Challenges in the statistical modelling of data on large river networks

Alastair Rushworth

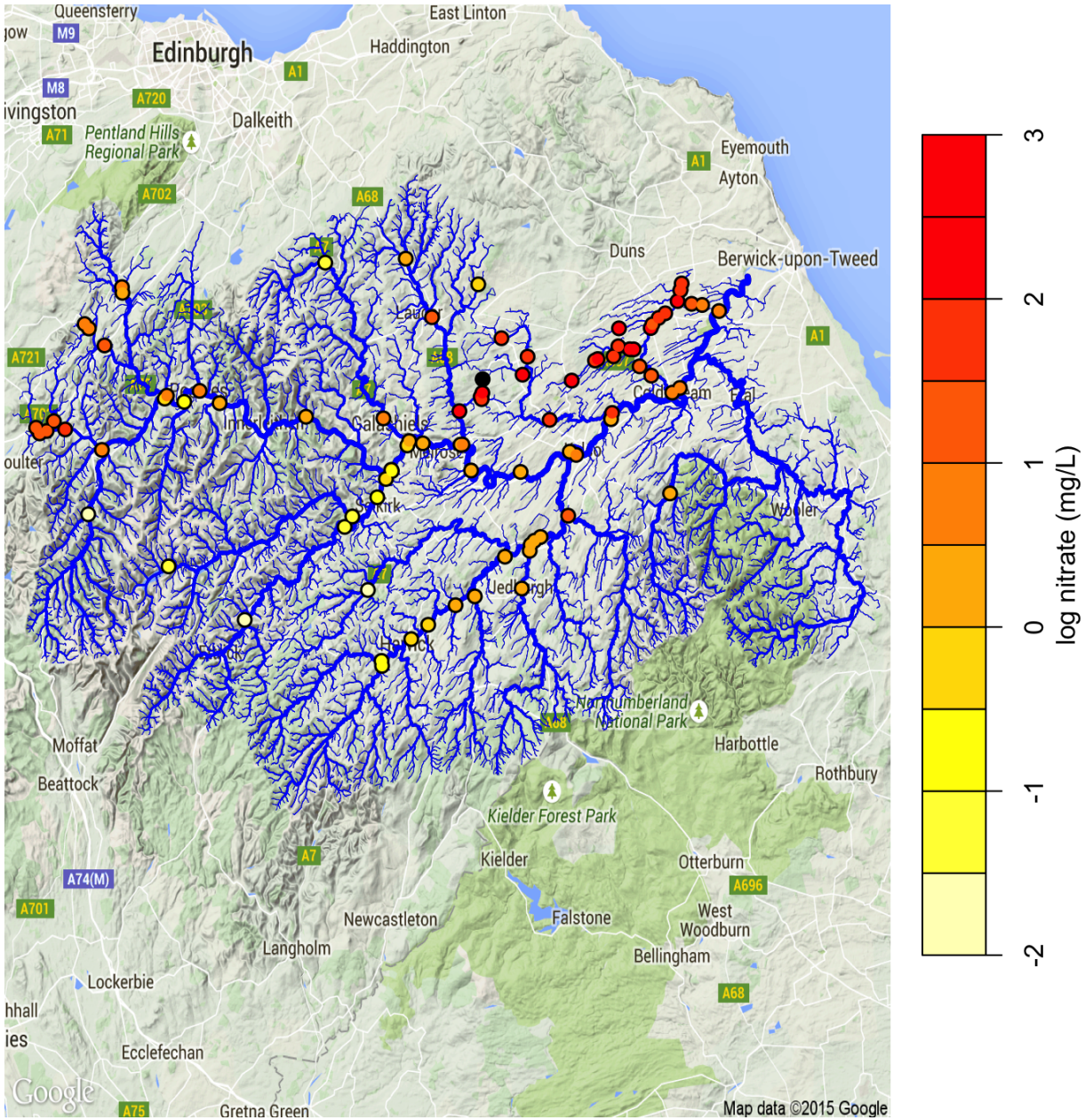
Basque Centre of Applied Mathematics

Tuesday 1st December, 2015

Talk outline

- **Special features of river networks**
- **Example: The River Tweed, Scotland**
- **Current approaches for spatial models**
- **Computational challenges and future developments**

River Tweed, Scotland

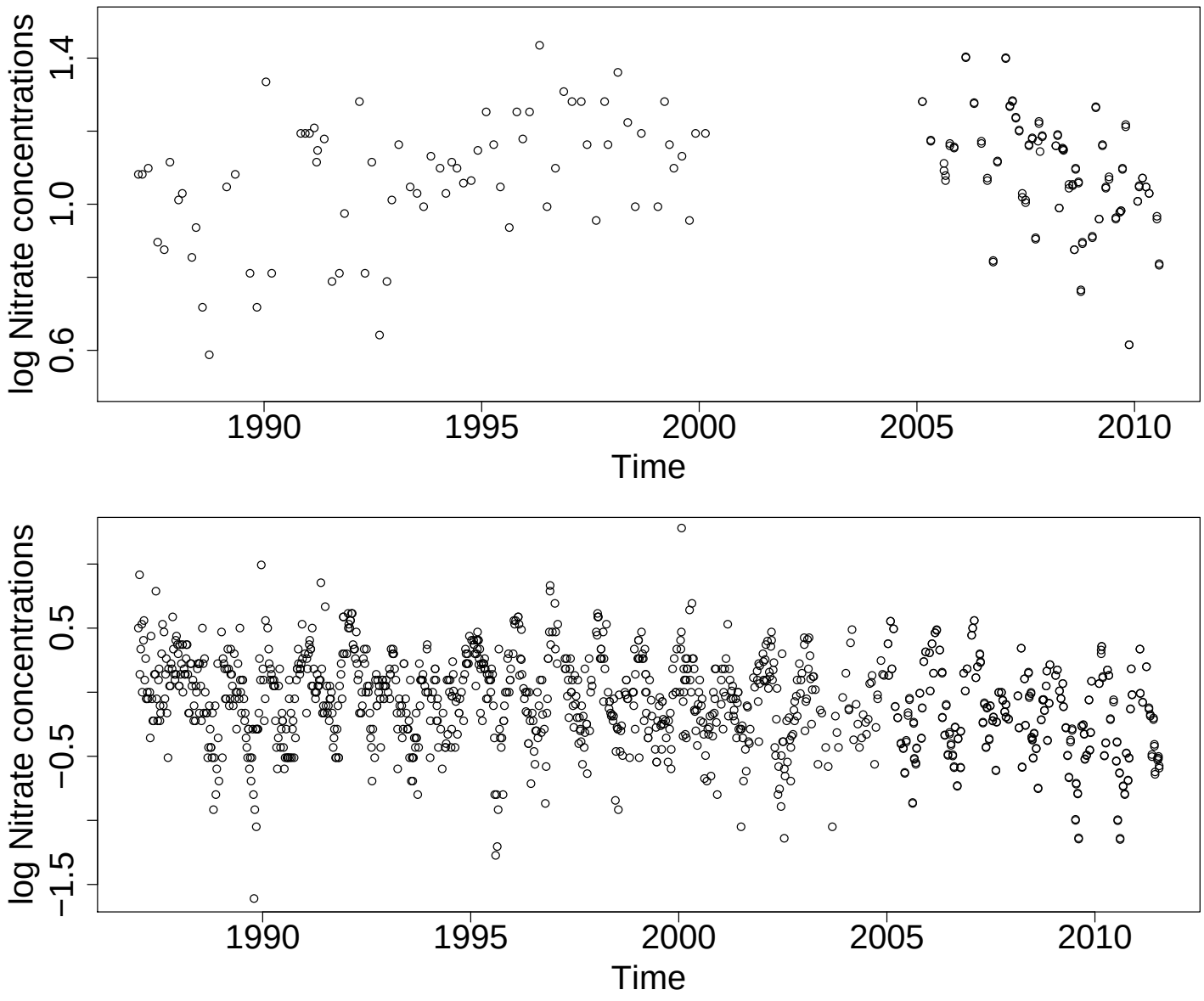


Goals of this work

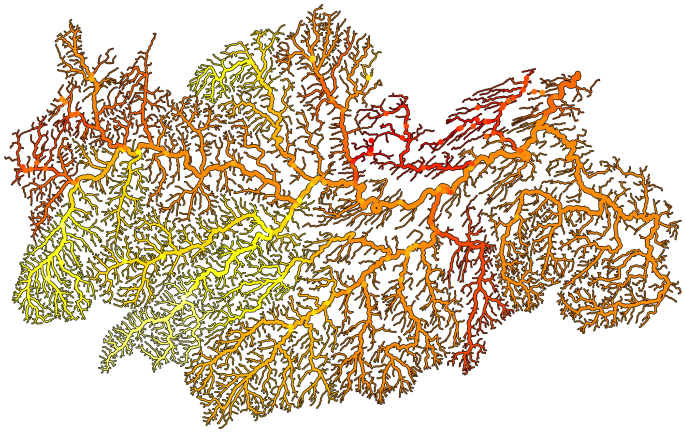
- **Long-term trend estimation**
- **Spatial prediction**
- **Fast and simple to use software**

```
# fit a penalised model with smooth trend and seasonal terms
tweed_model <- smnet(formula = log(Value_) ~ m(day, cyclic = T)
                     + m(year)
                     + network(adjacency, weight = "autoShreve"),
                     control     = list(approx = 100),
                     data.object = tweed,
                     netID       = 2)
```

Data on River Networks

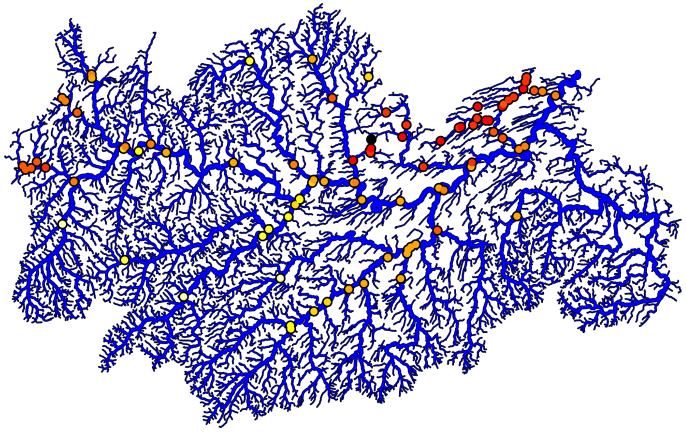


River Tweed, Scotland



Goals

- Estimation of dissolved pollutant at unmonitored locations
- Identify temporal trend
- Uncertainty in estimates
- Spatial prediction

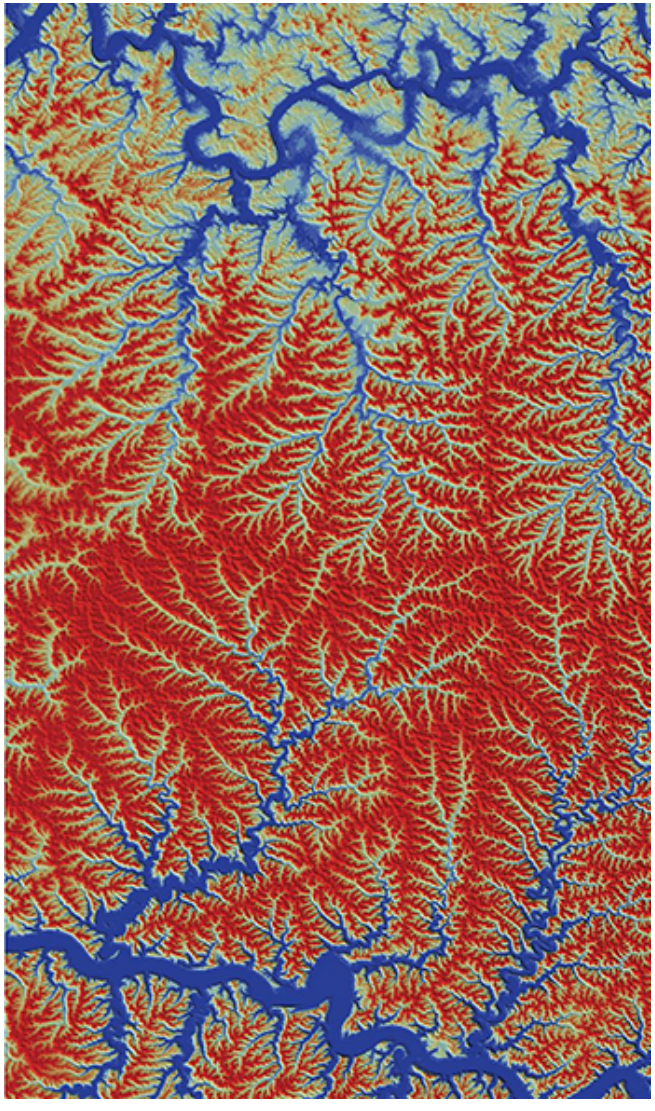


Challenges

- Data are often not small: $n >$15000. Over 9000 channels.
- Flow direction and volumes are important
- Seasonal, and other complex temporal variation

Spatial features: branching and flow

Image source: Taylor Perron / MIT



Spatial features: confluences

Two approaches for modelling river network data

1. Geostatistical approach using stream distance and flow

- Ver Hoef, Peterson and Theobald. (2006) *Spatial statistical models that use flow and stream distance*
- Good: detailed spatial structure can be captured
- Bad: computationally costly for large data

2. Penalised regression over stream segments

- O'Donnell, Rushworth, Bowman, Scott and Hallard. (2014) *Flexible regression over river networks*
- Good: fast, allows flexible model structure
- Bad: detail can be lost when small-scale variation present

Geostatistical framework

A traditional geostatistical model for observations made at locations (x_i, y_i) is

$$\begin{array}{rcccccc} Y(x_i, y_i) & = & \text{Covariates} & + & \text{Spatial process} & + & \text{error} \\ & = & \sum_{j=1}^{p} x_{ij}\beta_j & + & Z(x_i, y_i) & + & \epsilon_i \end{array}$$

The covariance matrix of Z is determined by a covariance function that depends on Euclidean distance seperation of observation points.

Z can be adapted for a river network...

Ver Hoef, Peterson & Theobald (2006)

$$\text{Cov}(r_k, s_l|\boldsymbol{\theta}) = \begin{cases} \pi_{k,l} C_t(h|\boldsymbol{\theta}) & \text{if } r_k \text{ and } s_l \text{ are flow-connected,} \\ 0 & \text{if } r_k \text{ and } s_l \text{ are flow-unconnected,} \end{cases}$$

Definitions:

1. **Stream segments** k and l determine membership of a particular channel on the network
2. **Upstream distances** r and s are defined along the path of the river from the network outlet
3. **Stream distance** $h = |r - s|$ is the separation between points r_k and s_l along the path of the river.

Care must be taken when choosing the covariance function C_t.

Ver Hoef, Peterson & Theobald (2006)

$$\text{Cov}(r_k, s_l|\boldsymbol{\theta}) = \begin{cases} \pi_{k,l} C_t(h|\boldsymbol{\theta}) & \text{if } r_k \text{ and } s_l \text{ are flow-connected,} \\ 0 & \text{if } r_k \text{ and } s_l \text{ are flow-unconnected,} \end{cases}$$

Properties

- C_t behaves like an ordinary spatial covariance function except using with stream distance as a seperation metric.
- Pairs of locations only correlated if they are connected by flow
- $\pi_{i,j} \in [0, 1]$ represents the effects of intervening flow that has 'diluting' effect on the covariance

Called the 'tail-up' model, since fitted values are spatially weighted averages of flow connected upstream observations, with weights that decrease for values further upstream.

Ver Hoef, Peterson & Theobald: summary

Powerful framework:

- Usual computational tools for obtaining estimates (here, REML)
- Non-Gaussian errors
- Flexible mixtures of covariance based on Euclidean and stream distance is handled

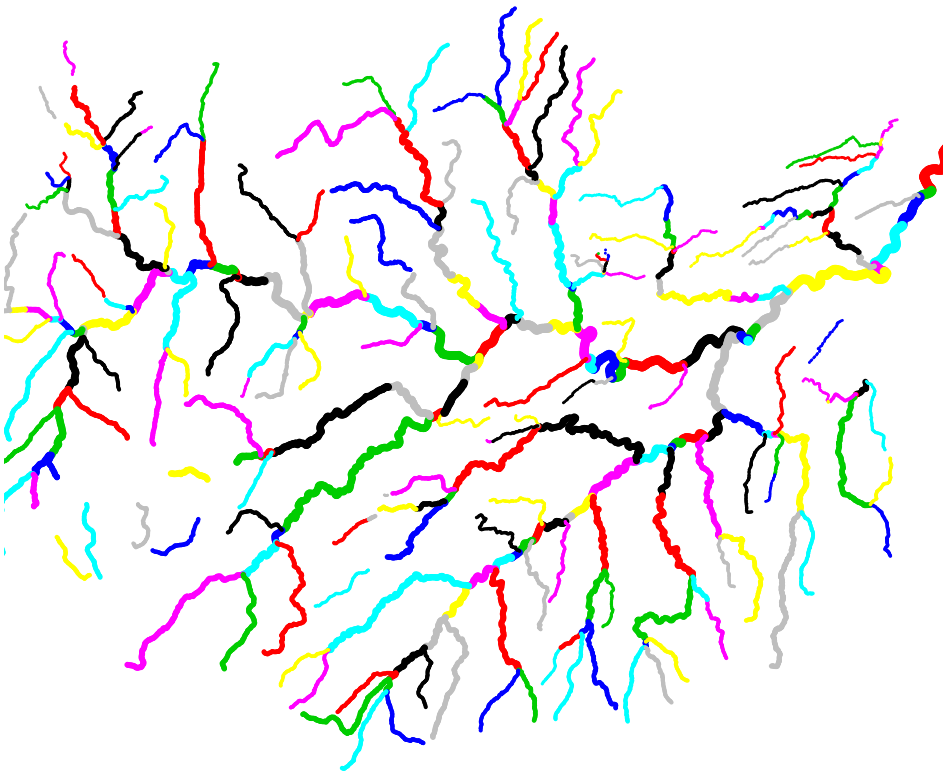
On the other hand:

- Working directly with a huge distance / covariance matrix is slow
- Spatio-temporal modelling is expensive
- Stationarity is required

Penalised regression offers an alternative to handling the distance matrix

Penalised regression approach - O' Donnell et al. (2014)

We can use the idea of using and enumerating stream segments to build **models for segments rather than points**.



- Each point belongs to a stream segment i, for which we associated a mean pollutant level, b_i.
- Many more segments to estimate than monitoring sites, so we must apply a penalty or constraint to estimate each b_i.

Building a model for stream segments

Suppose we have set of n observations, each observed at one of a set of p stream segments, then

$$E(\mathbf{y}) = \underbrace{\boldsymbol{\alpha}}_{\text{optional intercept}} + \underbrace{\mathbf{Xb}}_{p \text{ dummy variables}}$$

where

$$\mathbf{X} = \begin{pmatrix} 0 & 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Building a model for stream segments

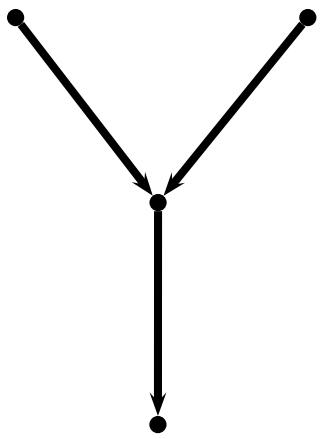
Suppose we have set of n observations, each observed at one of a set of p stream segments, then

$$E(\mathbf{y}) = \underbrace{\boldsymbol{\alpha}}_{\text{optional intercept}} + \underbrace{\mathbf{Xb}}_{p \text{ dummy variables}}$$

This could be ok, *but*

- Only observe a subset of the p segments
- Flow or stream order data provides critical information

Building a penalty for stream segments



An idealised confluence with stream units 1, 2 **and** 3, **associated mean pollutant levels** b_1, b_2 **and** b_3.

Values upstream
should be similar to those directly downstream
subject to their respective flow contributions:

$$\frac{f_1}{f_3} b_1 + \frac{f_2}{f_3} b_2 \approx b_3$$

Assuming that $f_1 + f_2 = f_3$,

$$\frac{f_1}{f_3} + \frac{f_2}{f_3} = 1$$

Intuitively attractive to rearrange these in terms of differences between upstream and downstream pollutant levels.

Derivation of the confluence penalty

$$\frac{f_1}{f_3} b_1 + \frac{f_2}{f_3} b_2 \approx b_3 \qquad \text{and} \qquad \frac{f_1}{f_3} + \frac{f_2}{f_3} = 1$$

so

$$\frac{f_1}{f_3} b_1 + \frac{f_2}{f_3} b_2 - b_3 \approx 0 \qquad \text{and} \qquad b_3 \left(\frac{f_1}{f_3} + \frac{f_2}{f_3} \right) = b_3$$

combining, we have

$$\frac{f_1}{f_3} b_1 + \frac{f_2}{f_3} b_2 - b_3 \left(\frac{f_1}{f_3} + \frac{f_2}{f_3} \right) \approx 0$$

rearranging,

$$\frac{f_1}{f_3} (b_1 - b_3) + \frac{f_2}{f_3} (b_2 - b_3) \approx 0$$

Building a model for stream segments

Let $\mathbf{y}$ be the vector of observed pollutant concentrations. A simple spatial model is $E(y_i) = \alpha + b_i$, and so

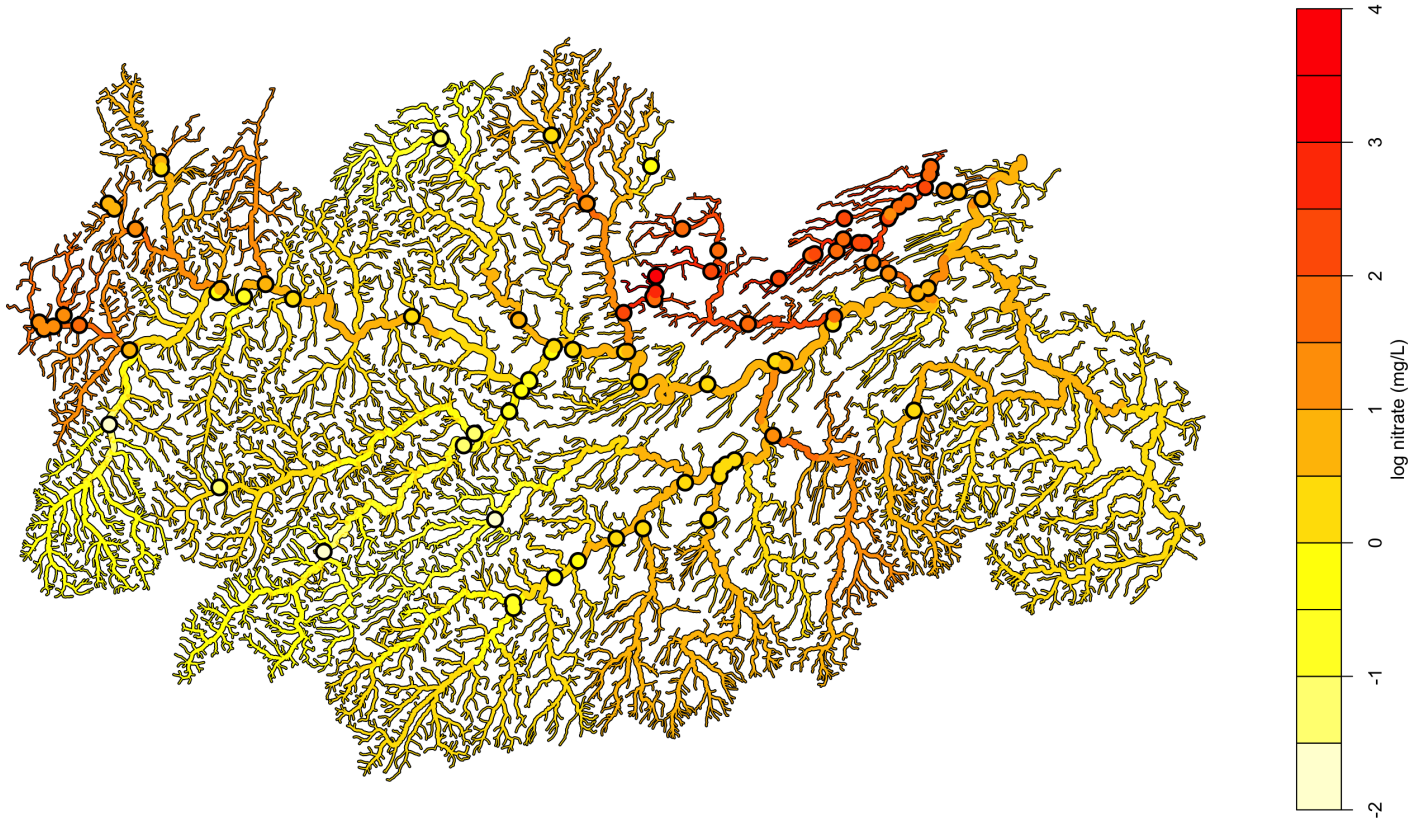
$$\mathbb{E}(\mathbf{y}) = \underbrace{\boldsymbol{\alpha}}_{\text{optional intercept}} + \underbrace{\mathbf{Xb}}_{\text{p dummy variables}}$$

where $\mathbf{X}$ is an indicator matrix identifying the 'stream unit' membership of each of the y_i. Then all we need is

$$\min_{\mathbf{b}} \left[(\mathbf{y} - \mathbf{Xb})^T (\mathbf{y} - \mathbf{Xb}) + \sum_{i,j \sim k} \lambda \left(\frac{f_i^2}{f_k^2} (b_i - b_k)^2 + \frac{f_j^2}{f_k^2} (b_j - b_k)^2 \right) \right]$$

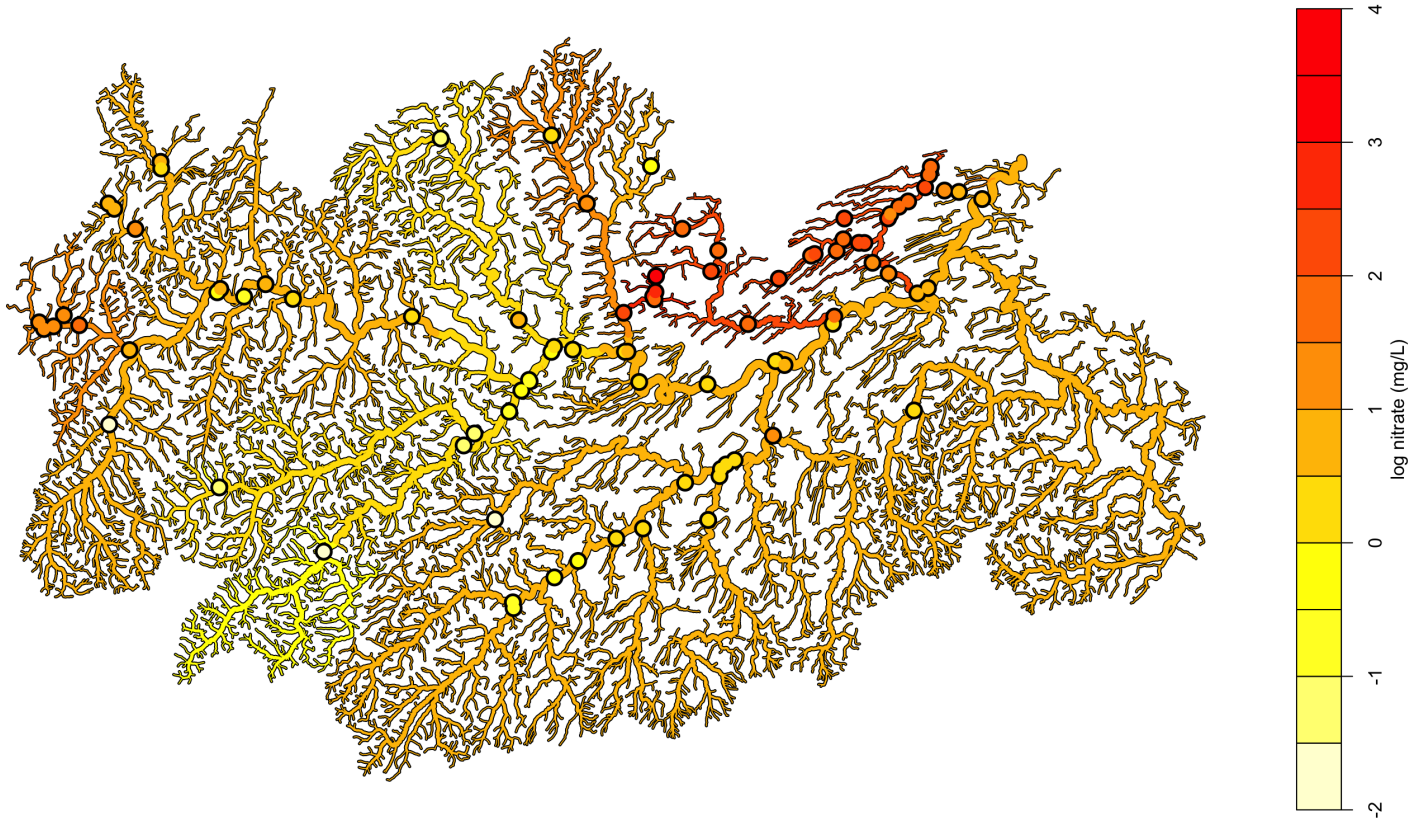
What happens if λ is varied?

Effect of varying the confluence penalty $\lambda = 0$



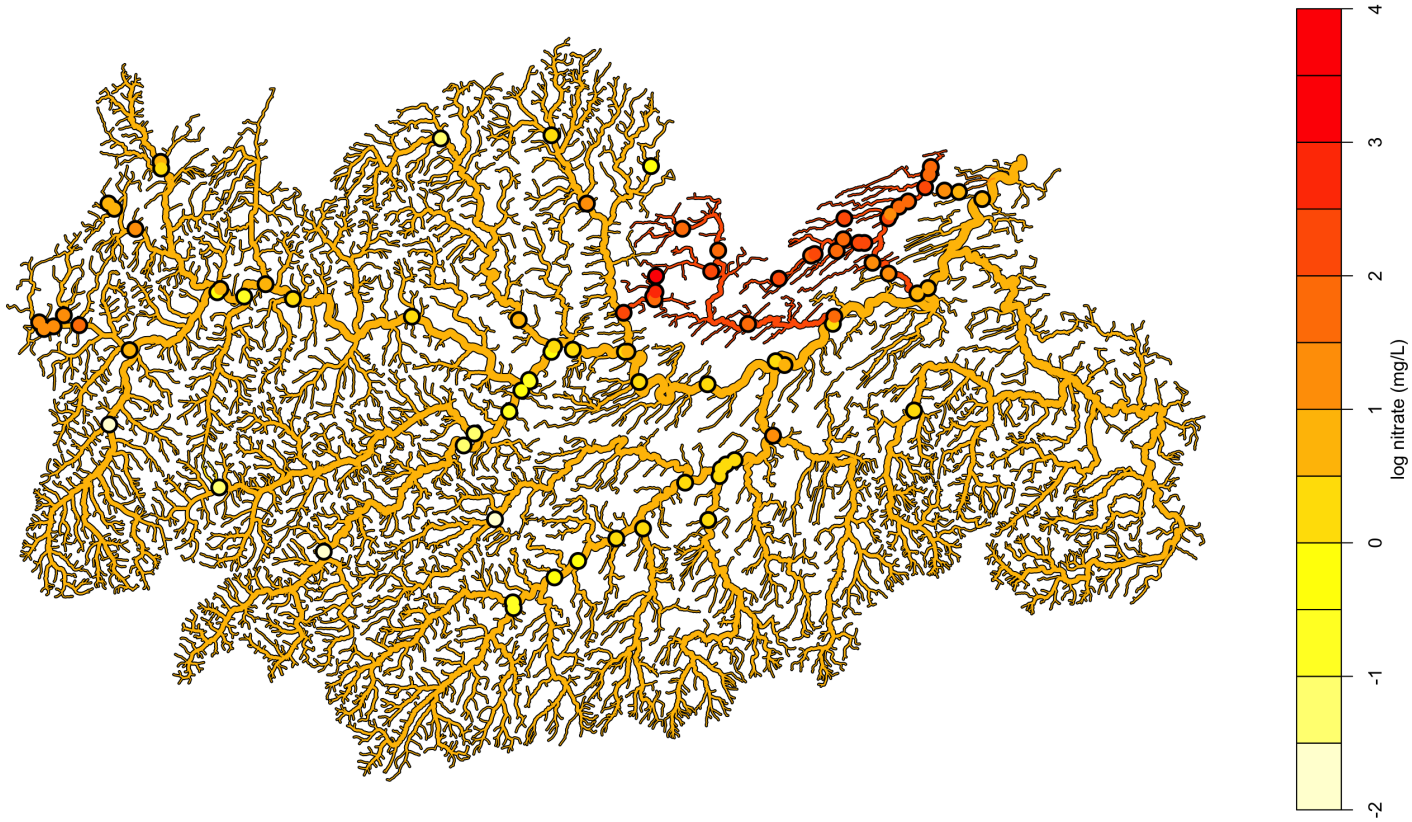
Weaker penalties interpolate or badly overfit the observed data

Effect of varying the confluence penalty $\lambda = +$



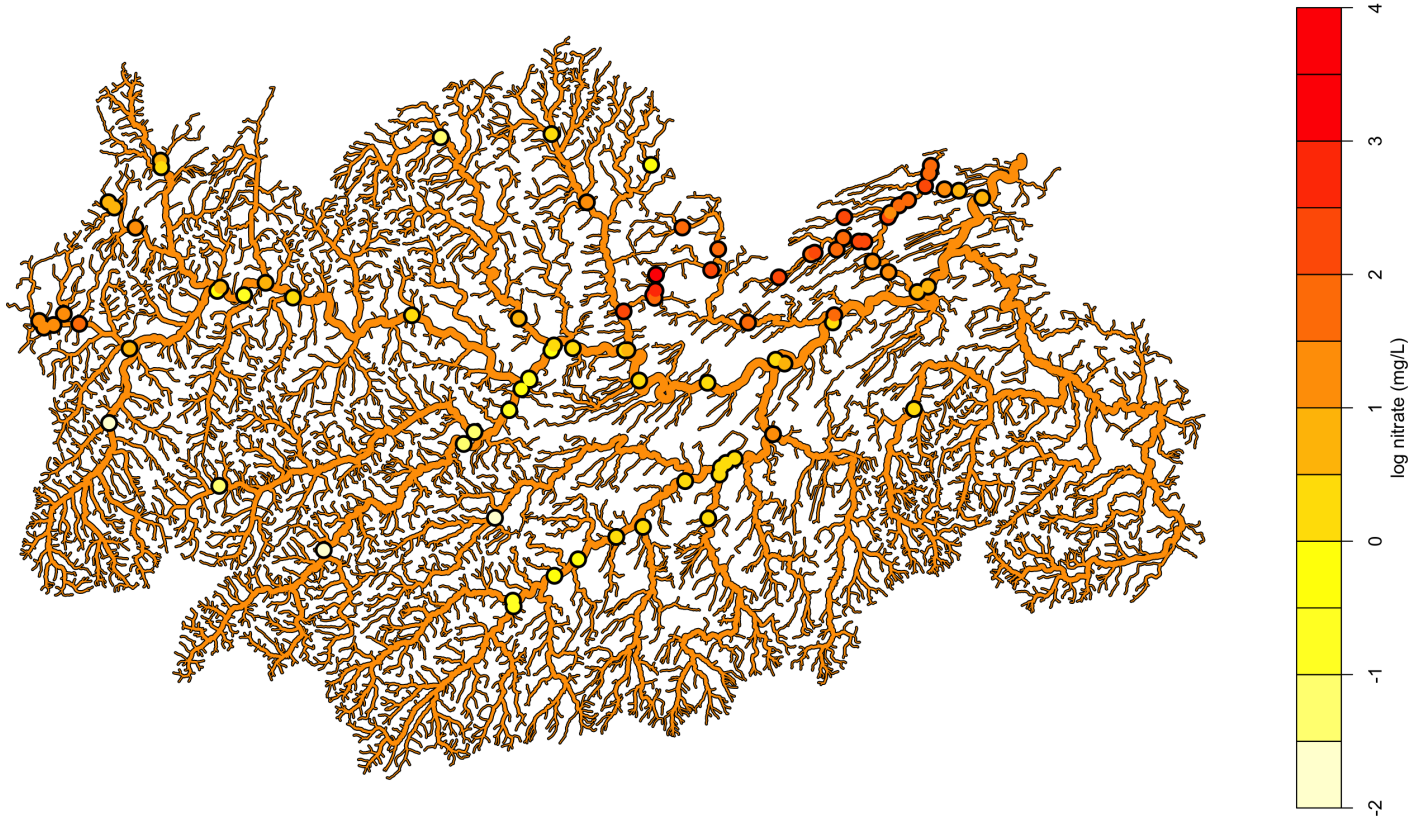
Weaker penalties interpolate or badly overfit the observed data

Effect of varying the confluence penalty $\lambda = +++$



Stronger penalties bring the estimates toward the *overall* mean

Effect of varying the confluence penalty $\lambda \to \infty$



A very strong penalty completely inhibits spatial variation in the estimates

Choosing good λ

In practise, one good way to choose λ would be to minimise some model fit criterion such as AICc:

$$\text{AICc} = \log(\hat{\sigma}^2) + 1 + \frac{2(\text{tr}(\boldsymbol{H}(\lambda)) + 1)}{n - \text{tr}(\boldsymbol{H}(\lambda)) - 2}$$

Some justification for this, since spatial prediction is an important goal.

$$\text{tr}(\boldsymbol{H}(\lambda)) \quad = \quad \text{tr}\left[\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\boldsymbol{D}^\top\boldsymbol{D})^{-1}\mathbf{X}^\top\right] \tag{1}$$

However, $\text{tr}(\boldsymbol{H}(\lambda))$ is expensive to calculate.

Getting tr($\boldsymbol{H}$) more cheaply

Hutchinson (1989) showed that

- For symmetric $\boldsymbol{A}_{n\times n}$
- and $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$, a vector of n independent samples from a random variable U
- $\mathbb{E}(U) = 0$ and $\mathrm{Var}(U) = \sigma^2$

$$\mathbb{E}(\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u}) = \sigma^2 \mathrm{tr}(\boldsymbol{A}) \tag{2}$$

Hutchinson showed that the U that results in the lowest possible variance estimator of tr($\boldsymbol{A}$) is a discrete random variable that takes the values $\{-1,\ 1\}$.

Getting tr($\boldsymbol{H}$) more cheaply

These results mean that we can approximate

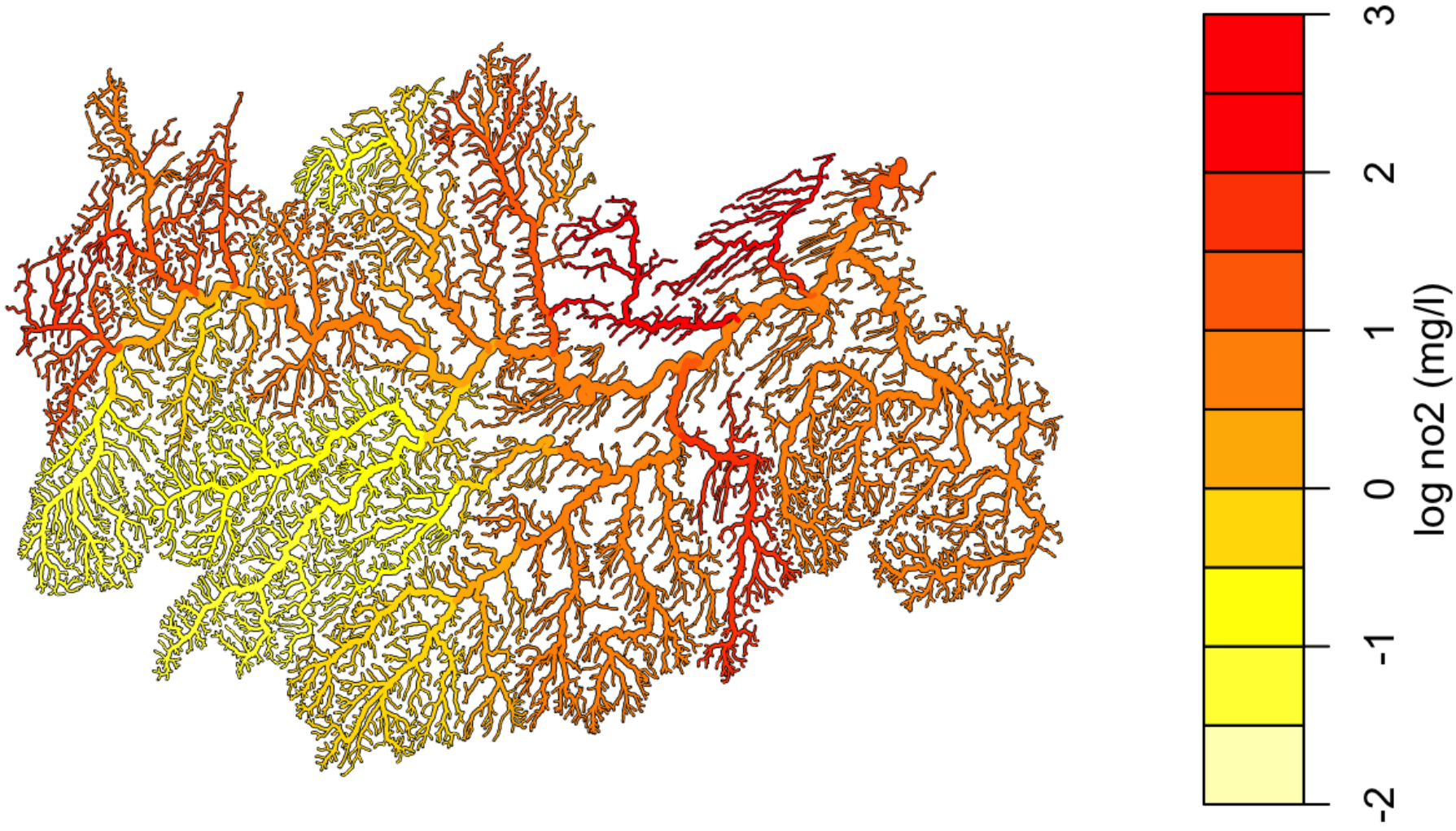
$$\begin{aligned} \text{tr}(\boldsymbol{H}(\lambda)) &= \text{tr}\left[\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\boldsymbol{D}^\top\boldsymbol{D})^{-1}\mathbf{X}^\top\right] \\ &\approx \frac{1}{s}\sum_{i=1}^{s}\boldsymbol{u}_i^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\boldsymbol{D}^\top\boldsymbol{D})^{-1}\mathbf{X}^\top\boldsymbol{u}_i \end{aligned}$$

- $\boldsymbol{u}_i^\top\mathbf{X} = (\mathbf{X}^\top\boldsymbol{u}_i)^\top$
- If s is reasonable large, calculate $\boldsymbol{u}_i^\top\mathbf{X}$ only once (for all possible λ), and resulting bias in tr($\boldsymbol{H}(\lambda)$) is usually small
- Use Cholesky $\boldsymbol{L}^{-1}\boldsymbol{L}^{-\top} = (\mathbf{X}^\top\mathbf{X} + \lambda\boldsymbol{D}^\top\boldsymbol{D})^{-1}$ for further speed up

Implemented in R package `smnet`.

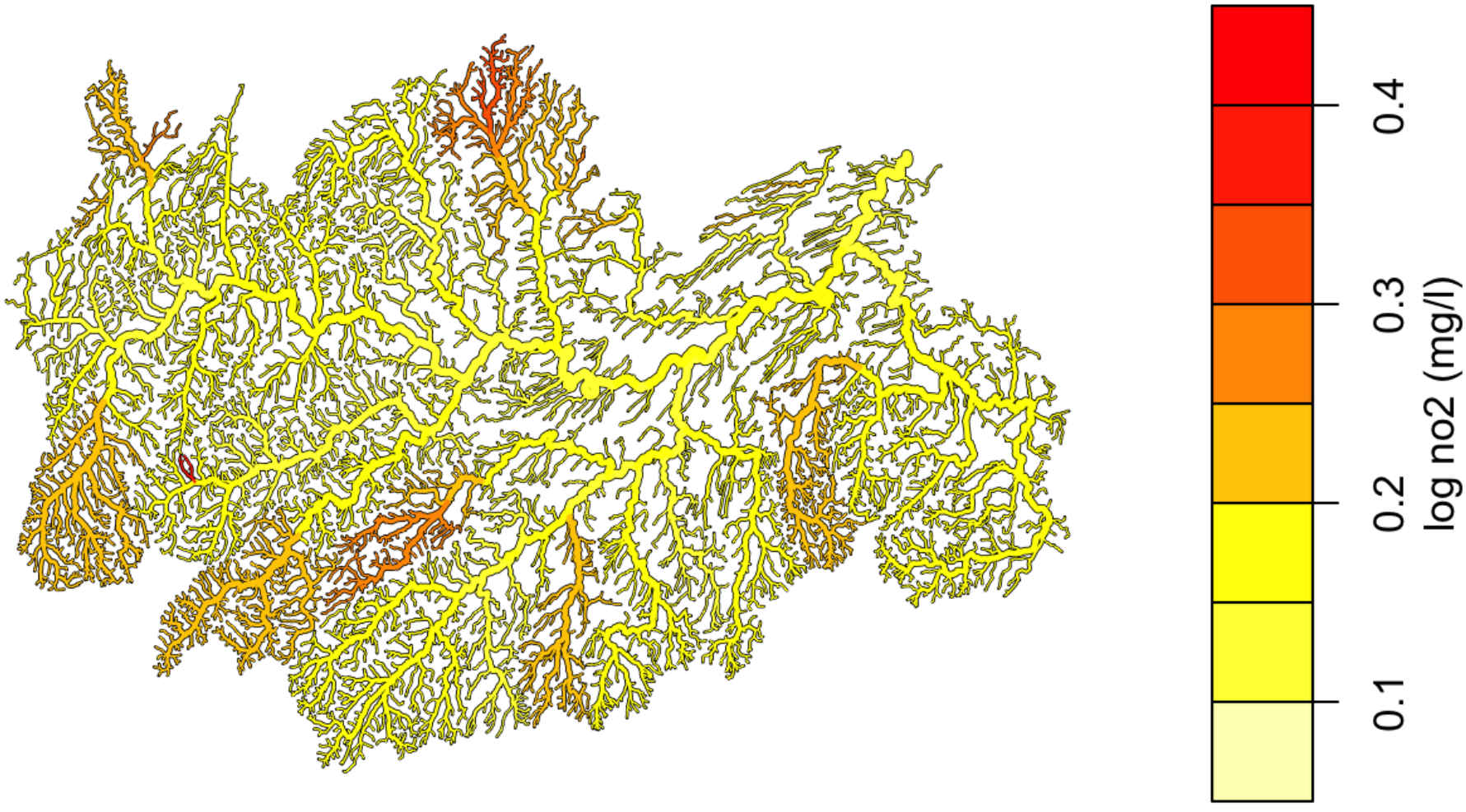
An example analysis using River Tweed data

Spatial fit



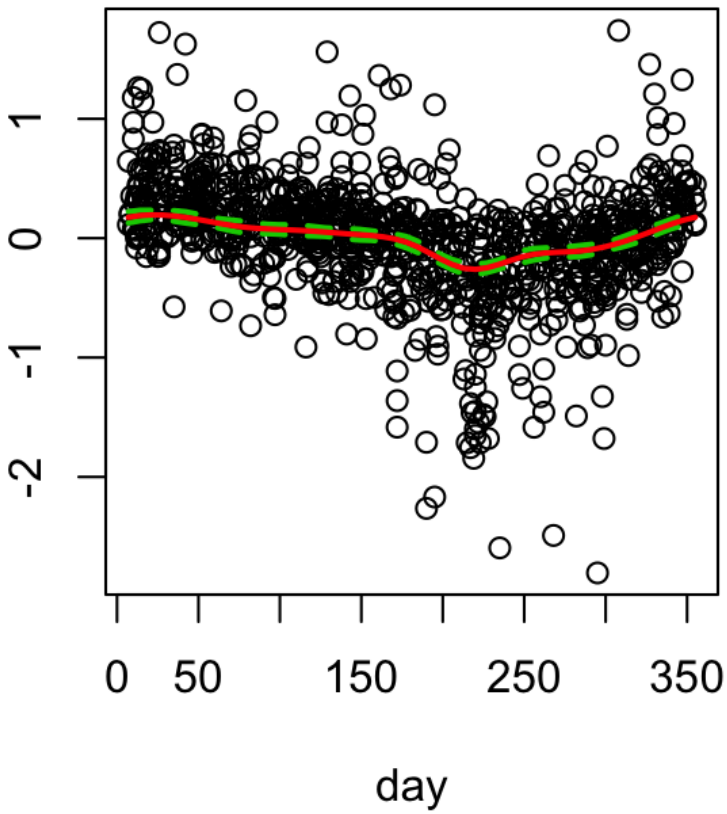
An example analysis using River Tweed data

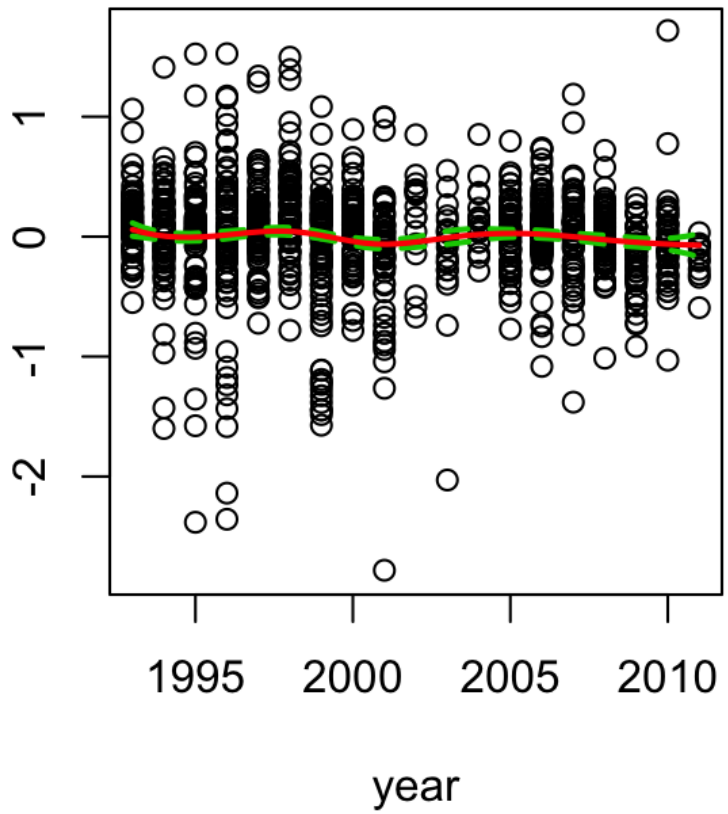
Spatial standard errors



An example analysis using River Tweed data

- Trend and seasonal pattern must be incorporated
- No reason to expect these are linear $\rightarrow$ use additive model
- Additive models permit smooth, non-linear effects of covariates





Comparison of the two approaches

Validation and comparison of geostatistical and spline models for spatial stream networks

Comparison of **geostatistical** and **penalised regression** was favourable (Rushworth et al. (2015)).

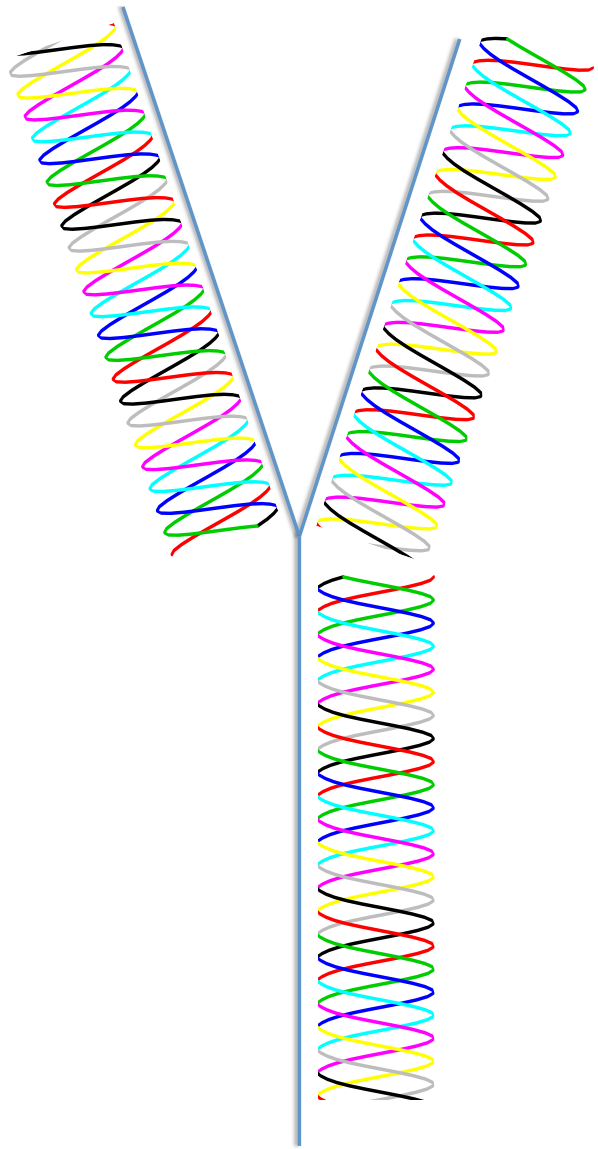
Main outcomes:

- Penalised model was very fast
- Large networks, sparsely sampled: results were close
- **Smaller networks, densely sampled: penalised model performed poorly**

This is an important point: many ecological studies involve higher sampling in smaller regions.

Comparison of the two approaches

- Smaller network with dense sampling, penalised regression performed poorly
- Possible solution: build B-spine basis along the channels of the network that respects flow.
- For larger networks, heavily tapered ‘tail-up’ model combined with PLS might be a good compromise that permits fast computation.



Current technical challenges

Flow contributions f_a, f_b and f_c are guessed, or roughly estimated.

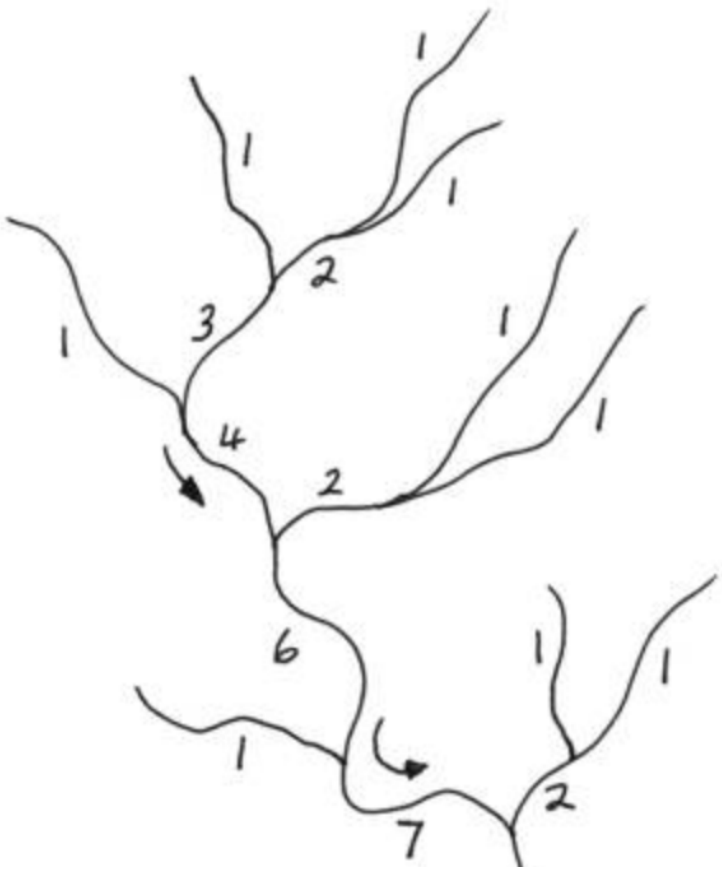


Figure: Illustration of Shreve order

- In reality, vary with time and are uncertain
- Important for space-time modelling with data at higher temporal resolution
- This becomes like an adaptive smoothing problem, heirarchical Bayes probably unavoidable and likely very intensive.

Current technical challenges

Modelling multiple networks

- Smoothing **on** and **between** networks
- Insight into national patterns
- Accelerated by huge increase in open data

AICc search complexity blows up with increasing numbers of λ.
Important for larger additive models.

- Estimation using REML is appealing, but mixed-model representation would break matrix sparsity (I think).
- Other options, backfitting, L-curve..

References

1. Ver Hoef JM, Peterson EE, Clifford D and Shah R. (2014). *SSN: An R package for spatial statistical modeling on stream networks.* Journal of Statistical Software 56(3).
2. Peterson, EE, et al. (2013) *Modelling dendritic ecological networks in space: an integrated network perspective.* Ecology Letters 16.5: 707-719.
3. O'Donnell D, Rushworth A, Bowman AW, Scott EM, Hallard M. (2014). *Flexible regression models over river networks.* Journal of the Royal Statistical Society: Series C (Applied Statistics) 63(1):47-63.
4. Ver Hoef JM, Peterson EE and Theobald D. (2006). *Spatial statistical models that use flow and stream distance.* Environmental and Ecological Statistics 13(4):449464.
5. Rushworth A, Peterson EE, Ver Hoef JM, Bowman AW. (2015) *Validation and comparison of geostatistical and spline models for spatial stream networks* Environmetrics 26(5):(327-338)