous sections.



(a) Seasonal component



(b) Overall trend

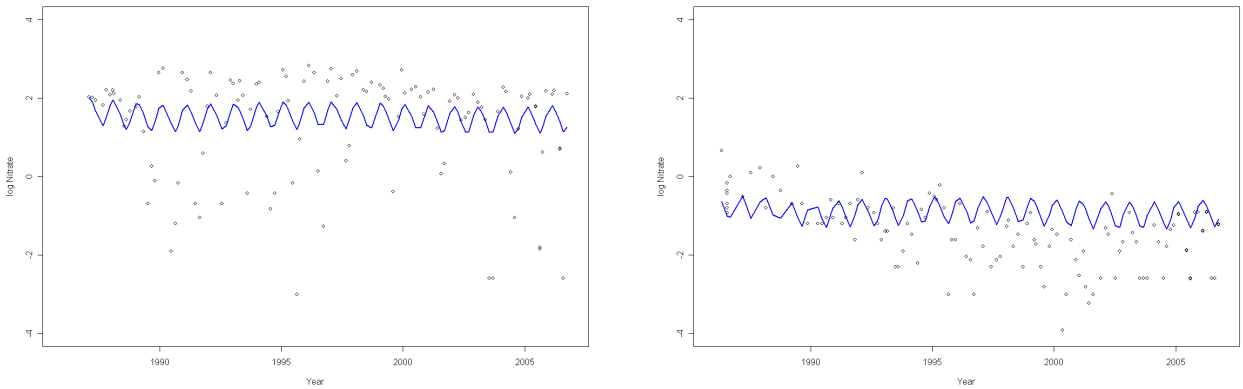


(c) Spatial trend

Figure 5.1. Plots of fitted spatial, seasonal and temporal trend effects in additive model(5.2)

This model is one of the simpler models that could have been fitted to the Tweed data. It can be seen to be a relatively poor fit to the data as it fits the same seasonal component and temporal trend to every location on the river, adjusted up or down slightly by the spatial pattern. This can be seen from the models fitted to the data at individual stations, two of which are shown in Figure 5.2.

As there are very few stations to exhibit anything more than a very slight trend, the model provides a reasonable fit to quite a lot of stations. The only exceptions to this are the areas that exhibit a strong seasonal component (several of which are located in and around the Turfford Burn area) and areas that show



(a) Charterpath bridge (b) Kingledores

Figure 5.2. Fitted values for main effects model (5.2)

a strong trend (Kingledores). So, in general, the fitted values from the additive model are reasonable but fail to pick up on complexities at a handful of stations. The two plots shown in Figure 5.2 show instances where the seasonal component and trend do not seem to be adequately accounted for by the very general fitted model. Figure 5.2(a) seems to have quite a strong seasonal component with a large amplitude, while Figure 5.2(b) clearly shows a downward trend that is not captured by the fitted model. It is clear from these plots that it would be a good idea to investigate a more complex model.

5.2.1 Building A More Complex Model

Model (5.16) is an obvious starting point in the additive modelling process due to its simplicity, but this simplicity makes it a poor fit to these data. Ideally, a model for these data would have an interaction between the spatial location and both the seasonal and temporal trends, allowing for a different seasonal pattern and overall trend at each station but smoothed over the spatial surface to allow for extrapolation to other locations. This is similar to the approach adopted by Bowman et al. (2009), and this is the first time such an approach has been applied

to river network data that uses the type of weighting proposed by Ver Hoef et al. (2006).

The model that is to be fitted is shown in (5.17). This model contains the same main effects as used in (5.2), but adds two interaction terms to account for the interactions between space and trend and space and season respectively. The interactions in (5.17) are constrained to sum to zero. Due to the added complexity of the interactions, the local mean method of smoothing (5.3) must be used. Equations (5.18) and (5.19) show the weightings used in the smoothing of each of the terms in the model. In each of these formulae, $x_{s_i} - x_s$ represents the spatial stream distance or lag (in kilometers in the Tweed example), while $x_{t_{1i}} - x_{t_1}$ represents the temporal lag in years (for (5.18) and $x_{t_{2i}} - x_{t_2}$ represents the temporal lag in days (for 5.19)), while h_s, h_{t_1} and h_{t_2} represent the spatial and trend term and seasonal bandwidths or smoothing parameters respectively. In (5.19), r_{t_2} is set at 365 to ensure that the seasonal component has a period of one year. It is worth remembering that the local mean smoothing in (5.3) includes the indicator term $\delta_i(x)$ that defines the connectivity of the network as well as containing additional flow-based weighting information. As with the simpler model, the errors ε are assumed to be correlated with mean 0 and variance $\Sigma\sigma^2$, where Σ is the correlation matrix. Section 5.4 will describe how the correlation matrix was estimated for each of the fitted additive models. It is not necessary to use the same bandwidths for the interaction terms as were used in the main effects, and different values could be chosen if desired. However, it does seem reasonable that if the main effect is influential over that time/distance then the interaction term would be too.

$$log(nitrate) = \begin{array}{c} \mu + m_1(spatial) + m_2(year) + m_3(day) + \\ m_4(spatial, year) + m_5(spatial, day) + \varepsilon \end{array} \tag{5.17}$$

$$w((x_{s_i} - x_s), ((x_{t_{1i}} - x_{t_1}); h) = exp(-\frac{1}{2}\frac{(x_{s_i} - x_s)^2}{h_s^2})exp(-\frac{1}{2}\frac{(x_{t_{1i}} - x_{t_1})^2}{h_{t_1}^2}) \quad (5.18)$$

$$w((x_{s_i} - x_s), ((x_{t_{2i}} - x_{t_2}); h) = exp(-\frac{1}{2}\frac{(x_{s_i} - x_s)^2}{h_s^2})exp(\frac{1}{h_{t_2}}cos(2\pi\frac{x_{t_{2i}} - x_{t_2}}{r_{t_2}})) \quad (5.19)$$

The choice of each of the three bandwidths must be carefully considered. For the spatial bandwidth, h_s, a bandwidth of 15km was used, in keeping with the spatial smoothing analysis in previous sections. For the trend bandwidth, h_{t_1}, it was decided that a bandwidth of 2 years would be sufficient to smooth without following the data too closely. This means that data 4 years either side of any particular point, and thus four years data in total, are used to calculate the trend. This seems appropriate given the 21 year range of available data. The seasonal bandwidth, h_{t_2} was taken to be 0.17, the equivalent of around 60 days and again in keeping with that used for the simpler additive model in model (5.2).

5.3 Assessing the Interaction Models

The novel, stream distance based interaction model defined in (5.17) will be used to model the River Tweed data. In addition to this, another two reduced models will be used, each containing one of the two interaction terms in addition to the main effects. Including the main effects model, this means that four models were used to describe the Tweed data, and summaries of all four are shown in Table 5.1.

The weighting structure used in the novel spatial component, present as both a main and interaction effect, made it impossible to perform the backfitting algorithm using existing additive modelling software. The size of the dataset (around $11,000$ observations) was initially thought to be prohibitive to performing backfitting using the matrix formulation shown in (5.10). A possible alternative approach that was considered would be to reduce the data to monthly observations at each station, creating a 250 (months) by 83 (stations) grid of data, as in Bowman et al. (2009). This method was ultimately rejected as in the monthly gridded form, there was no available data for around 50% of the station/month combinations. It was decided that the easiest way to carry out the analysis would be to use the data in its original form, though this required working with and manipulating 11000×11000 matrices. Computational efficiencies and sensible memory management were needed throughout in order to carry out backfitting and to subsequently obtain the residual sums of squares and degrees of freedom required for model testing. Sharing a 3GHz processor with 32Gb RAM (more if virtual memory is included), the backfitting algorithm converged in just under 24 hours for each model, while obtaining summary statistics (such as degrees of freedom and residual sums of squares) took nearer 36 hours, due to the need to adjust for spatio-temporal correlation.

As discussed in Section 5.1.3, adjustments must be made to the residual sums of squares and degrees of freedom for the error in each of the models in order to account for the residual spatiotemporal correlation in the data. The formulae in Section 5.1.3 were used to calculate these numbers for the models in Table 5.1, using the correlation matrices that will be calculated in Section 5.4.

Each of the models were fitted to the full 21 year dataset using the spatial and temporal bandwidth values that have been discussed previously. Table 5.1 shows the residual sums of squares and degrees of freedom for the error terms in the four

Table 5.1. Comparison of Sums of Squares for all Additive Models

Model	Name	Terms	*Adjusted RSS*	DF
1	Main Effects	$\mu + m_1(spatial) + m_2(year) + m_3(day) + \varepsilon$	2100.886	10976.75
2a	Season Int.	Main Eff+$m_4(spatial, day)$	1839.607	10790.90
2b	Trend Int.	Main Eff+$m_5(spatial, year)$	2061.465	10854.07
3	Full Model	Main Eff+$m_4(spatial, day) + m_5(spatial, year)$	1724.399	10663.15

Table 5.2. Results of F-tests for the Additive Models

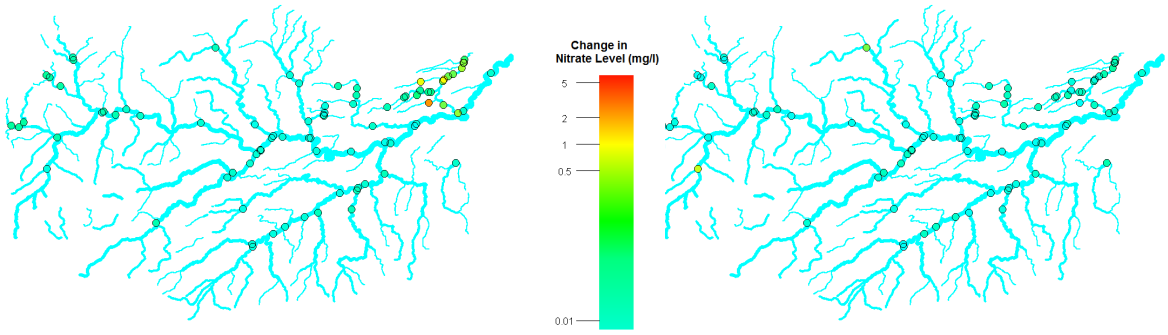
Model 1	Model 2	F-statistic	Critical Value
Full	Main Effects	7.373	4.262
Full	Season Int.	5.577	3.071
Full	Trend Int.	10.917	3.599
Season Int.	Main Effects	8.152	3.578
Trend Int.	Main Effects	1.663	3.042

fitted additive models. These statistics were then used to perform approximate F-tests (5.15) to compare the models, the results of which are shown in Table 5.2. It is interesting to see that the F-statistic is less than the critical value only for the test of the main effects model vs. the trend interaction model. This seems to reflect the fact that there are very few strong long-term trends present in the River Tweed data. However, when the trend interaction term is used in addition to the seasonal interaction term in the full model, it does seem to be explaining enough of the remaining variability to produce a significant F-statistic. The F-tests suggest that the full model is preferable to all of the reduced models as the F-statistic in each of the tests involving the full model is larger than the critical value. This suggests that the added complexity of the full model is worthwhile as it provides a significantly better model fit than could be obtained by using just one interaction term or just the main effects model.

Figure 5.3 shows the effect that the two interaction terms on the spatial component of the model, by plotting the maximum (absolute) value of the interaction over the 21 year period, for each of the interaction terms. The maximum is used as the interaction term changes over time so, by showing the maximum value, the

stations most affected by each interaction can be seen reasonably well. It can be seen that almost none of the stations experience any change in fitted values due to the trend interaction (Figure 5.3(b)). Only two stations experience any real change , and it is no surprise that the largest change is observed at the Kingledores station in the south west of the map. The other station to see a change, the Kilcouter Bridge station on the Heriot, is one of the only other stations with a slight trend in the data, albeit less than at Kingledores. The plot of maximum change in spatial term over the changing season, Figure 5.3(a), shows many more monitoring stations being affected by the season interaction than the trend interaction. It is also very interesting to see that the stations with the largest maximum change are all on, or are on tributaries of the Leet in the north east of the Tweed network area. Very slight changes are observed elsewhere, such as the cluster of stations to the far west, but they are all very minor compared to those seen on the Leet. This implies that the seasonal components in this area tend to be quite different to those observed elsewhere in the network.

Assessing the Fitted Model



(a) Maximum change in spatial effect arising from seasonal interaction

(b) Maximum change in spatial effect arising from trend interaction

Figure 5.3. Maximum value of interaction term plotted over space

Figures 5.4, 5.5, 5.6 and 5.7 show the fitted values and effects from the full (two interaction) model for a selection of four stations. Figure 5.8 displays the fitted model at eight more stations but without showing the seasonal and temporal trends. In each figure, the four panels correspond to (from top left to bottom right) the fitted model, the fitted seasonal component, the fitted trend and the location of the station on the network. The red dotted line around each of the fitted lines corresponds to twice the standard error, which is calculated using the formulae in Section 5.1.3 along with the spatiotemporal correlation matrix Σ that is estimated in Section 5.4.

The values of the fitted model, seasonal component and trend are calculated using the projection matrices P_j^l after convergence of the backfitting algorithm (after "l" iterations). The full interaction model has six components; the mean; spatial, seasonal and trend effects; and the space-season and space-trend interactions. The fitted value for each of the components of the model is calculated by multiplying the projection matrices by a vector y containing the data i.e. $\hat{m}_j^l = P_j^l y$. The fitted values for the data are then calculated by adding together the $\hat{m}_j^l$ for each of the six components, and these are shown alongside the observed data in the top left panel on each plot. The plots of the fitted seasonal and trend components are not quite as straightforward to obtain. For the fitted seasonal component, the fitted line is calculated by adding together the $\hat{m}_j^l$ for just the seasonal component and the interaction term between space and season. This means that each spatial location has a different seasonal component associated with it. This is plotted against the residuals that are obtained by subtracting the four remaining $\hat{m}_j^l$ from the observed data. The plot of the fitted trend is obtained in the same way, but using the $\hat{m}_j^l$ from the trend and the space-trend interaction instead of the seasonal components.

Figure 5.4 shows the full interaction model for the Norham gauging station.

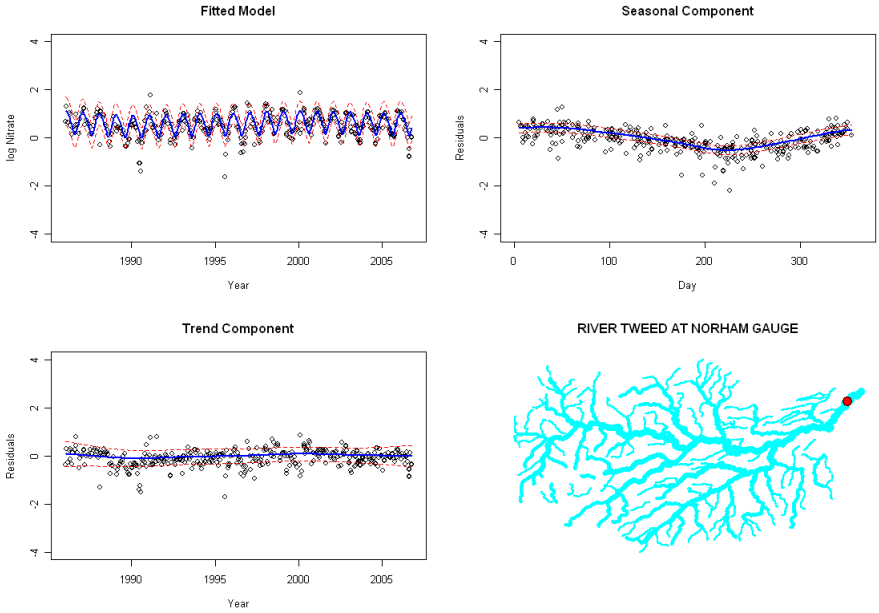


Figure 5.4. Fitted effects for full additive model at Norham

This station has again been chosen as it is the furthest downstream monitoring station and, as such, provides the best estimates for nitrate levels being discharged into the North Sea. For this station, although not shown, the main effects model produces a very similar fitted model to that shown here for the full interaction model. Given that this station is the furthest downstream, and therefore is fed into by all other monitoring stations, it follows that the 'overall average' model fit provided by the main effects model would be a good fit to the data. Looking at the seasonal component for the full model, there is a small difference between the lowest and highest nitrate levels. Levels are at their lowest in summer (around August) and at their highest around February. The trend component for this station is very flat, with only slight inclines or declines over the 21 year period. This is typical of the trends fitted at most stations, and a similar trend can be seen in many of the fitted models shown in Figure 5.8. Most fitted trends on the river network show a slight increase between the early and late 1990s followed by a slight decline after 2000. This is consistent with earlier results and findings.

While the full interaction model provides a good fit to the data at this station, its benefits over the single interaction models and the main effects model can be more easily seen by looking at other stations.

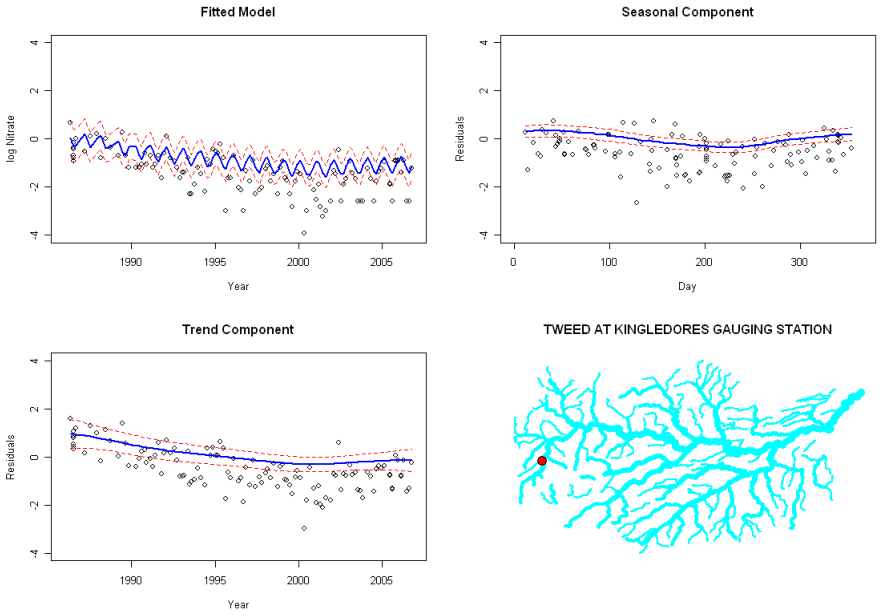


Figure 5.5. Fitted effects for full additive model at Kingledores

Figure 5.5 shows the effects and model fit from the full interaction model at Kingledores, and so can be directly compared with Figure 5.2(b) to see the effect that the interaction terms have had on the fitted values. The interaction model fits a noticeable, but not too steep downward trend that captures some of the trend that seems to be present at this station. This is a definite improvement on the main effects model (Figure 5.2(b)), which does not allow for a difference in trend across the spatial locations. Despite being an improvement on the main effects model, the fitted model is still not as good a fit to the data as might have been hoped. This is because the fitted trend is smoothed across space and the lack of any definite trends elsewhere, let alone trends similar to the one seen here, has the effect of flattening out the fitted trend. The fitted model at this station is one of the poorest at any location due to the inability to handle the trend. This

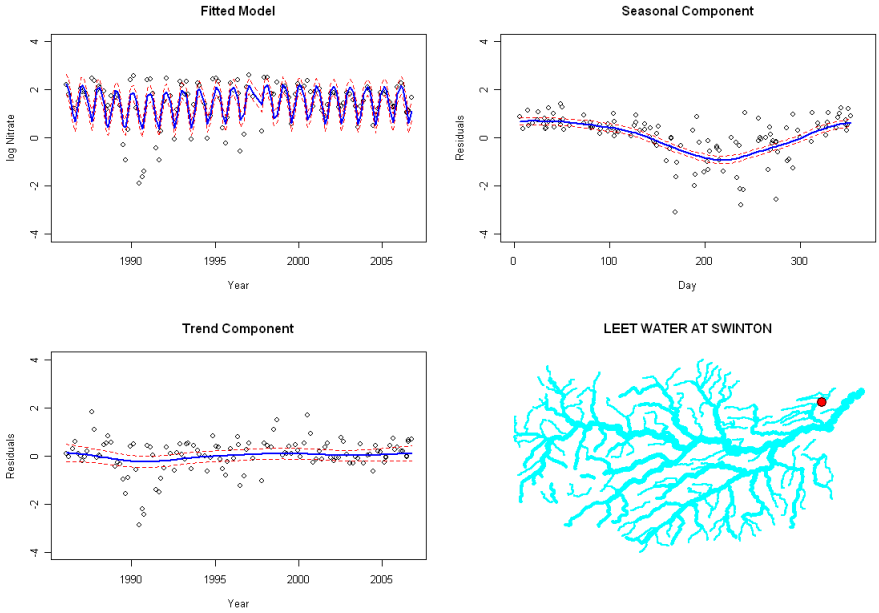


Figure 5.6. Fitted effects for full additive model at Charterpath bridge on the Leet

problem is unavoidable for the Tweed dataset as this station is something of an 'unusual' observation in trend terms when compared to the rest of the monitoring stations. One way round this problem would be to fit separate additive models at each of the monitoring stations. This would lead to a better description of what is happening at the individual station level but the results would not give a further of the overall spatial patterns.

Figure 5.6 shows the fitted model and its seasonal and trend components for the monitoring station on the Leet at Charterpath Bridge. This is the same station that is shown for the main effects model fit in Figure 5.2(a) and so again, the two fitted models can be compared. The difference between the amplitudes of the fits in the two models demonstrate by far the biggest advantage of using the interaction model for the Tweed data: the space-season interaction allows the seasonal pattern to change across space. The benefits of the trend-space interaction were minimal due to the fact that trends were only observed at a small

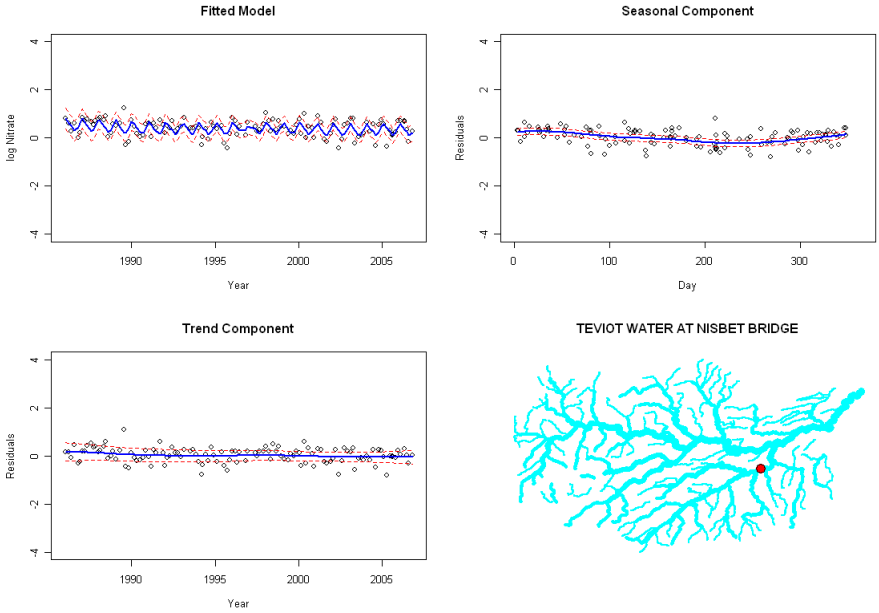


Figure 5.7. Fitted effects for full additive model at Teviot

number of isolated stations, but the space-season interaction seems to benefit the vast majority of monitoring stations when compared to the main effects model. There seem to be distinct clusters of monitoring stations on the various tributaries around the Tweed that share similar seasonal patterns, and this pattern lends itself very well to the way the values are fitted in the full interaction model. The fitted model in figure 5.6 does a reasonable job of capturing the observed data, with the amplitude of the fitted curve just about large enough to capture the majority of the variation across a year. The only problem seems to be that the variability of the observed values seems to dramatically increase in the summer, meaning that several observations fall quite far away from the fitted model. It would be possible to investigate the accuracy of the estimated confidence bands, but such an investigation was outwith the scope of the investigation here. This could potentially be carried out by conducting a study similar to those found in Chapter 3.

The rest of the monitoring stations on the Leet and its tributaries are also modelled very well by the interaction model, such as the station at Coldstream which is shown in Figure 5.8(b). All stations in this region tend to have a larger amplitude to their seasonal component than would have been fitted by the main effects model. This benefits the interaction model fit as the smoothing of the seasonal component over space is smoothing over stations with very similar amplitudes. This smoothing was a hindrance when it came to the trend at Kingledores, where the observed trend was effectively smoothed over, but it is a benefit to the model fit here, borrowing strength from nearby stations. The only potential downside to this is that it may lead to 'quirks' at individual station level being lost. For instance the fitted model at Coldstream, Figure 5.8(b), fits reasonably overall but tends to miss a few very low values in the summer. During the summer of 1999, daily sampling took place at this location (as was discussed earlier), and they seem to form a well defined 'trough' shape, which the fitted model does not capture. This may suggest that the actual seasonal component may have a larger amplitude than has been fitted, or possibly that the amplitude has changed over time. Of course, it is possible that these observations in 1999 are an exception rather than the rule, but observations at this time of the year in other years suggest otherwise. This problem is at its most pronounced at this station, although there is evidence that it might be the case at some other stations on the Leet (albeit to a much lesser extent). Again, if a more descriptive model was required to investigate this station then the data could be modelled separately for individual stations.

Figure 5.7 shows the fitted model and effects for the Nisbet Bridge monitoring station on the Teviot. A further example of a station on this tributary can be seen in figure 5.8(g). The station at Nisbet Bridge is fairly typical of other stations on this and other tributaries to the south of the main body of the River Tweed, in that it has very little seasonal component or temporal trend. Looking at the

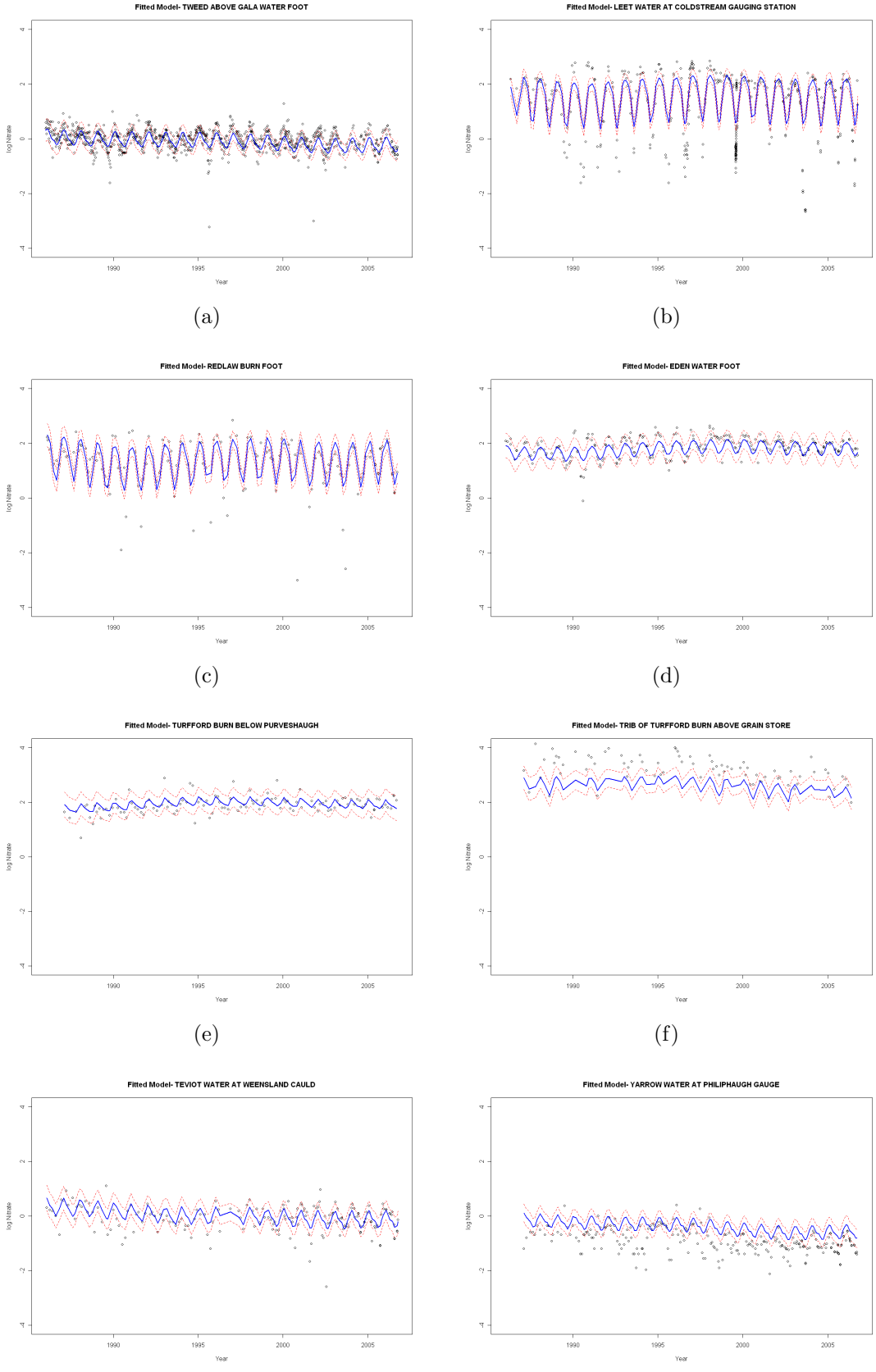


Figure 5.8. Full Interaction Model Fitted Values at a selection of sites

seasonal component, there is very little difference between the highest and lowest points of the cyclical curve especially when compared to the very pronounced seasonal component seen in Figure 5.6. The very flat trend fitted at Nisbet Bridge again illustrates that there are very few trends present in any station's data in the Tweed network.

Figure 5.8 contains the fitted models from a selection of monitoring stations around the Tweed, in order to try to give a general feel for the fitted models across the network. The fitted spatial component in the interaction model has not been discussed yet, and some of the plots shown here demonstrate locations where the spatial component does not seem adequate to capture the data. Figures 5.8(f) and 5.8(h) are two such locations where the spatial component does not fit particularly well. In the model, the spatial component raises and lowers the mean value at a particular station relative to the other stations. In Figure 5.8(f), it does not raise the mean value quite enough to provide a good fit for the data. This station lies in the Turfford Burn area, a very high nitrate zone, and has some of the highest nitrate levels of any of the stations in that area. The fitted spatial component for the area is very large, but as it is smoothed over the nearby stations it is not as high as it would need to be to accurately model the data. In contrast, Figure 5.8(e) shows another monitoring station in the Turfford Burn area but the fitted spatial component is sufficient for the model to accurately reflect the data. The Philiphaugh station on the Yarrow, shown in Figure 5.8(h), demonstrates exactly the opposite problem as the fitted spatial component is too large to accurately fit the data.

The plots in Figure 5.8 also demonstrate how the width of the confidence bands changes from station to station. The confidence bands represent the fitted model plus or minus twice the spatiotemporally adjusted standard error. The confidence bands vary from being relatively wide, as in Figures 5.8(d) and 5.8(e),

to much tighter, and more closely following the fitted model, such as in Figures 5.8(b) and 5.8(c). For the latter two, the confidence bands at the peaks and troughs of the seasonal cycle are relatively wide, but are much closer to the fitted model during the rest of the cycle. It is not clear if this reflects the fact that variability of observed nitrate tends to be higher at the peaks and troughs or a lack of observations outwith these times of the year, or both. There does seem to be fewer observations at these times of year than during the peak and trough periods, but the model seems to fit these observations quite well (except in Figure 5.8(d) for the summers between 1998 and 2000 inclusive where, for some reason, the summer observations are more in keeping with those that would be expected in winter).

5.4 Correlation in Space and Time

Models for spatio-temporal correlation can be divided into two classes – separable (5.20) and non-separable. Separable models are based on the assumption that it is possible to separate out the components of the covariance structure whereas a non-separable model does not make this assumption. A non-separable structure would mean that the spatial correlation structure changes over time, and so the components would have to be estimated simultaneously. Separability, on the other hand, means that each component could be estimated separately and then combined to make a valid spatio-temporal covariance structure.

Cressie (1991) defines a separable covariogram as a covariogram with $C(h) = \prod_{i=1}^{D} C_i(h_i)$ where D is the number of dimensions. (5.20) shows a separable model for the covariogram, where C_s is the spatial correlation structure, with stream distance h_s between the two points, and C_t the temporal correlation structure,

with a time difference of h_t between the two points.

$$C_{st}(h_s, h_t) = C_s(h_s)C_t(h_t) \tag{5.20}$$

It was believed that the errors in the additive models that were created in this chapter were very likely to be correlated, whether that be spatially, temporally or both spatially and temporally. The reported errors (in the form of the error bands on the additive models), in addition to the residual sums of squares and degrees of freedom reported in Tables 5.1 and 5.2 have all been adjusted to account for spatio-temporal correlation. This section will discuss the methods used to estimate the residual correlation left in the data after the additive modelling process. Covariograms will be estimated to describe both the spatial and temporal covariance in the residuals from the additive modelling process, and these will then be multiplied together as in (5.20) and will be converted to correlations in order to adjust the results.

5.4.1 A Test for Presence of Correlation

Diblasi and Bowman (2001) describe a test for the presence of correlation. With a set of n observations d_{ij}, which in this example would represent the residuals from the additive model, with mean value $\bar{d}_{ij}$, then (5.21) denotes a test statistic that can be used to check for the presence of correlation in the d_{ij}. In (5.21), d contains the d_{ij} in vector form while $B = (I - W)^T(I - W)$ and $A = I - L - B$. In this formulation, the rows of W contain the smoothing weights used to construct $\bar{d}_{ij}$, L is the matrix with all entries $2/(n(n-1))$ and I is the identity matrix.

The test statistic tests the null hypothesis that there is no correlation present

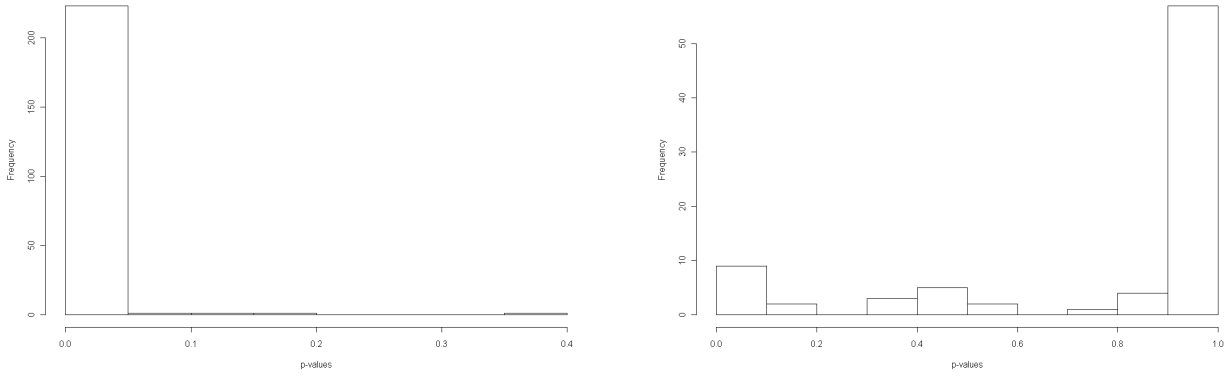
in the residuals against the alternative hypothesis that there they are correlated. The test statistic is essentially the sum of n χ^2 random variables, and this is not easy to evaluate so an approximate distributions is used, and the test performed numerically. The distribution of this test statistic is "equivalent to that of $e^T Ae/e^T Be$, where e has a multivariate normal distribution with mean zero and covariance matrix Σ".

$$T = \frac{d^T Ad}{dTBd} \tag{5.21}$$

This test can be carried out using the 'sm' package in R (Bowman and Azzalini, 2007), and this was used to test the additive model residuals for presence of residual spatial and residual temporal correlation. The aim of these tests was to obtain a subjective impression about whether the residuals were likely to be correlated in space and/or time. Therefore, the residual spatial correlation was tested using Euclidean distances to form the smoothing matrix W. Stream distances were considered but the added complexity of flow connectivity and weighting made this metric overly complex for the purposes of a subjective impression.

Individual tests of spatial correlation were carried out for the residuals at snapshots in time (that had sufficient monitoring data available). Similarly, individual tests of temporal correlation were carried out for the residuals at each station. This means that there were 83 different tests of temporal correlation and around 250 tests of spatial correlation. Under the null hypothesis of no residual correlation, it would be expected that the p-values for each set of tests would be uniformly distributed between zero and one, and so deviation from this may suggest that the data are correlated.

Figure 5.9 shows histograms of the p-values for the spatial and temporal correlation tests. The spatial correlation tests, shown in Figure 5.9(a), give a clear



(a) Spatial correlation (b) Temporal Correlation

Figure 5.9. Results of tests for presence of spatial and temporal correlation

indication that the residuals from the additive model are spatially correlated, according to Euclidean distances. Almost all of the p-values are less than 0.05, suggesting that there is a significant correlation between spatial locations at the majority of snapshots in time. If stream distances were used, this overall conclusion is unlikely to be different.

The overall conclusion for residual temporal correlation is very different and requires more interpretation. Figure 5.9(b) shows that the majority of p-values are greater than 0.9, implying that there is no significant evidence for the presence of residual temporal correlation. Closer inspection of the p-values may give some insight into the reasons for this result. The stations with lower p-values tend to be monitored on a more regular basis than most of the others, suggesting that the evidence against residual correlation may be a result of the regularity of the monitoring at those locations. It is widely accepted that environmental data collected more than two weeks apart are not significantly correlated in time (Van Belle and Hughes, 1984), and so it would follow that those locations with less regular monitoring (and observations generally more than 2 weeks apart) would not have significant residual temporal correlation, as these tests would imply.

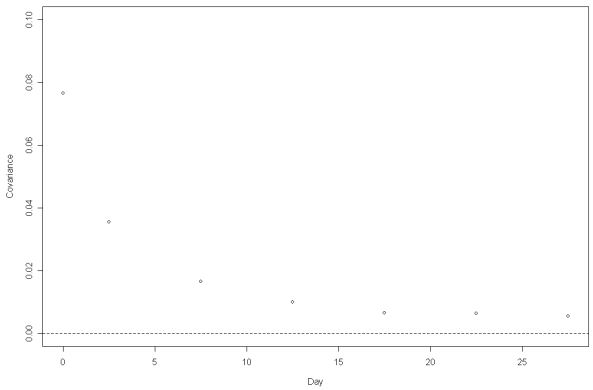
Consequently, there is little doubt that spatial correlation should be accounted for when adjusting the residuals from the additive models. The evidence from the tests carried out in this section may suggest that there is no need to account for residual temporal correlation, but this may be as a result of the sampling frequency at certain stations. Therefore, a conservative course of action would be to account for the temporal correlation in the additive model residuals too.

5.4.2 Modelling the Residual Correlation

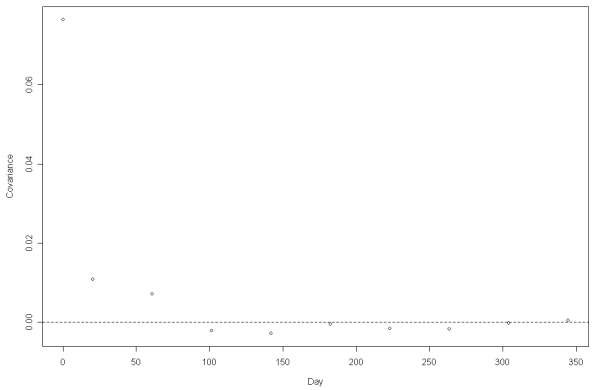
All of the work done so far in this chapter has accounted for correlated errors, and this section will explain how the spatio-temporal covariance structure used to do this was estimated.

The equations used to account for correlation in the standard errors and summary statistics in the additive model are outlined in Section 5.1.3, but in order to use them it is necessary to to first estimate the correlation matrix Σ. This is done by first fitting the additive model and then extracting the residuals to see if they are correlated in either space or time. The residual correlation was assessed using covariograms, produced as described in Chapter 2. Figure 5.10 shows the residual covariance in time, over a period of 30 days and one year respectively, and then the residual covariance in space, of the residuals from the fitted main effects model (5.2). As described in Chapter 2, the spatial covariances have been adjusted to account for the river distance weighting structure.

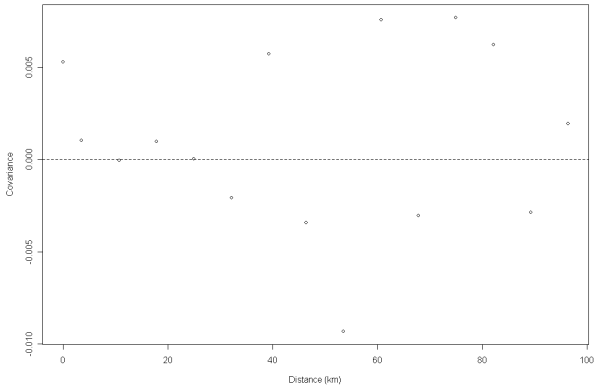
For the main effects model, it would appear that there is not too much residual temporal correlation left in the data. Figure 5.10(b) shows the correlation over a period of 365 days, and suggests that the covariance is very small after a lag of 30 or so days and is zero at lags of over 100 days. Looking at the lower lags in finer detail in Figure 5.10(a) suggests that the covariance drops off significantly



(a) Residual Temporal Covariance - One Month



(b) Residual Temporal Covariance - One Year



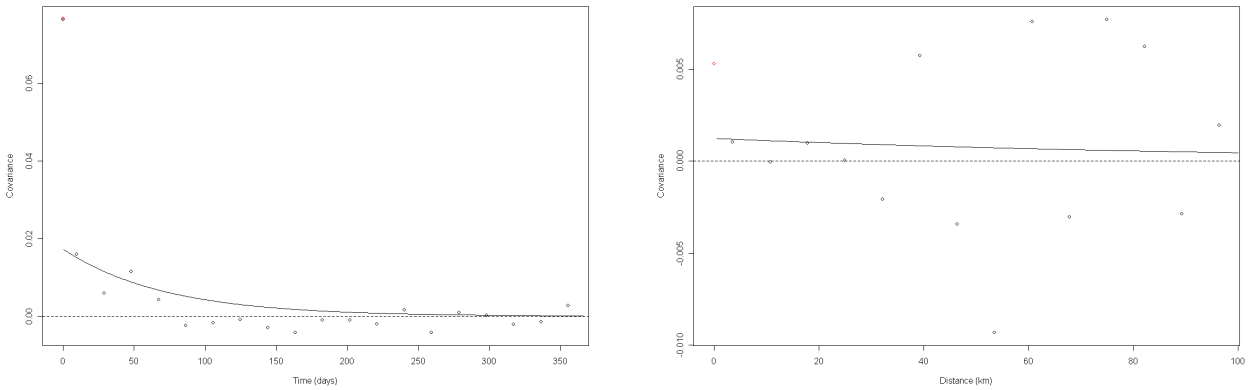
(c) Residual Spatial Correlation

Figure 5.10. Residual Covariance From Main Effects Additive Model

within the first two weeks and levels off somewhat at that point. Given that very few observations in the Tweed dataset are made less than two weeks apart, this suggests that there is very little residual temporal correlation left in the data.

For the spatial correlation, the pattern is far less clear. For the lags between zero and twenty kilometres, the covariances seem to drop off to around zero after one or two lags . For lags of over thirty kilometres the picture is far more noisy. Large covariance values are observed at these lags, and the covariances seem to alternate from negative to positive from lag to lag. The lack of any systematic pattern in this plot makes it difficult to say if this is due to noise or if there is actually some spatial covariance left in the data after modelling. The fact

that the values are alternating between positive and negative may suggest that it is just noise, and therefore that the covariance is effectively zero after a lag of around 30km. This is not particularly satisfactory, but it may be explained if there were some temporal correlation left in the mixture model, which is still observed in the spatial plot but does not make much sense when plotted against spatial lag.



(a) Temporal (b) Spatial

Figure 5.11. Fitted Covariance Models for Main Effects Model Residuals

The temporal and spatial covariances were modelled in the same way as in Chapter 2 using exponential functions. Figure 5.11 shows the models fitted to the observed covariances. For the temporal covariance, the fitted nugget took a value of 0.0594, the partial sill was 0.0173 and the range was 71.3 days. For the spatial covariogram, the nugget was estimated as 0.00407, the partial sill was 0.00124 and the range was 104km. These values are shown in Table 5.3. The difference in the magnitudes of the sills (i.e. nugget + partial sill) in each of the fitted covariogram models may suggest that the temporal covariance is stronger than the spatial covariance, though neither has particularly high values. There is perhaps an argument that there is very little covariance left in the data at all after the first lag, at least visually from the plots. There definitely seems to be a distinct pattern of temporal correlation present in the residuals, but this drops

Table 5.3. Estimated Covariance Parameters for Additive Models

Model	Temporal			Spatial		
	Nug	Sill	Range	Nug	Sill	Range
Full	0.0434	0.0171	69.2	0.00212	0.00200	13.6
Season Int.	0.0330	0.0128	438	0.00224	0.00096	68.3
Trend Int.	0.0334	0.0105	511	0.00940	0.00211	59.8
Main Effects	0.0594	0.0173	71.3	0.00407	0.00124	104

off significantly after just a week and is practically zero after just two weeks. Given the frequency at which the Tweed data was collected this means that only a tiny minority of residuals are likely to be correlated in time and could be used as an argument against adjusting for temporal correlation. Given the rapid drop in the spatial covariance followed by the noise after 30km, a similar argument could be made about the spatial covariance. However, the fitted models for each were used to define the residual correlation after the additive modelling and were subsequently used to adjust the standard errors, degrees of freedom and sums of squares as described previously.

Table 5.3 shows the estimated covariance parameters for the main effects model along with those estimated for the the two single-interaction models and the full model. The fitted covariograms for the remaining additive models are also shown in Figure 5.12.

Looking first at the fitted temporal covariograms, there seems to be very little difference between the full and main effects models in terms of either the observed covariances or the fitted model. Both follow a very similar pattern, following a reasonably well defined curve and with a very large jump at lag zero. The fitted covariance parameters are all very similar in both sets of residuals. This is in contrast to the covariograms of the two single-interaction models, which are very similar to each other but very different to the full and main effects model covariograms. Both single interaction models have very noisy plots that do not seem to tend to zero, implying residual temporal correlation lasting longer than

one year.

Moving on to the spatial covariograms, all three interaction models have less noisy covariograms than the main effects model. As was the case with the temporal covariograms, the spatial covariograms for the two single-interaction models look very similar to one another. Neither has a particularly strong pattern, though both start relatively high before tending to a value slightly above zero. Neither suggests that it will eventually tend to zero, but this could again be due to noise rather than being a sign that the spatial covariance does not trend to zero at all.

In general, it is interesting to see that the most distinct patterns observed in any of the models' covariograms are seen in those belonging to the full model. The fitted covariance models correspond better to the observed covariances and this could be seen as further proof that the full model provides a better fit to the data. The resulting covariogram is more regular which therefore suggests that some of the noise in the other covariograms may have been accounted for by the interaction terms. Even though the errors and summary statistics of the additive models fitted earlier have been adjusted using these estimated covariances, it is not quite clear whether this was required. The magnitudes of the covariance models used are quite low and seem to tend to zero quite rapidly. In the case of the full model, this is estimated to be the case after a temporal lag of just 70 days or a spatial lag of around 14km. As these values are reasonably low there are not many pairs of residuals that fall underneath one of these boundaries. Nevertheless, the estimated covariances were used to adjust the output from the additive models that were fitted and discussed previously.

5.5 Conclusions

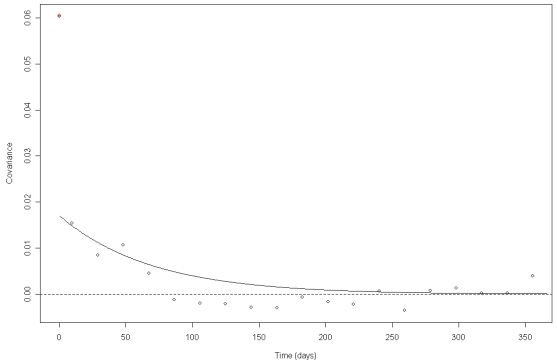
The River Tweed data has been modelled through space and time using additive modelling techniques. A novel spatial smoothing component incorporating the ideas used in the tail-up spatial model proposed by Ver Hoef and Peterson (2010), such as stream distance, flow volume and flow connectivity, has been created in order to tailor the additive model to be more suitable to the context of river networks. The spatial smoothing term was then used alongside a cyclical smoothing term and a local linear smoothing term in order to create an additive model with spatial, seasonal and temporal trend terms. This model seemed an inadequate description of the variability in the Tweed data though, and so two novel interaction terms were formulated to allow for differing seasonal and temporal trend terms across space. The main effects model, the two single-interaction models and the model containing both interactions were then tested using approximate F-tests to determine which provided the most suitable fit to the Tweed data. This showed that the model with the two novel interaction terms provided a significantly better fit to the data

In order to account for the potential spatio-temporal correlation present in the data, the standard errors in addition to the residual sums of squares and degrees of freedom for the error (required in the F-tests in order to compare the models) had to be adjusted. In practice this meant first fitting the models, then fitting a model to the correlation still present in the residuals from the models and finally refitting the models and adjusting by incorporating the correlation matrix V. It was not clear whether it was necessary to adjust for the temporal correlation, given subjective tests suggested that this was negligible for the residuals at the majority of monitoring stations. However, it was decided that the conservative course of action would be to adjust the residuals to allow for temporal correlation anyway.

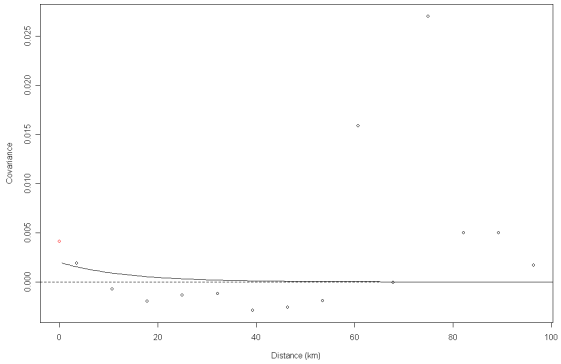
In general, Figure 5.8 in addition to Figures 5.4, 5.5, 5.6 and 5.7 give a good picture of how well the fitted full interaction model reflected the observed nitrate values at those and the the remaining 71 monitoring stations. Strong temporal trends were not prevalent on the River Tweed and so fitted trends such as those that were seen at Kingledores (Figure 5.5) are rare. Most fitted trends tended to be very slight increases or decreases over time such as in Figure 5.4. Approximate F-tests on the models with and without interaction terms seemed to back this up, as the model with only a space-trend interaction was rejected in favour of the main effects model. The trend interaction was significant when used in conjunction with the season interaction.

Seasonal patterns were much more varied over the River Tweed monitoring stations than trends. The seasonal trend at many stations seemed to have amplitudes similar to Norham (Figure 5.4), and this was reflected in almost all plots in Figure 5.8 with the exception of Figures 5.8(b) and 5.8(c). It should be noted that while Kingledores had almost the only major trend and was used as an example for this reason, Figures 5.8(b) and 5.8(c) were just two of several examples of larger amplitude seasonal components on the Tweed. This was reflected in the fact that the approximate F-tests suggested that the model with the space-season interaction was favoured when compared to the main effects model. The varying amplitudes only tell part of the story, as the season interaction will also allow the seasonal component to vary in terms of the start date, giving an even greater degree of flexibility to the model. It is unlikely given the observed data that this will have made a huge improvement to the fitted values as the peaks and troughs in the nitrate levels tended to be around the same times for each monitoring station. Nevertheless, the greater degree of flexibility will allow slight adjustments to the cyclical component over space, which could reflect the different nuances in the data (for instance, if a farmer in one location on the river tended to fertilise his crops slightly earlier than one at a different location).

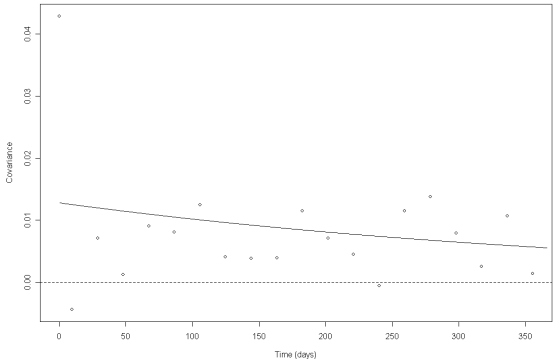
The results of the approximate F-tests suggest that the interactions that have been introduced in this chapter have significantly improved the effectiveness of the fitted model in capturing the change in nitrate level over time on the River Tweed. This technique could be applied to any other river network, and the significance of the space-trend and space-season interactions would depend on the variability of the trends and seasonal components in the new dataset.



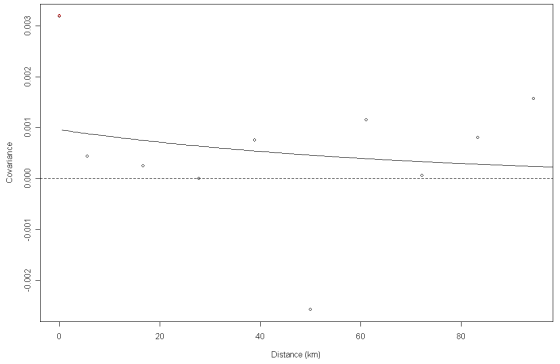
(a) Full Model- Temporal



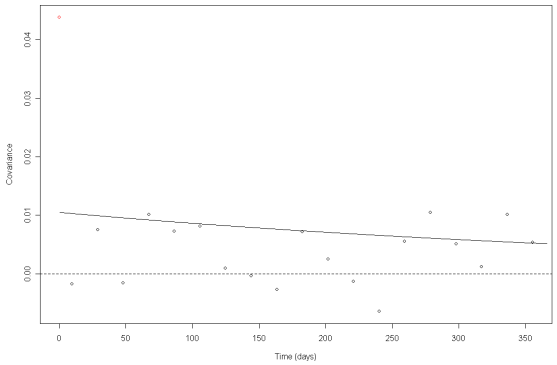
(b) Full Model- Spatial



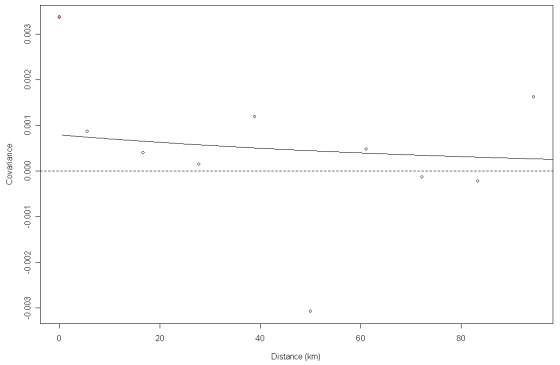
(c) Season Interaction- Temporal



(d) Season Interaction- Spatial



(e) Trend Interaction- Temporal



(f) Trend Interaction- Spatial

Figure 5.12. Fitted Covariograms for Full, Season Interaction and Trend Interaction Models

Chapter 6

Conclusions and Extensions

6.1 Conclusions – Spatial Prediction

The unique characteristics of river networks provide interesting challenges to both predicting levels of pollutants at unobserved locations and modelling pollutants through space and time.

The 'tail-up' covariance structure (Ver Hoef and Peterson, 2010) was used as the basis for all of the methods used in the analysis of the River Tweed data. The tail up structure was the first valid covariance model to be proposed that allowed for the use of stream distance in place of Euclidean distance. Previous analyses of river network data had either used the Euclidean distance, or had used stream distance with the standard covariance structures designed for use with Euclidean distance. In addition to the tail-up structure providing a valid model for covariance using stream distance, it also had several other desirable properties that further link the covariances to the intrinsic properties of the river itself. The connectivity of the network is accounted for so that two locations that are not flow-connected (i.e. that it would not be possible for the water

at either location to reach the other without having to go back upstream at some point, against the direction of flow) are assigned to have a covariance of zero. This essentially assumes that the level of a pollutant at one location will have no impact on the predicted value at a flow-unconnected location. The tail up structure also incorporates a weighting system that can be based on flow volumes, stream orders or some other surrogate for the size of the stream. This means that the volume of water contributed at the point of confluence by each of the two merging streams is taken into account, so that the larger stream will contribute more to the resulting prediction than the smaller stream.

Cressie et al. (2006) contains one of the earlier uses of the tail-up structure in the literature, and proposes further that it could be used in conjunction with a Euclidean distance based covariance structure to create a mixture model of the two. At first it may be unclear as to why this would be desirable, as the tail-up model provides a much more intuitive description of what is likely to be happening on a river. However desirable it may be to specify zero covariance between flow-unconnected locations from an ecological point of view, in practice this is unlikely to be representative to environmental covariates (such as land use) causing nearby but flow-unconnected locations to have similar characteristics. This is where a mixture of the tail-up structure and a Euclidean distance based model might be beneficial. Later literature (for example Peterson and Ver Hoef (2010)) has shown that a mixture of stream and Euclidean distance structures can lead to more accurate predicted values, though it was still unclear whether this would be the case when work on the Tweed data commenced. The tail-down model structure, which uses stream distance but allows nonzero covariances between flow-unconnected locations, has also been used in such mixture models in the literature, but was not used in analysis of the Tweed data. This is because it was felt that the tail-down structure was more suited to data on fish stocks or determinands linked to fish, such as ammonia, and that the modelling of nitrate

levels would not benefit from incorporating this approach.

6.1.1 Trend and Parameter Estimation

One of the first things to be considered in the analysis of the Tweed data was the possibility of detrending the data prior to spatial prediction. The absence of trend (beyond an overall nonzero mean) in the data is one of the assumptions of kriging, but previous literature has used parametric trends as part of universal kriging, which seem unlikely to adequately account for this assumption. Also, the river network structure throws up the further complication of whether to remove a trend based on stream or Euclidean distance. A novel trend based on stream distance was created, using the same kind of structure as used in the tail-up model, and this would later form the basis for the stream distance smoothing functions in additive modelling. There was little difference in the accuracy of the predictions created using the Euclidean and stream distance based trends in terms of the lowest root mean squared errors obtained using a cross-validation procedure, suggesting that neither was necessarily a 'better' metric for use with the trend than the other. This seemed to be partly because the covariance mixture model that was being used to create the kriging predictions balanced out the trend; so the stream distance trend was balanced by more weight being put on the Euclidean component of the covariance mixture and vice versa. Therefore it is likely that choice of which metric to use for the trend would come down to whichever metric was likely to better reflect what an ecologist (for example) thought the underlying trend represented. For instance, a Euclidean trend may better represent the trend coming from land use or land composition. Given there was very little difference between the detrending methods and uncertainty as to which would better represent the trend of a river, further kriging analysis was carried out using both Euclidean and stream distance detrending.

6.1.2 Studies into the Mixture Model for Covariance

Four studies were carried out in order to investigate different facets of the estimated trends and covariance structures.

Mixture Model and Detrending Investigation

The first study investigated whether a mixture of stream and Euclidean distances in the fitted covariance model would provide more accurate kriged predictions, again assessed using a cross-validation procedure. A mixture model for river network data was proposed in Cressie et al. (2006), but it was unclear whether this would provide a more accurate fit than using a single distance metric. Analysis on the River Tweed data found that, for both stream and Euclidean distance based detrending methods (as well as when no detrending was performed prior to analysis) a mixture model provided a lower crossvalidated root mean squared error than either model using just one distance metric. The lowest RMSPEs for each of the two detrending metrics were very similar, but the mixing parameter (used to decide how much weight to put on each of the distance metrics in the covariance model) at which they were found was quite different for each. The Euclidean detrended example tended to put more weight on the tail-up model in the mixture, while the stream distance detrending put slightly more weight on the Euclidean model. This seemed to suggest further that the choice of detrending metric did not matter as the covariance model would effectively balance it out to account for the Euclidean and stream distance based elements of variability in the data. These findings were consistent with those in subsequent literature that also found that a mixture of different covariance models (including the tail-down model as well as tail-up and Euclidean) could lead to a reduction in RMSPE.

A follow-up to this study was also carried out. This simultaneously estimated

both the Euclidean and stream distance covariance parameters in the mixture model for each mixing parameter in a specified range, before comparing them in terms of RMSPE. The results for all mixing parameters (except those corresponding to just the Euclidean or stream distance models on their own) were almost identical. This further suggested that allowing certain parameters to change and selecting based on lowest RMSPE, had a something of a 'balancing' effect, where fixing some parameters (in this case the mixing parameter) would lead to the 'optimal' values of the other parameters changing in order to balance out the effect of fixing the others. The RMSPEs produced in this part of the study were larger than those found in the earlier part, but this was likely to be due to slightly different covariogram estimation techniques. It also provided the motivation to go on and construct a more thorough study of covariance parameters in Study 2.

Covariance Parameter Investigation

In order to investigate the covariance parameters as potential sources of error, the second study attempted to show the roles played by different covariance models and parameters on prediction error. Firstly, it was demonstrated that the nugget and sill do not have any effect on the predicted values (and thus the RMSPE) from kriging, and altering these only serves to change the kriging standard errors. Therefore the range parameter became the focus of further investigation. In order to do this, another cross-validation procedure was constructed. This time, the covariance models using Euclidean and stream distance in the mixture model were allowed to follow the exponential, spherical or linear with sill models, instead of just the exponential model. This was performed using data that had been detrended using the stream distance method. All 9 combinations of covariance models for both elements of the mixture model were tested, with a variety of mixture parameters, and a variety of range parameters used for each

element of the covariance mixture model. The lowest RMSPE found used the linear with sill model for both distance metrics, had a Euclidean covariance range of 38km, a stream distance range of 58km and a mixing parameter of 0.2 (putting more weight on the Euclidean covariance). However, the difference between the optimal models for each combination of models (Exponential, Spherical, Linear) was very slight. The poorest RMSPE of any of these combinations was found by using the Exponential model for both distance metrics, and this was 0.3220 as opposed to the lowest combination RMSPE of 0.3133. The differences observed between the model combinations were small enough that it is possible that they were due to the ranges of values checked for each parameter– mixing parameters in increments of 0.05 and range parameters in increments of 2km were used– and that fine-tuning would produce even more similar results. This suggested that covariance model type does not have a huge impact on the RMSPE. The only slight caveat to this is the observed volatility of the linear with sill model when used for the Euclidean portion of the mixture model. No reason could be found as to why this was the case, but through observing the condition numbers of the matrices used in kriging it seemed that certain combinations of range and mixing parameters and excluded stations (as part of the cross-validation) led to the matrix becoming ill conditioned, though not actually singular, in certain circumstances.

Trend Bandwidth Investigation

The third study looked into the covariance parameters that had to be estimated for both detrending methods and, within each, investigated the bandwidths that had to be selected for the trend components. Both of these were a potential source of additional error and so additional work was carried out to assess how much these affected the outcomes of other studies. For the trend

bandwidth, a sensitivity study was carried out by fitting a trend surface using a range of bandwidths and then looking at the lowest crossvalidated RMSPEs that were produced from each. The results suggested that altering the bandwidth did not have very much effect on the RMSPE, but seemed inconclusive if the object was finding an optimum bandwidth. This could again be due to the best mixture of covariances balancing out the change in trend surface. There may also be an issue with the cross-validation procedure used to conduct this study as it may be affected by the correlated errors making

Simulation Study

The final study was a simulation study, created in order to assess the mixture model on simulated networks, and also to assess the impact of preferential sampling on the River Tweed results. Initially, attempts were made to simulate the geography of a random river network but it proved too complex a task to do this without resorting to perfectly straight stream segments (which meant that several Euclidean distances were exactly the same as stream distances) and so the River Tweed geography was used and data generated for locations on it. The simulation study simulated from three different sampling schemes on the river; one using the sampled locations in the real data and performing cross-validation (the "Tweed sample"); one that simulated one value at each of the 298 stream segments and performed cross-validation on those (the "full sample"); and finally another that simulated from the full 298 locations but only used the 85 Tweed sampling locations to predict the remaining 213 simulations (the "other sample"). The simulated values were randomly generated draws from a multivariate normal distrubution with zero mean and a covariance structure given by the mixture model over a range of mixing parameters but using the covariance parameters estimated from the Tweed data. This approach was chosen in order

to produce realistic simulated values for a river network. The approach was repeated for model mixing parameters ranging between 0 and 1 in increments of 0.05 in order to see whether the lowest root mean squared error was generally found when using the underlying true mixing parameter.

For all three simulation scenarios, and over all the 21 underlying mixing parameters tested, the lowest RMSPE was generally found at or very near to the underlying true mixing parameter. There was no systemic drift of the optimal mixing parameter away from the underlying true mixing parameter for any of the sampling methods. Since a mixture of covariance structures produced the lowest RMSPEs for the real Tweed data, the results of this simulation study were encouraging as they implied that the 'true' underlying covariance structure of the data was likely to be found close to, if not exactly at the calculated mixing parameters.

In addition to ensuring that the estimated mixing parameters for the data were likely to be consistent with the underlying covariance structures in the data, the simulation study was also designed to look at whether the preferentially sampled nature of the data was likely to bias the mixing parameter estimation or the root mean squared errors. There did not appear to be any biasing effect on the mixing parameter, as there was no suggestion that any of the sampling schemes differed in terms of how often the lowest RMSPE was located at the true mixing parameter. However, the preferential sampling of the Tweed data did seem to have an effect on the RMSPEs. When compared to the full sample, which was regarded as something of a benchmark for a complete sampling structure, the Tweed sample and other sample seemed to show signs of bias. The other sample, as expected, always produced higher RMSPEs than the other two sampling schemes. The Tweed sample produced some very interesting, and somewhat surprising results. For mixing parameters of less than 0.5 or so (i.e. mixtures that favoured the

Euclidean covariance), the Tweed sample tended to produce higher RMSPEs than the full sample, whereas for mixtures greater than 0.5 or so the Tweed sample tended to have slightly lower errors. This suggests that the effect of having a preferentially sampled network may vary depending on the underlying covariance structure of the data, with mixtures closer to the tail-up stream distance model tending to produce lower errors. This is likely to be due to the slightly denser sampling and the extra information being provided due to the structure of the river (flow based weightings and flow connectivity). Therefore, it is likely that the impact of preferential sampling on the River Tweed analysis has led to bias in the reported root mean squared errors, but it is unlikely that the conclusions on the use of a mixture model or issues such as covariance structures were affected significantly.

6.1.3 Predicted Values on the River Tweed

Having assessed the accuracy of the various types of models that could be used to interpolate the Tweed data at unobserved locations, the next step was assessing these fitted values obtained using these methods. As with other analysis, this was done separately for Euclidean and stream distance based detrending methods. While it has been shown that there is very little difference between the detrending methods in terms of the accuracy of predictions, differences are slightly easier to see when observing the fitted values across the entire river networks. The differences tend to be in remote areas with few monitoring locations, where the stream distance approach tends to the mean more quickly as it generally has fewer nearby (flow connected) monitoring stations from which to predict. Again, there is no evidence for a preference between these approaches, and choice will come down to what one might expect of a river network.

As both detrending methods produced very similar results in the more densely monitored areas of the network, the conclusions for the nitrate levels across the river network are the same for each. In many monitored areas, there seems to be a very slight increase in nitrate levels over time between 1986 and the late 1990s, followed by a slight decline after 2000. The unusual year in this sense is 1998, which seems to be the peak of this increase. However, any change is very slight overall. The predicted nitrate levels also allow estimates to be made about which streams have levels above the limits suggested by the Nitrates Directive. Of most concern was the Turfford Burn area, which had one tributary with nitrate levels over 30 mg/l at its worst (with the upper limit set by the Nitrates Directive being 11.3 mg/l). All of the other tributaries around Turfford Burn displayed similarly high values, starting in 1989 with levels of 11.7 mg/l or more and mostly rising even higher by 1998. By 2005 these levels had fallen significantly though, with 14.6 mg/l being the maximum value on any tributary around this area. The only other area of any real concern was the Leet area, consisting of the stream joining the Tweed at Coldstream and its tributaries. This area is not as bad as Turfford Burn and, at its worst, nitrate levels are still mainly between the lower Nitrates Directive threshold (of 9.04 mg/l) and the upper threshold, with only a couple of stream segments slightly exceeding the upper. In general (and certainly towards the end of the monitoring period) the nitrate levels are below even the lower limit, though they are higher than the rest of the monitoring locations in general.

6.2 Spatio-Temporal Modelling

The final section of analysis of the River Tweed data focused on modelling the entire dataset through space and time. This was an area that had not been

covered much in the literature, especially with regards modelling the spatial component using the tail-up model structure. It was decided to tackle the problem using an additive model with standard seasonal and (temporal) trend components in addition to a novel stream distance based smooth function for the spatial trend, based on the tail-up covariance structure. This smoothing function was a modified version of that used in the stream distance based trend for spatial prediction.

In this kind of analysis, the major complicating factor is the presence of spatio-temporally correlated residuals, and the results from all models fitted to the Tweed data were adjusted to allow for this. It was unclear whether the residual correlation temporal correlation was particularly strong, as tests suggested that there was no residual correlation in the residuals from the majority of monitoring stations. Despite this, models were fitted to each and then combined in a separable spatio-temporal correlation structure that was then used to adjust the residual sums of squares, degrees of freedom and standard errors calculated from the various additive models.

The initial model with spatial trend, seasonal and temporal trend terms fitted the same seasonal component and temporal trend to every monitoring station and raised and lowered the mean value according to the spatial component. This was a good starting point, but looking at the fitted values it became clear that the one size fits all approach was insufficient for many of the monitoring stations. In order to allow the seasonal and trend components to vary over the network space, novel space-season and space-trend interaction terms, based on the stream distance smooth function, were added to the model and seemed to improve the model fit. Approximate F-tests were carried out in order to test the significance of the 2-interaction model compared to the two 1-interaction models and the main effects model, and these suggested that the 2-interaction model was a significant

improvement over all three simpler models. The season interaction seemed to be the most beneficial to the model as the season interaction model was a significant improvement on the main effects model while the trend interaction model was not. Despite this, the trend interaction was significant in addition to the season interaction in the 2-interaction model.

In terms of the fitted values, the 2-interaction model was seen to be very beneficial, when compared to the main effects model, to the fitted values at certain stations with a more pronounced seasonal component than the rest. This was an example of the season interaction improving the fit. Examples of the trend interaction doing the same were harder to see due to the lack of very strong trends in the Tweed data (at least when viewed on the log scale). Only one station showed a strong trend, and this was modelled much better by the interaction model. However, the lack of similar nearby trends meant that the fitted trend was smoothed somewhat and so was still not a particularly good fit to the data at certain locations. This did not tend to be so much of a problem when smoothing the seasonal components as stations with similar seasonal components tended to be clustered together.

Overall, the interaction model provides a good fit to the data and represents an interesting first step into modelling river network data using additive models.

6.3 Extensions

There are several possible extensions to the work on river networks that has been carried out.

In general, the more interesting extensions seem to be in the spatio-temporal modelling setting, but there are some improvements that could be made to the

spatial prediction work on the Tweed. Firstly, the (co)variogram estimation used seems to be one of the major weaknesses of this type of analysis. The approach used to model it in this work has focused on the empirical covariogram and estimation of it using weighted least squares. However, it would be very interesting to see what parameters were obtained when using likelihood based methods instead and comparing and contrasting the two. It would also be interesting to see what knock-on effects this had on the subsequent analysis, especially given that Cressie et al. (2006) concluded that a mixture model would not provide a more accurate covariance structure for their data based on maximum likelihood estimates. This covariogram modelling could also attempt to simultaneously estimate the parameters of both the Euclidean and stream distance components of the mixture model, as well as the mixing parameter λ. This has not been considered in the literature and comparison with the parameters estimated separately for the covariance structures would be very interesting.

One further extension to the analysis of spatial predictions on the River Tweed would be the inclusion of the tail-down covariance model structure on its own or as part of a mixture model. The tail-down model was not used in analysis as it was felt that it was unsuited to the context of nitrates data, but it would be interesting to see if this proved to be the case. The tail-down structure could be incorporated into the Tweed analysis, and perhaps could even include looking at another determinand that might be more suited to this structure. There are now several papers in the literature exploring the tail-down model using various other networks and determinands (Garreta et al., 2009; Money et al., 2009; Peterson and Ver Hoef, 2010; Ver Hoef and Peterson, 2010, for example, see) and so research in this area is already fairly advanced.

Most of the natural extensions to this work from the additive model as the

methods proposed represent one of the few examples in the literature of modelling river network data through space and time while using stream distance and keeping the flow connectivity and weighting that were so attractive in the tail-up model. Given the importance of the mixture model in earlier work, the most natural extension may be to experiment with the spatial smoothing terms (as both main effect and in the interaction) to see if a mixture model would provide a better fit to the data. This could also incorporate the tail-down structure in some way. However, given the fact that the spatial term is defining spatial trend and that there seemed to be little difference between the detrending methods (stream or Euclidean distance) in earlier analysis in terms of the final models, it is unclear what impact this is likely to have. In terms of spatial trend, it may seem reasonable to use just stream distance but the interaction terms may be suited to an interaction with a Euclidean term. Currently the model smooths the interaction terms over space, but the spatial smoothing still incorporates flow connectivity so that nearby but unconnected stations do not borrow weight from each other's seasonal components and trends in these interactions. It may be interesting to smooth the interaction terms over Euclidean space rather than stream distance, reflecting the fact that clusters of similar seasonal components seem to be similar due to surrounding land use.

It would also be interesting to see how the additive model would perform on a dataset with more temporal trends present in the data. As it stands in the Tweed data, the space-trend interaction does not have too much to do as there are few definite trends present in the data. It may be expected that if there were more trends present in the data, the space-trend interaction would adapt to fit these and certainly this should be the case. However, it is possible that the trends may not 'cluster' in the way that the seasonal components seem to in the Tweed data. If this was not the case and being close in space was no guarantee of a similar trend then it may be that this interaction term is redundant. At the moment

this is all hypothetical though, and so it would be interesting to see if this was the context of a river network.

Improvements could almost certainly be made to the smoothing terms used to fit the terms in the additive model. P-splines were initially considered instead of a local mean for the spatial term, however the added complexity of the river network structure seemed to make their use far more complex than normal. Rather than a certain (reasonably small) number of splines in total, it seemed that a river network would require a number of splines for each individual segment of the river, and these would then have to be able to 'split' (or merge, depending on point of view) at a point of confluence. One approach to this would be to adopt the same method used by the tail-up model to split the tails of the moving averages when moving back upstream (thus giving the name 'tail-up'). This would involve moving upstream through the river network and, at the points the river splits in two, the remaining tail of the spline splits into two pieces. This split could be done using flow or stream order data as in the tail-up model, so that more weight in the smooth function is given to the more prominent streams (which have higher flows). It is unclear how this could be accomplished in practice, and it may be likely that the computational efficiencies obtained by adopting a spline-based approach to smoothing may be lost due to the added complexity needed to adapt them to a river network.

It would also be very interesting to build more complex models using the current interaction model as a basis. This could hypothetically involve more covariates such as land use, rainfall data etc. On the River Tweed, there is limited availability of covariates, especially in the form of a time series. Most covariate data is available simply as a static figure for a region and so may struggle to account for any of the remaining variability. Most of the other data available as a time series on the River Tweed are other chemical determinands

such as phosphorous and dissolved oxygen, but chemical measures such as these do not seem suitable for use as explanatory variables. It would be very interesting to be able to take a dataset with covariates, such as rainfall which is linked to nitrate levels, available from at least one location but preferably at several sites around the network, and include them in such a model. This would allow a more descriptive model, which may suggest the factors behind the pollutant levels.

Following on from this, it is worth noting that the flow data used throughout analysis has been in the form of single, estimated average values for each monitoring location. It would be very interesting, as suggested in Cressie and O'Donnell (2010), to be able to dynamically alter the weightings in the tail-up covariance structure or the spatial smoothing term in the additive model according to the changing flows observed on the river. This would potentially negate the desire to include rainfall data as the increasing flows would reflect the additional runoff and likely increase in values that would be seen due to this, and vice versa. Sadly, the likelihood of obtaining such densely monitored flow data in addition to chemical determinands is probably quite low, but would make a very interesting example if available.

The analysis is also very computationally intensive in its current form and it is likely that further work could reduce this. The problem is mainly due to the interaction terms, which themselves are not too complex, but the use of the tail-up style spatial weighting means that it is difficult to bin or grid the data as is done in automatic additive modelling functions. In order to use this model on a wider scale, computational efficiencies would have to be made by finding a way to bin the data. In saying this, the Tweed analysis is fairly large scale with 21 years worth of data and over 80 monitoring stations with a total of over 11000 observations modelled through space and time (and the 11000×11000 correlation matrices required for that). However, this very much pushed the boundaries of

what was possible with the processing power available, and larger datasets would almost certainly need some form of computational efficiencies. Also, the available computing power for fitting the additive model was considerable, and for this analysis to be repeated on an ordinary PC, some sort of binning or gridding algorithm would be required to simplify the problem.

Finally, if improvements were made in the computational efficiency of the additive models it would be possible to predict at sever unsampled locations to produce predicted value plots over the entire network, such as those show for the spatial prediction in Chapter 4. This was considered for the two interaction model produced in Chapter 5 but, even with a hugely thinned out network for only a handful of snapshots in time, it was estimated that the processing time required would run into several weeks. Being able to produce such plots would be an excellent tool for the visualisation of the fitted additive model, especially if they could subsequently be turned into an animation.

Bibliography

Addiscott, T. and N. Benjamin (2004). Nitrate and human health. *Soil Use and Management 2*, 98104.

Akita, Y., G. Carter, and M. L. Serre (2007). Spatiotemporal nonattainment assessment of surface water tetrachloroethene in new jersey. *Journal of Environmental Quality 36*(2), 508–520),.

Banachiewicz, T. (1937). Zur berechnung der determinanten, wie auch der inversen, und zur darauf basierten auflosung der systeme linearer gleichungen. *Acta Astronomica, Seri C 3*, 41–67.

Barry, R. and J. M. Ver Hoef (1996). Blackbox kriging: spatial prediction without specifying the variogram. *Journal of Agricultural, Biological, and Environmental Statistics 1*, 297322.

Bernstein, D. (2005). *Matrix Mathematics*. Princeton University Press.

Boltz, H. (1923). Entwickelungs-verfahren zum ausgleichen geoddtischer netze nach der methode der kleinsten quadrate. veroffentlichungen des preussischen. *Geodatischen Institutes*.

Bowman, A., M. Giannitrapani, and E. M. Scott (2009). Spatiotemporal smoothingand sulphur dioxide trends over europe. *Applied Statistics 58*(5), 737–752),.

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

Bowman, A. W. and A. Azzalini (2007). *R package* `sm`*: nonparametric smoothing methods (version 2.2)*. University of Glasgow, UK and Universitá di Padova, Italia.

Brown, P. E., P. J. Diggle, M. E. Lord, and P. C. Young (2002). Space-time callibration of radar rainfall data. *Applied Statistics 50*(2), 221–241),.

Carroll, R. J. and D. Ruppert (1982). A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. *Journal of the American Statistical Association 75*, 878882),.

Cheney, E. W. and D. R. Kincaid (2007). *Numerical mathematics and computing*. Cengage Learning.

Clayton, J. (1997). The biology of the river tweed. *Science of the Total Environment 194/195*, 155162.

Clement, L. (2007). *Statistical Validation and spatio-temporal modelling of river monitoring networks*. Ph. D. thesis, Gent University, Gent, Belgium.

Clement, L. and O. Thas (2007). Spatio-temporal statistical models for river monitoring networks. *Journal of Agricultural, Biological, and Environmental Statistics 12*(2), 161–176),.

Clement, L., O. Thas, P. Vanrolleghem, and J. Ottoy (2006). Spatio-temporal statistical models for river monitoring networks. *Water Science and Technology 53*(1), 9–15),.

Cressie, N. (1985, July). Fitting variogram models by weighted least squares. *Mathematical Geology 17*(5), 563–586.

Cressie, N. (1991). *Statistics For Spatial Data*. Wiley.

Cressie, N., J. Frey, B. Harch, and M. Smith (2006, June). Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics 11*(2), 127–150.

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

Cressie, N. and D. O'Donnell (2010, March). Statistical dependence in stream networks. *Journal of the American Statistical Association 105*(489), 18–21.

Diblasi, A. and A. Bowman (2001, March). On the use of the variogram in checking for independence in spatial data. *Biometrics 57*(1), 211–218.

Diggle, P. J., R. Menezes, and T.-l. Su (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society, Series C 59*, 191–232.

European Parliament (1991, December). Council directive 1991/676/EEC, concerning the protection of waters against pollution caused by nitrates from agricultural sources. *Official Journal of the European Communities L375*, 0001–0008.

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.

Ferguson, C. A. (2007). *Univariate and Multivariate Statistiacl Methodologies for Lake Ecosystem Modelling*. Ph. D. thesis, The University of Glasgow.

Fuentes, M. (2006). Testing for separability of spatial-temporal covariance functions. *Journal of Statistical Planning and Inference 136*, 447–466),.

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

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

Garreta, V., P. Monestiez, and J. M. Ver Hoef (2009). Spatial modelling and prediction on river networks: up model, down model or hybrid? *Environmetrics*.

Giannitrapani, M. (2006). *Nonparametric Methodologies for Regression Models with Correlated Data*. Ph. D. thesis, The University of Glasgow.

Giannitrapani, M., A. Bowman, and M. Scott (2005). Additive models with correlated errors. Technical report.

Goodwin, T. H., A. Young, M. G. Holmes, H. Musgrave, and D. Pitson (2004). Development and assessment of methods to estimate flow statistics at the ungauged site for use within lf2000 scotland. Technical report.

Guillaume Blanchet, F., P. Legendrea, and D. Borcarda (2008). Modelling directional spatial processes in ecological data. *Ecological Modelling 15*, 325–336),.

Guttorp, P., W. Meiring, and P. D. Sampson (1994). A space-time analysis of ground-level ozone data. *Environmetrics 5*(3), 241–254),.

Hastie, T. and R. Tibshirani (1990). *Generalised Additive Models*. Chapman and Hall.

Horn, R. A. and C. R. Johnson (1985). *Matrix Analysis*. Cambridge University Press.

Huang, H.-C. and N. Cressie (1996). Spatio-temporal prediction of snow water equivalent using the kalman filter. *Computational Statistics and Data Analysis 22*(2), 159–175.

Jarvie, H., A. Wade, D. Butterfield, P. Whitehead, C. Tindall, W. Virtue, W. Dryburgh, and A. McGraw (2002). Modelling nitrogen dynamics and distributions in the river tweed, scotland: an application of the inca model. *Hydrology and Earth System Sciences 6*, 433–453.

Jobson, J. D. and W. A. Fuller (1980). Least squares estimation when the covariance matrix and parameter vector are functionally related. *Journal of the American Statistical Association 75*, 176181),.

Loader, C. (1999). *Local regression and likelihood*. Springer.

Loecher, M. (2011). *RgoogleMaps: Overlays on Google map tiles in R*. R package version 1.1.9.6.

McMullan, A. (2004). *Non-Linear and Nonparametric Modelling of Seasonal Environmental Data*. Ph. D. thesis, The University of Glasgow.

Miller, K. S. (1981). On the inverse of the sum of matrices. *Mathematics Magazine 54*, 67–72),.

Money, E., G. P. Carter, and M. L. Serre (2009). Using river distances in the space/time estimation of dissolved oxygen along two impaired river networks in new jersey. *Water Research 43*, 19481958),.

Peterson, E. E., A. A. Merton, D. M. Theobald, and N. Urquhart (2006). Patterns of spatial autocorrelation in streamwater chemistry. *Environmental Monitoring and Assessment 121*, 613636),.

Peterson, E. E., D. M. Theobald, and J. M. Ver Hoef (2007). Geostatistical

modelling on stream networks: developing valid covariance matrices based on hydrologic distance and stream flow. *Freshwater Biology 52*, 267279.

Peterson, E. E. and N. Urquhart (2006, October). Predicting water quality impaired stream segments using landscape-scale data and a regional geostatistical model: a case study in maryland. *Environmental Monitoring and Assessment 121*, 571–596),.

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

Priestley, M. (1981). *Spectral Analysis and Time Series*. Academic Press.

Reyjol, Y., P. Fischer, S. Lek, R. Rsch, and R. Eckmann (2005). Studying the spatiotemporal variation of the littoral fish community in a large prealpine lake, using self-organizing mapping. *Canadian Journal of Fisheries and Aquatic Sciences 62*(10), 22942302),.

Robson, A. and C. Neal (1997). Regional water quality of the river tweed. *Science of the Total Environment 194/195*, 173192.

Robson, A., C. Neal, J. Currie, W. Virtue, and A. Ringrose (1996). The water quality of the tweed and its tributaries. Technical report.

Romano, N. and C. Zeng (2007). Acute toxicity of sodium nitrate, potassium nitrate and potassium chloride and their effects on the hemolymph composition and gill structure of early juvenile blue swimmer crabs. *Environmental Toxicology and Chemistry 26*, 19551962.

Scottish Environment Protection Agency (2009a). Condition of the water environment in the tweed area. `http://www.sepa.org.uk/water/river_basin_`

`planning/area_advisory_groups/tweed/condition_and_objectives.aspx`.

Scottish Environment Protection Agency (2009b, January). Solway tweed draft river basin management plan environmental report. Technical report.

Shaddick, G. and J. Wakefield (2002). Modelling daily multivariate pollutant data at multiple sites. *Applied Statistics 51*(3), 351–372),.

Shreve, R. (1967). Infinite topographically random channel networks. *Journal of Geology 75*, 178–186.

Strahler, A. N. (1952). Hypsometric (area-altitude) analysis of erosional topography. *Geological Society of America Bulletin 63*(11), 1117–1141.

Theobald, D., J. Norman, E. E. Peterson, and S. Ferraz (2005). Functional linkage of watersheds and streams (flows): network-based arcgis tools to analyze freshwater ecosystems. In *Proceedings of the ESRI User Conference 2005*.

Thorp, J. H., M. C. Thom, and M. D. Delong (2006). The riverine ecosystem synthesis: Biocomplexity in river networks across space and time. *River Research and Applications 22*, 123–147),.

Tweed River Purification Board (1996). Tweed river purification board annual report annual report. Technical report.

Van Belle, G. and J. Hughes (1984). Nonparametric tests for trend in water quality. *Water Resources Research 20*, 127–136),.

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

Ver Hoef, J. M., E. Peterson, and D. Theobald (2006, December). Spatial statistical models thats use flow and stream distance. *Environmental and Ecological Statistics 13*(4), 449–464.

Webster, R. and M. A. Oliver (2001). *Geostatistics for Environmental Scientists*. Wiley.

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