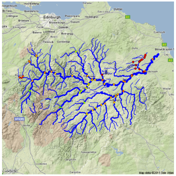
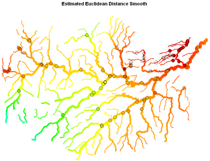


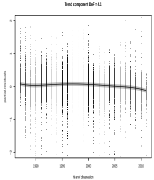

Going with the flow

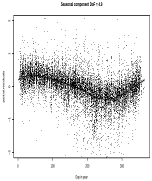
Regression models for river networks

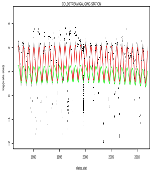
Adrian Bowman & Alastair Rushworth

University of Glasgow | School of Mathematics & Statistics







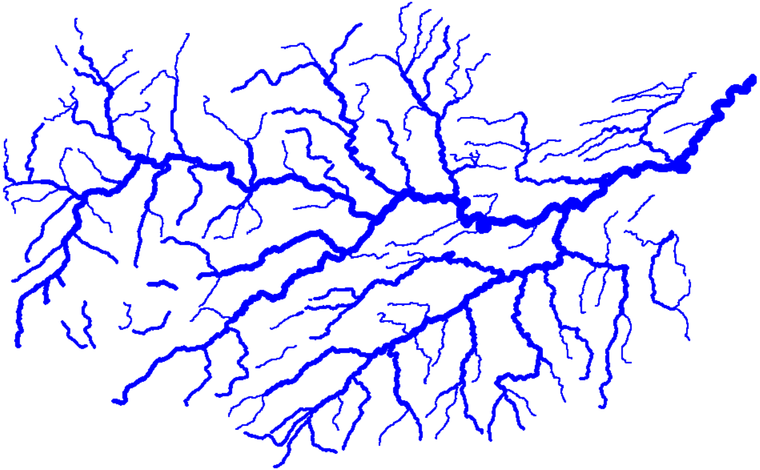
The River Tweed

David O'Donnell, Marian Scott, Mark Hallard

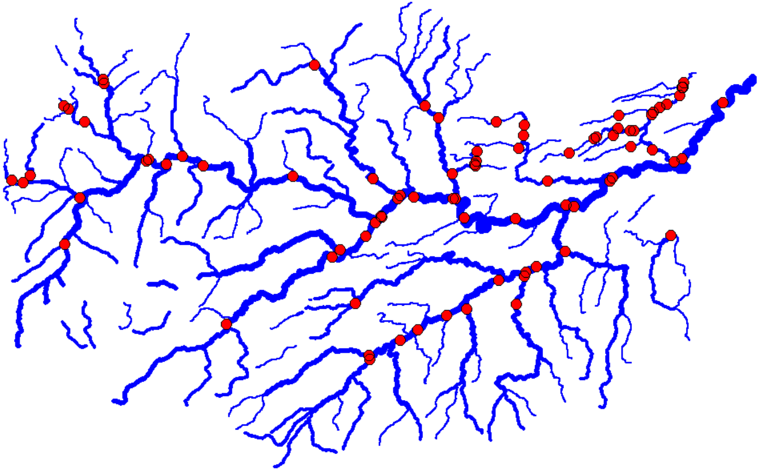
The River Tweed

The River Tweed

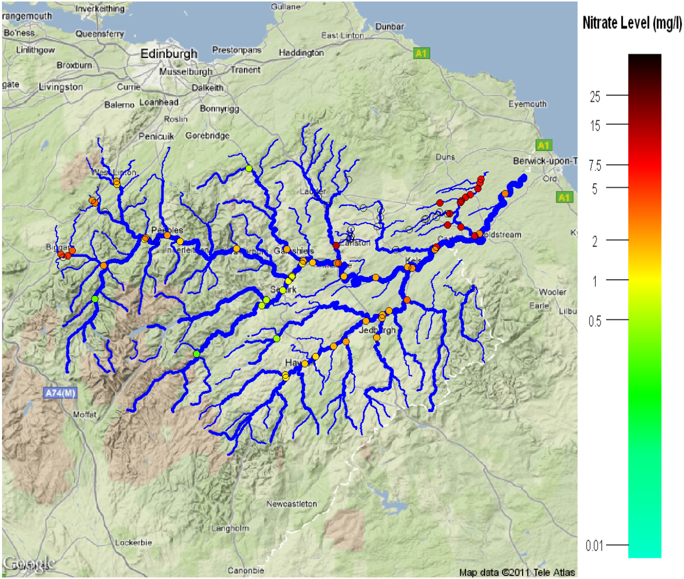
The River Tweed



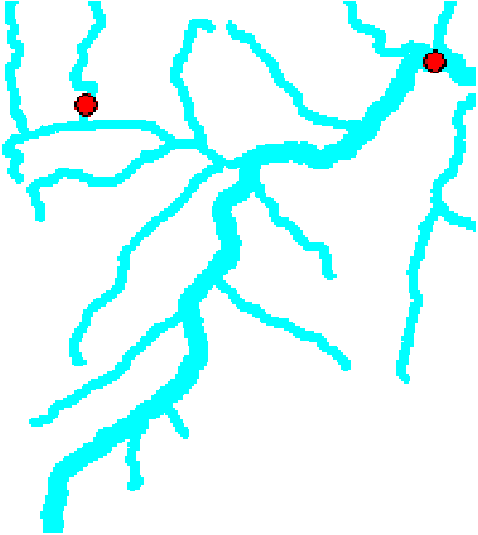
The River Tweed



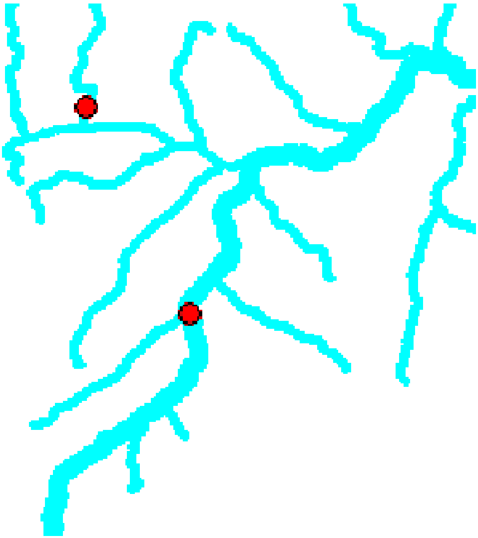
The River Tweed



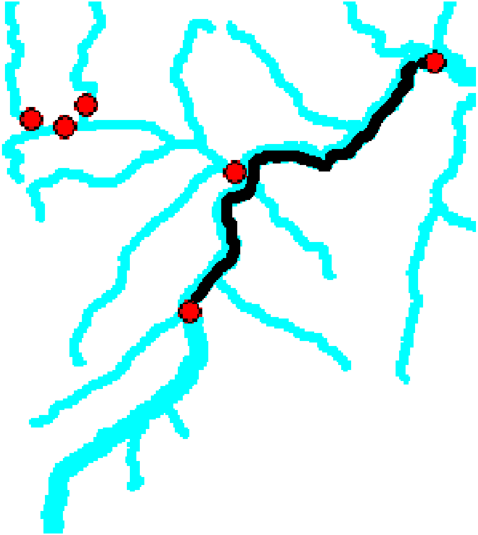
Flow connected



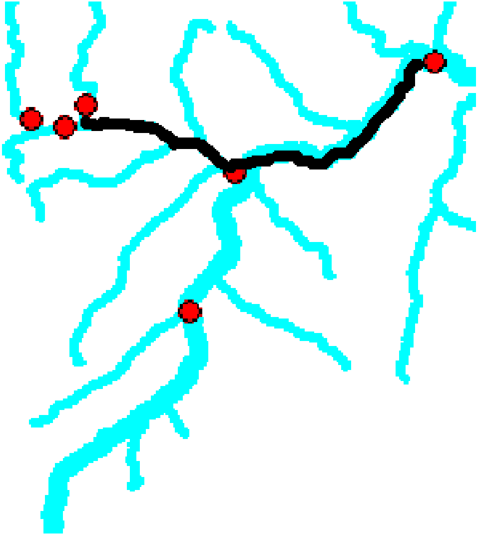
Not flow connected



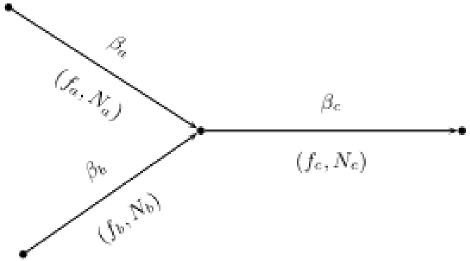
River distance



River distance



Confluences



The ver Hoef model

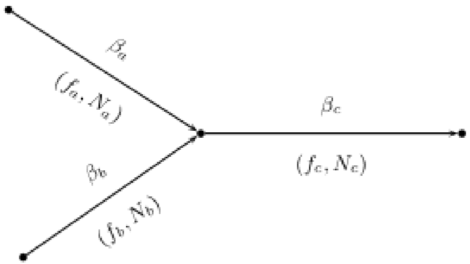
Covariance function

$$C(s_i, t_j) = \begin{cases} 0 & \text{if } s \text{ and } t \text{ are not flow connected;} \\ c_0 + c_1 & \text{if } s = t; \\ w \;\, c_1 \exp\left(-\frac{d_{s,t}}{c_2}\right) & \text{otherwise.} \end{cases}$$

where

- $d_{s,t}$ denotes the river distance between stations s and t;
- c_0, c_1, c_2 denote the nugget, partial sill and range parameters;
- $w = \prod_{k \in B_{s,t}} \sqrt{\omega_k}$
- $B_{s,t}$ denotes the set of all water stretches between s and t;
- ω_k denotes the proportion of flow contributed by water stretch k to its subsequent confluence.

Confluences



The $\sqrt{\omega}$ ensures that the variance in all water stretches is constant.

$$\mathrm{var}\{\sqrt{\omega_a} N_a + \sqrt{\omega_b} N_b\} = (\omega_a + \omega_b)\sigma^2 = \sigma^2,$$

where σ^2 denotes the variance of each measurement.

Flexible regression - weight functions

Data $\{(y_i, x_i), i = 1, \ldots, n\}$ can be modelled by a flexible regression

$$y_i = m(x_i) + \varepsilon_i,$$

where m denotes a smooth function and the ε_i denote error terms.

Flexible regression - weight functions

Data $\{(y_i, x_i), i = 1, \ldots, n\}$ can be modelled by a flexible regression

$$y_i = m(x_i) + \varepsilon_i,$$

where m denotes a smooth function and the ε_i denote error terms.

A very simple form of estimator is

$$\hat{m}(x) = \sum_i v_i y_i,$$

where the weights v_i decrease with distance from x.

Flexible regression - weight functions

Data $\{(y_i, x_i), i = 1, \ldots, n\}$ can be modelled by a flexible regression

$$y_i = m(x_i) + \varepsilon_i,$$

where m denotes a smooth function and the ε_i denote error terms.

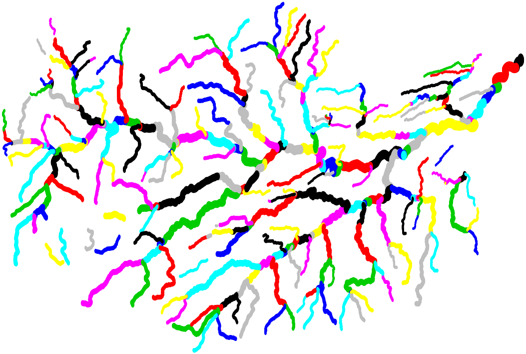
A very simple form of estimator is

$$\hat{m}(x) = \sum_i v_i y_i,$$

where the weights v_i decrease with distance from x.

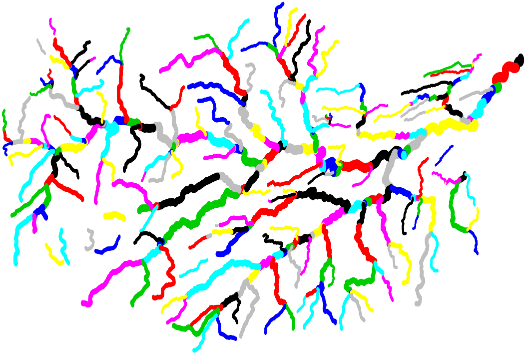
To adapt to a river network, we can use river distance and also incorporate the flow weights $w = \prod_{k \in B_{s,t}} \sqrt{\omega_k}$ into the weighting scheme.

Stream segments



An alternative approach to the estimation of smooth functions is to consider the network as a collection of small stream segments.

Stream segments

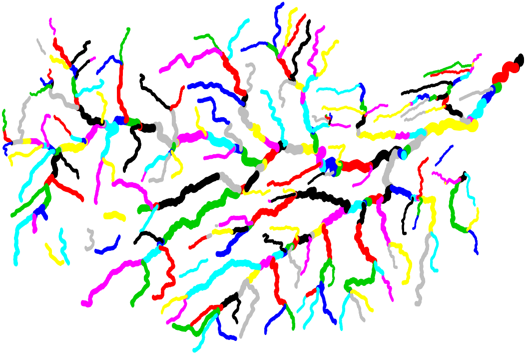


An alternative approach to the estimation of smooth functions is to consider the network as a collection of small stream segments.

An estimator is then available as

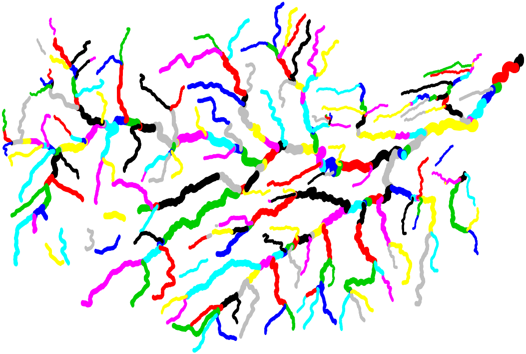
$$\hat{m}(x) = \beta_j, \quad \text{where } x \text{ lies in stream segment } j$$

Flexible regression - penalised splines



Formally, this is a b-spline approach of order 0.

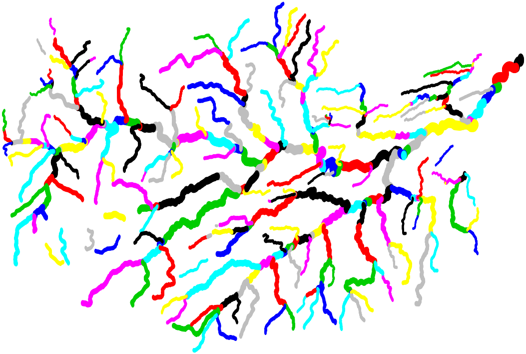
Flexible regression - penalised splines



Formally, this is a b-spline approach of order 0.

Smoothness is induced by use of a penalty, making this a *p-spline*.

Flexible regression - penalised splines

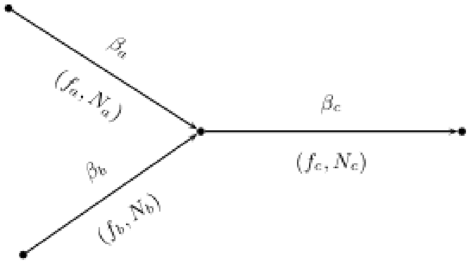


Formally, this is a b-spline approach of order 0.

Smoothness is induced by use of a penalty, making this a *p-spline*.

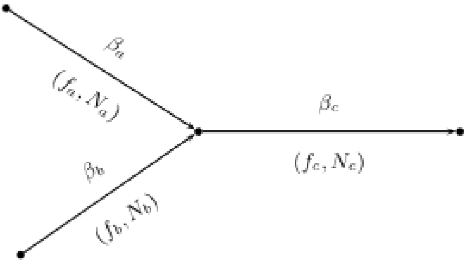
The 'smoothness' of β-values corresponding to adjacent stream units j and k, with no intervening confluence, can be measured by $(\beta_j - \beta_k)^2$.

Flexible regression - penalised splines



Where a confluence point is involved, the measure of smoothness needs to reflect the relative levels of flow in the contributing streams *a* and *b*.

Flexible regression - penalised splines

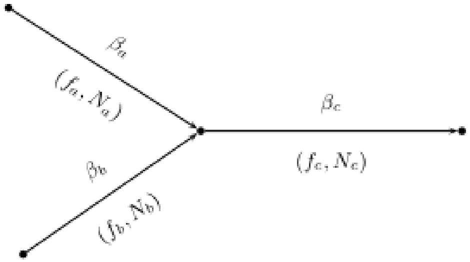


The relative flows of the inputs are $\omega_a = f_a/f_c$ and $\omega_b = f_b/f_c$.

The combined pollution input $\omega_a\beta_a + \omega_b\beta_b$ and output β_c are identical, following the principle of mass balance, if

$$\omega_a(\beta_a - \beta_c) + \omega_b(\beta_b - \beta_c) = 0.$$

Flexible regression - penalised splines



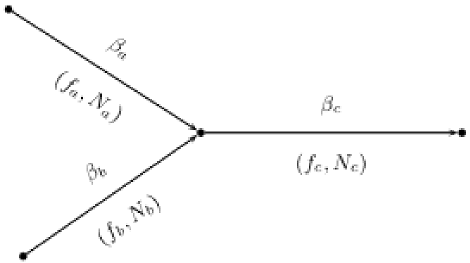
The combined pollution input $\omega_a\beta_a + \omega_b\beta_b$ and output β_c are identical, following the principle of mass balance, if

$$\omega_a(\beta_a - \beta_c) + \omega_b(\beta_b - \beta_c) = 0.$$

Smoothness across the confluence can therefore be achieved through the penalty

$$\omega_a^2(\beta_a - \beta_c)^2 + \omega_b^2(\beta_b - \beta_c)^2.$$

Flexible regression - penalised splines



Smoothness across the confluence can therefore be achieved through the penalty

$$\omega_a^2(\beta_a - \beta_c)^2 + \omega_b^2(\beta_b - \beta_c)^2.$$

This has the attractive form of combining penalties for smoothness across each flow path of the confluence, with weights determined by the relative volumes.

Flexible regression - penalised splines

- A p-spline model can be formulated as a regression model

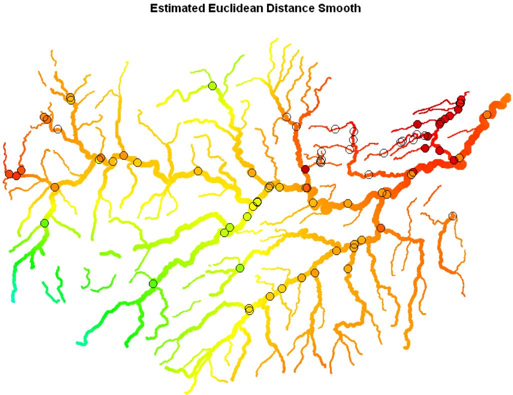
$$y = B\beta + \varepsilon,$$

  where the matrix B is simply an indicator matrix whose rows identify the stream unit containing y_i.
- The model is fitted by minimising the penalised sum-of-squares

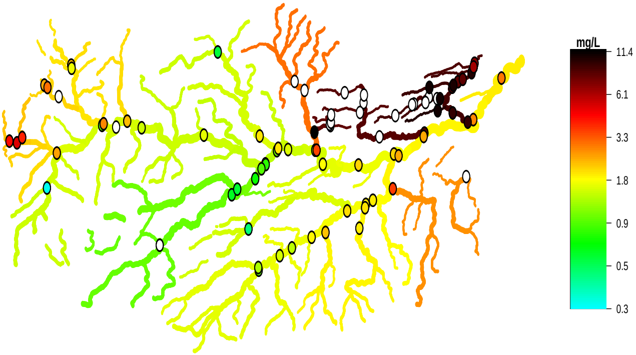
$$(y - B\beta)^T(y - B\beta) + \lambda\beta^T D^T D\beta$$

  with respect to β. The matrix D generates the penalty.
- The penalty parameter λ controls the degree of smoothing.
- The solution to this least squares problem is easily shown to be $\hat{\beta} = (B^T B + \lambda D^T D)^{-1} B^T y$.
- The linear form of this expression allows an approximate degrees of freedom to be computed as the trace of the 'hat' matrix.

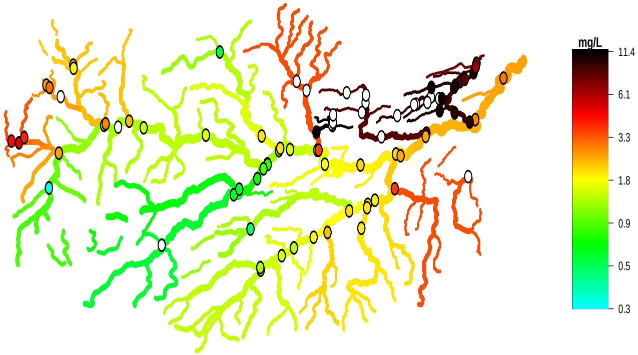
The River Tweed



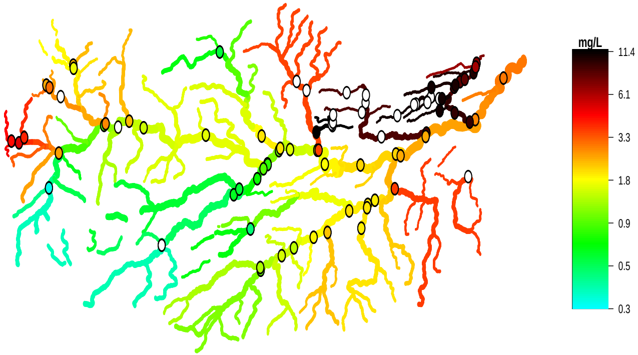
The River Tweed



The River Tweed



The River Tweed



Additive models

In order to incorporate time t_i and day of the year z_i,

$$y_i = \mu + m_s(s_i) + m_t(t_i) + m_z(z_i) + \varepsilon_i \tag{1}$$

where the three functions m_s, m_t, m_z describe spatial, temporal and seasonal trends..

Additive models

In order to incorporate time t_i and day of the year z_i,

$$y_i = \mu + m_s(s_i) + m_t(t_i) + m_z(z_i) + \varepsilon_i \tag{1}$$

where the three functions m_s, m_t, m_z describe spatial, temporal and seasonal trends..

- If each each of the trend functions is estimated by b-splines then they can be represented as $B_s\beta_s$, $B_t\beta_t$, $B_z\beta_z$ where the columns of the design matrices evaluate each basis function at the observed values of the relevant covariate.

Additive models

In order to incorporate time t_i and day of the year z_i,

$$y_i = \mu + m_s(s_i) + m_t(t_i) + m_z(z_i) + \varepsilon_i \quad (1)$$

where the three functions m_s, m_t, m_z describe spatial, temporal and seasonal trends..

- If each each of the trend functions is estimated by b-splines then they can be represented as $B_s\beta_s$, $B_t\beta_t$, $B_z\beta_z$ where the columns of the design matrices evaluate each basis function at the observed values of the relevant covariate.
- B-splines of order 0 can be used for the spatial network, while cubic b-splines would be a good choice for the temporal and seasonal effects, which operate in more standard sample spaces.

Additive models

In order to incorporate time t_i and day of the year z_i,

$$y_i = \mu + m_s(s_i) + m_t(t_i) + m_z(z_i) + \varepsilon_i \tag{1}$$

where the three functions m_s, m_t, m_z describe spatial, temporal and seasonal trends..

- If each each of the trend functions is estimated by b-splines then they can be represented as $B_s\beta_s$, $B_t\beta_t$, $B_z\beta_z$ where the columns of the design matrices evaluate each basis function at the observed values of the relevant covariate.
- B-splines of order 0 can be used for the spatial network, while cubic b-splines would be a good choice for the temporal and seasonal effects, which operate in more standard sample spaces.
- In the presence of an overall mean parameter μ, identifiability can be achieved simply be requiring the parameter vector for each term to sum to 0.

Additive models

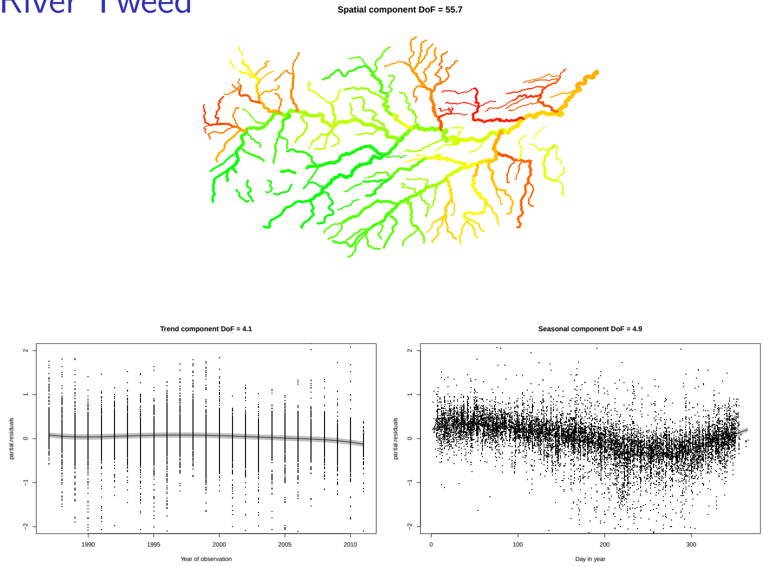
In order to incorporate time t_i and day of the year z_i,

$$y_i = \mu + m_s(s_i) + m_t(t_i) + m_z(z_i) + \varepsilon_i \quad (1)$$

where the three functions m_s, m_t, m_z describe spatial, temporal and seasonal trends..

- If each each of the trend functions is estimated by b-splines then they can be represented as $B_s\beta_s$, $B_t\beta_t$, $B_z\beta_z$ where the columns of the design matrices evaluate each basis function at the observed values of the relevant covariate.
- B-splines of order 0 can be used for the spatial network, while cubic b-splines would be a good choice for the temporal and seasonal effects, which operate in more standard sample spaces.
- In the presence of an overall mean parameter μ, identifiability can be achieved simply be requiring the parameter vector for each term to sum to 0.
- The full model can be represented as $y = B\beta + \varepsilon$, where B combines the columns of the individual design matrices.

The River Tweed



Interaction terms

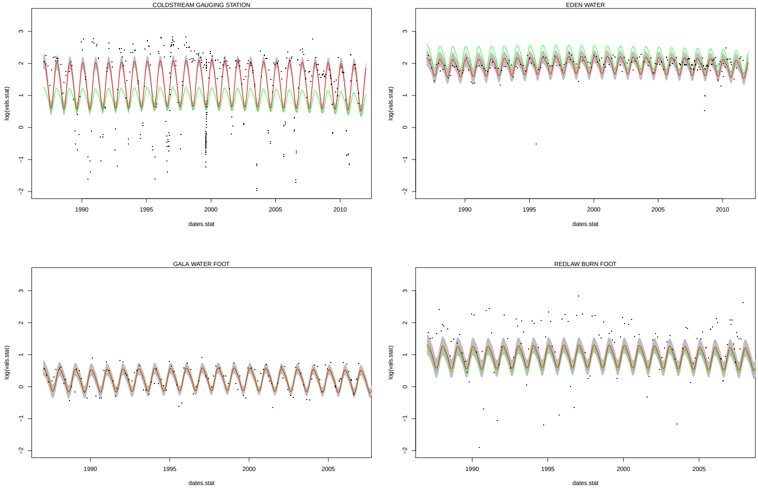
An interaction model has the form

$$y_i = \mu + m_s(s_i) + m_t(t_i) + m_z(z_i) + m_{s,t}(s_i, t_i) + m_{s,z}(s_i, z_i) + \varepsilon_i,$$

where the functions $m_{s,t}$ and $m_{s,z}$ encapsulate the adjustments required to capture the changes in time trend and seasonal effects over the river network.

The model can be fitted by exactly the same mechanism described above.

The River Tweed

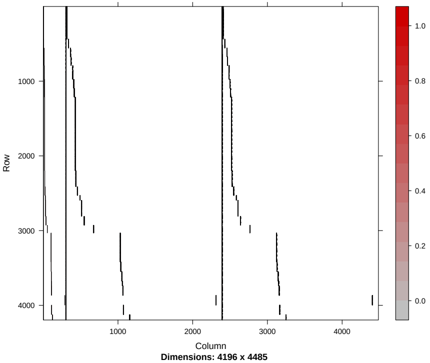


The River Tweed

Video

Computational issues

Calculations require the tensor product of an $n \times p$ matrix with n 1's, the rest 0's, which renders all model objects *very* sparse.



- The matrix $\mathbf{X}$ pictured above is 99.57% sparse.
- Using the R Matrix package of Bates and Mächler, $\mathbf{X}$ takes up only 1Mb of RAM and $\mathbf{X}^T\mathbf{X}$ takes 0.03s to calculate (compared to 150Mb and 66s uncompressed).
- Using a Cholesky factorisation further reduces calculation

Further research

- Seasonal-temporal interactions
- Residual autocorrelation
- Use of land-use covariate data
- Flow data (currently modelled values assumed constant)
- Bayesian heirarchical formulation - avoids smoothing parameter selection