UNIVERSITEIT GENT

BIOSTAT
Department of Applied Mathematics,
Biometrics and Process Control

Spatio-temporal state-space models for river network data: two extension

L. Clement and O. Thas

19th Annual Conference of The International Environmetrics Society
TIES 2008, Kelowna, Canada, June 8-13, 2008

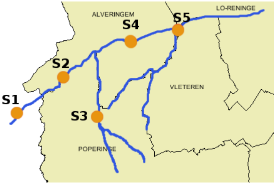
Outline

1. Introduction
2. Extension 1: Nonparametric trends in rivers
   1. Motivation & Modifications
   2. Implication on Parameter Estimation and Inference
   3. Case Study
3. Extension 2: Marginalised GLMM for River Networks
   1. Motivation
   2. Marginalised GLMM: Model Formulation
   3. Case Study
4. Conclusions and related research

State and Observation Model

State space representation

$$\begin{cases} \mathbf{S}_t = \mathbf{\Phi}\mathbf{S}_{t-1} + \boldsymbol{\delta}_t \\ \mathbf{Y}_t = \mathbf{X_t}\beta + \mathbf{S}_t + \boldsymbol{\epsilon}_t \end{cases}$$

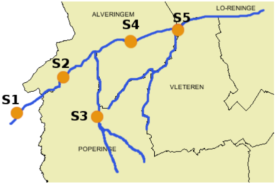


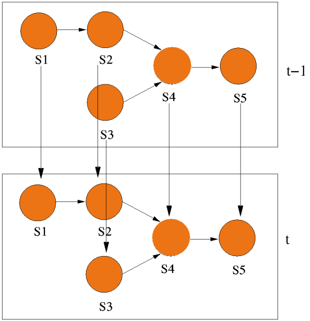
State and Observation Model

State space representation

$$\begin{cases} \mathbf{S}_t = \boldsymbol{\Phi}\mathbf{S}_{t-1} + \delta_t \\ \mathbf{Y}_t = \mathbf{X}_t\beta + \mathbf{S}_t + \epsilon_t \end{cases}$$

$$\mathbf{S}_t = \mathbf{A}\mathbf{S}_t + \mathbf{B}\mathbf{S}_{t-1} + \eta_t$$

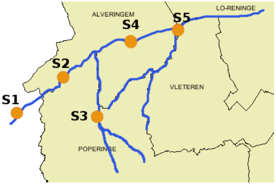




State and Observation Model

State space representation

$$\begin{cases} \mathbf{S}_t = \mathbf{\Phi}\mathbf{S}_{t-1} + \boldsymbol{\delta}_t \\ \mathbf{Y}_t = \mathbf{X_t}\boldsymbol{\beta} + \mathbf{f_t} + \mathbf{S}_t + \boldsymbol{\epsilon}_t \end{cases}$$

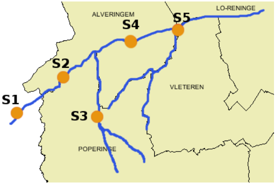


- Extension I: nonlinear trend

State and Observation Model

State space representation

$$\begin{cases} \mathbf{S}_t = \mathbf{\Phi}\mathbf{S}_{t-1} + \boldsymbol{\delta}_t \\ g\{\mathrm{E}\{\mathbf{Y}_t\}\} = \mathbf{X_t}\beta + \mathbf{S}_t \end{cases}$$



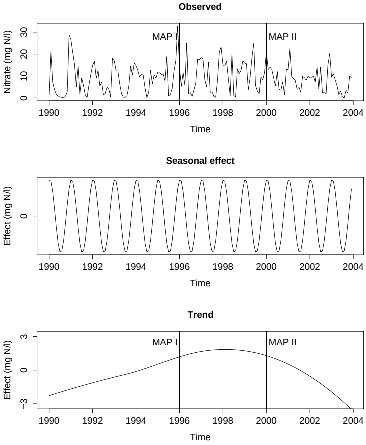
- Extension I: nonlinear trend
- Extension II: Response distributed according to other member of the exponential family

$\Rightarrow$ GLMM

Outline

1. Introduction
    1. Original Model
    2. Extensions
2. Extension 1: Nonparametric trends in rivers
    1. Motivation & Modifications
    2. Implication on Parameter Estimation and Inference
    3. Case Study
3. Extension 2: Marginalised GLMM for River Networks
    1. Motivation
    2. Marginalised GLMM: Model Formulation
    3. Case Study
4. Conclusions and related research

Motivation & Modification

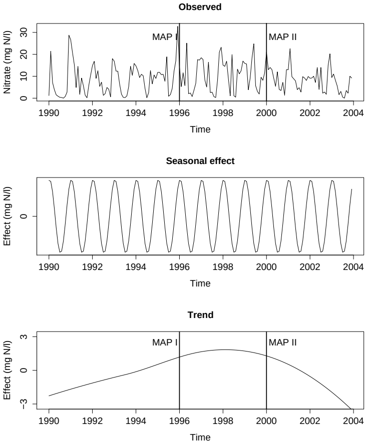


- Model:

$$\begin{cases} \mathbf{S}_t = \mathbf{\Phi}\mathbf{S}_{t-1} + \delta_t \\ \mathbf{Y}_t = \mathbf{X_t}\beta + \mathbf{S}_t + \epsilon_t \end{cases}$$

- Nitrate Data: ST + nonlinear trend

Motivation & Modification



- Model:

$$\begin{cases} \mathbf{S}_t = \mathbf{\Phi}\mathbf{S}_{t-1} + \delta_t \\ \mathbf{Y}_t = \mathbf{X_t}\beta + \mathbf{f(t)} + \mathbf{S}_t + \epsilon_t \end{cases}$$

- Nitrate Data: ST + nonlinear trend

- Adjust mean model

$$\mathrm{E}\left\{\mathbf{Y}_t\right\} = \mathbf{X}_t\beta + \mathbf{f(t)}$$

Fourier + smoother

1. Impact on P.E.
2. Inference on first derivative $\mathbf{f^{(1)}(t)}$
3. Multiplicity correction

Modifications to original ECM algorithm

1. ECM algorithm
   1. Choose initial estimates: $\mathbf{\Psi}^0$
   2. **E-step**: Calculate $Q(\mathbf{\Psi}, \mathbf{\Psi}_\alpha^k, \beta^k) = \mathrm{E}\left\{ l_c(\mathbf{\Psi}) | \mathbf{Y}_N, \mathbf{\Psi}^k \right\}$
   3. **CM-step 1**: Find the covariance parameters $\mathbf{\Psi}_\alpha^{k+1}$ that maximise $Q(\mathbf{\Psi}, \mathbf{\Psi}_\alpha^k, \beta^k)$
   4. **CM-step 2**: Find β^{k+1} that maximises $Q(\mathbf{\Psi}, \mathbf{\Psi}_\alpha^{k+1}, \beta^k)$
   5. Repeat steps 2-4 until convergence

Modifications to original ECM algorithm

1. ECM algorithm
   1. Choose initial estimates: $\mathbf{\Psi}^0$
   2. **E-step**: Calculate $Q(\mathbf{\Psi}, \mathbf{\Psi}_\alpha^k, \beta^k, \mathbf{f}^k) = \mathrm{E}\left\{l_c(\mathbf{\Psi})|\mathbf{Y}_N, \mathbf{\Psi}^k\right\}$
   3. **CM-step 1**: Find the covariance parameters $\mathbf{\Psi}_\alpha^{k+1}$ that maximise $Q(\mathbf{\Psi}, \mathbf{\Psi}_\alpha^k, \beta^k, \mathbf{f}^k)$
   4. **CM-step 2**: Find β^{k+1} and $\mathbf{f}^{k+1}$ that maximises $Q(\mathbf{\Psi}, \mathbf{\Psi}_\alpha^{k+1}, \beta^k, \mathbf{f}^k)$
   5. Repeat steps 2-4 until convergence

Modifications to original ECM algorithm

1. Estimate β and $\mathbf{f}$ by OLS: $\hat{\beta}$ and $\hat{\mathbf{f}}$
2. ECM algorithm
   1. Choose initial estimates: $\mathbf{\Psi}^0$
   2. **E-step**: Calculate $Q(\mathbf{\Psi}, \mathbf{\Psi}_\alpha^k, \hat{\beta}, \hat{\mathbf{f}}) = \mathrm{E}\left\{ l_c(\mathbf{\Psi}) | \mathbf{Y}_N, \mathbf{\Psi}^k \right\}$
   3. **CM-step 1**: Find the covariance parameters $\mathbf{\Psi}_\alpha^{k+1}$ that maximise $Q(\mathbf{\Psi}, \mathbf{\Psi}_\alpha^k, \hat{\beta}, \hat{\mathbf{f}})$
   4. **CM-step 2**: Redundant
   5. Repeat steps 3-4 until convergence

Modifications to original ECM algorithm

1. Estimate β and $\mathbf{f}$ by OLS: $\hat{\beta}$ and $\hat{\mathbf{f}}$
2. ECM algorithm $\Rightarrow$ EM on residuals of marginal mean model
    1. Choose initial estimates: $\mathbf{\Psi}^0$
    2. **E-step**: Calculate $Q(\mathbf{\Psi}, \mathbf{\Psi}_\alpha^k, \hat{\beta}, \hat{\mathbf{f}}) = \mathrm{E}\left\{l_c(\mathbf{\Psi})|\mathbf{Y}_N, \mathbf{\Psi}^k\right\}$
    3. **M-step**: Find the covariance parameters $\mathbf{\Psi}_\alpha^{k+1}$ that maximise $Q(\mathbf{\Psi}, \mathbf{\Psi}_\alpha^k, \hat{\beta}, \hat{\mathbf{f}})$
    4. Repeat steps 3-4 until convergence

$\Rightarrow$ Kalman Filter: $\mathbf{v_t} = (\mathbf{Y}_t - \mathbf{a}_{t|t-1} - \mathbf{X}_t\hat{\beta} - \hat{\mathbf{f}}(\mathbf{t}))$

$\Rightarrow$ CM-step 1: $\mathbf{Y}'_t = \mathbf{Y}_t - \mathbf{X_t}\hat{\beta} - \hat{\mathbf{f}}_(\mathbf{t})$

Mean model: parameter estimation by means of OLS

$$\mathrm{E}\left\{\mathbf{Y}_t\right\} = \mathbf{X}_t\beta + \mathbf{f(t)}$$

- OLS $\Leftrightarrow$ GLS (smoother matrix changes for every iteration $\Rightarrow$ computationally demanding)
- Estimate marginal mean: OLS (Hastie and Tibshirani, 1990)

$$\begin{cases} \hat{\beta} = (\mathbf{X}^T(\mathbf{I}-\mathbf{S}_f)\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I}-\mathbf{S}_f)\mathbf{Y} \\ \hat{\mathbf{f}} = \mathbf{S_f}(\mathbf{Y}-\mathbf{X}\hat{\beta}) = \mathbf{H_f}\mathbf{Y} \end{cases}$$

where $\mathbf{S_f}$ is the smoother matrix and
$\mathbf{H_f}$ is the projection matrix

Inference

- Assess local trends using $\mathbf{f}(t)$
- Tests on first derivative $\mathbf{f}^{(1)}(t)$ ($\mathbf{\Sigma}_{f^{(1)}} = \mathbf{H}_{f^{(1)}}\mathbf{\Sigma}_{Y_N}\mathbf{H}^T_{f^{(1)}}$)
- Many simultaneous tests
- Tests are dependent: classical multiplicity corrections to conservative
- Incorporate dependence between the tests explicitly: Adapt free step-down resampling method (algorithm 2.8 of Westfall and Young 1993)

Trend test: Multiplicity correction

1. Rank original p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$
2. Initialise the count variables: $COUNT_i = 0, i = 1, \ldots, n$
3. Generate a vector $(p^*_{(1)}, \ldots, p^*_{(n)})$ under H_0. (Note that sequence $\{(j)\}$ is fixed).
4. Successive minima to enforce the same monotonicity

$$q^*_n = p^*_{(n)}, \ldots, q^*_{n-i} = \min(q^*_{n-i+1}, p^*_{(n-i)}), \ldots, q^*_1 = \min(q^*_2, p^*_{(1)}).$$

5. If $q^*_i \leq p_{(i)}$, then $COUNT_i = COUNT_i + 1$.
6. Repeat (3)-(5) B times, adjusted p-values: $\overset{\sim(B)}{p}_{(i)} = \frac{COUNT_i}{B}$.
7. Enforce monotonicity of $\overset{\sim(B)}{p}_{(i)}$

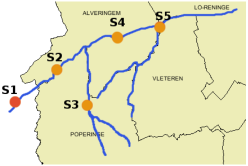
Problem: Step 3, simulate p^* under H_0

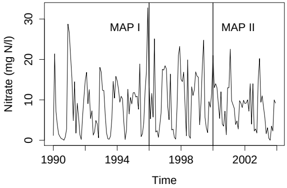
Trend test: Multiplicity correction

Problem: Step 3, simulate p^* under H_0
Solution: simulate from $\hat{F}_0^{f^{(1)}}$.

1. sampling a new set of derivatives $\mathbf{f}^{(1)*}$ under H_0 from $MVN(\mathbf{0}, \hat{\mathbf{\Sigma}}_{f^{(1)}})$
2. calculating the p-values p_k^* that correspond to each of the simulated derivatives $f_k^{(1)*}$, and
3. ranking these p-values according to the *original ranked* p-values $(p_{(1)}, \ldots, p_{(n)})$ to obtain $(p^*_{(1)}, \ldots, p^*_{(n)})$.

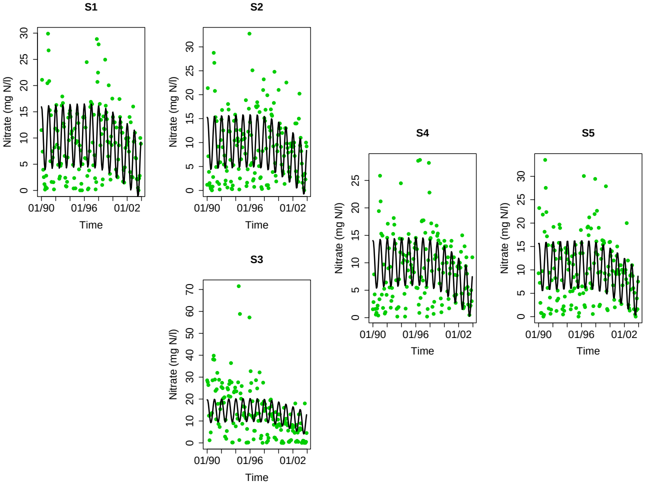
Case Study



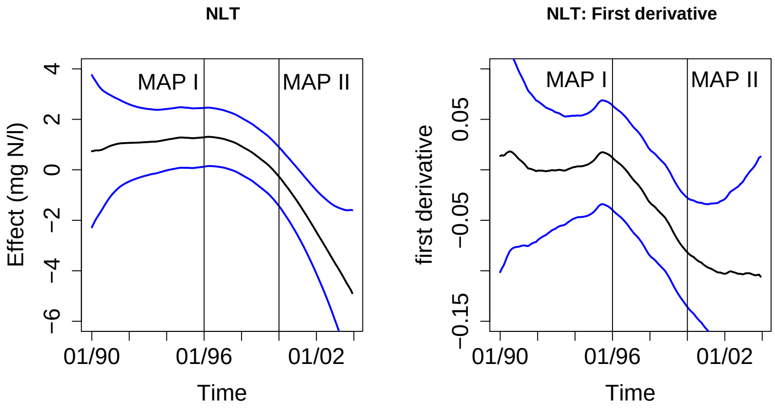


- Introduction of two Manure Decrees (1996 & 2000)
- Has the water quality improved?
- Mean model:
  $\mathrm{E}\{Y\} = \mathbf{X}\beta + \mathbf{f(t)}$
- $\mathbf{f(t)}$ local polynomial regression second order
  $\Rightarrow$ Assess first derivative

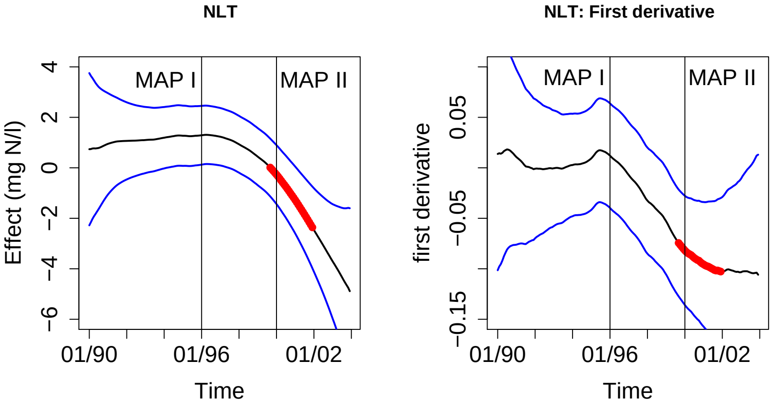
Fit



Trend test: multiplicity correction Westfall & Young

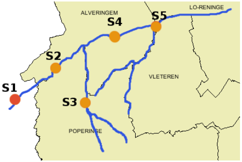


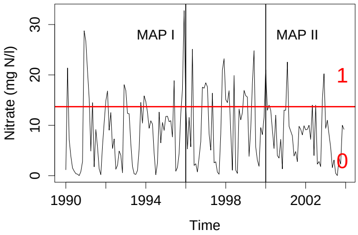
Trend test: multiplicity correction Westfall & Young



Outline

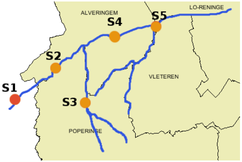
Motivation

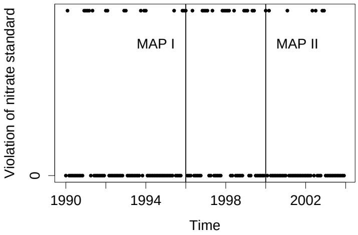




1. River Yzer: Restrict nutrient pollution
2. Previous actions
   1. 1996: MAP I
   2. 2000: MAP II
3. Use of environmental thresholds Nitrate: <13 mg NO_3^--N/l
   1. Above standard: not good (1)
   2. below standard: good (0)

Motivation





1. River Yzer: Restrict nutrient pollution
2. Previous actions
   1. 1996: MAP I
   2. 2000: MAP II
3. Use of environmental thresholds
   Nitrate: <13 mg NO_3^--N/l
   1. Above standard: not good (1)
   2. below standard: good (0)

$\Rightarrow$ Binary response for regulator

Generalised linear mixed models (GLMM)

1. Random component: $y_{it}|\mathbf{x}_{it}, S_{it} \sim \mathrm{B}(\mu^c_{it})$
2. Conditional mean $\mathrm{E}\{y_{it}|\mathbf{x}_{it}, S_{it}\} = \mu^c_{it}$
3. Systematic component: $\nu^c_{it} = \mathbf{x}_{it}\boldsymbol{\beta}^c + S_{it}$
4. Link: $\nu^c_{it} = g(\mu^c_{it})$
5. Spatio-temporal latent process:
   $\mathbf{S}_N = (S_{11}, \ldots, S_{p1}, \ldots, S_{1n}, \ldots, S_{pn})^T$
   $\mathbf{S}_N \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_{S_N})$

$\Rightarrow$ Problem conditional model

Marginal models

1. Marginal mean: $\text{E}\{y_{it}|\mathbf{x}_{it}\} = \mu_{it}^m$
2. Systematic component: $\nu_{it}^m = \mathbf{x}_{it}\boldsymbol{\beta}^m$
3. link: $\nu_{it}^m = g(\mu_{it}^m)$ with $g(.)$ as before

$\Rightarrow$ $\boldsymbol{\beta}^m$ correct marginal interpretation

$\Rightarrow$ GEE is commonly used to fit such marginal models

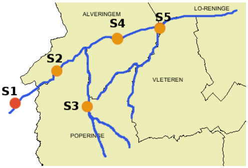
$\Rightarrow$ cannot be applied here: dependence among sampling locations

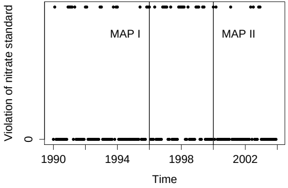
$\Rightarrow$ Solution: obtain marginalised GLMM via integration over latent variable
$\mu_{it}^m = \text{E}\{y_{it}|\mathbf{x}_{it}\} = \text{E}_S(\text{E}\{y_{it}|\mathbf{x}_{it}, S_{it}\}) = \text{E}_S(\mu_{it}^c)$

Marginalised generalised linear mixed models

1. Random component: $y_{it}|\mathbf{x}_{it}, S_{it} \sim \mathrm{B}(\mu_{it}^c)$
2. Marginal mean: $\mathrm{E}\{y_{it}|\mathbf{x}_{it}\} = \mu_{it}^m$
3. Conditional mean: $\mathrm{E}\{y_{it}|\mathbf{x}_{it}, S_{it}\} = \mu_{it}^c$
4. Systematic components: $\nu_{it}^m = \mathbf{x}_{it}\boldsymbol{\beta}^m$
   $\nu_{it}^c = \Delta_{it} + S_{it}$
5. Link: $\nu_{it}^m = g(\mu_{it}^m)$
   $\nu_{it}^c = g(\mu_{it}^c)$
   - $\Rightarrow$ For probit link $g() = \Phi() \Rightarrow \Delta_{it} = \sqrt{1 + S_{it}^2}\mathbf{x_{it}}\boldsymbol{\beta}^m$
   - $\Rightarrow$ Conditional model induces the marginal model of interest
   - $\Rightarrow$ Fit conditional model to obtain marginal model parameters (here in a Bayesian context)
6. Spatio-temporal latent variable: $\mathbf{S}_N \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_{S_N})$

Case Study

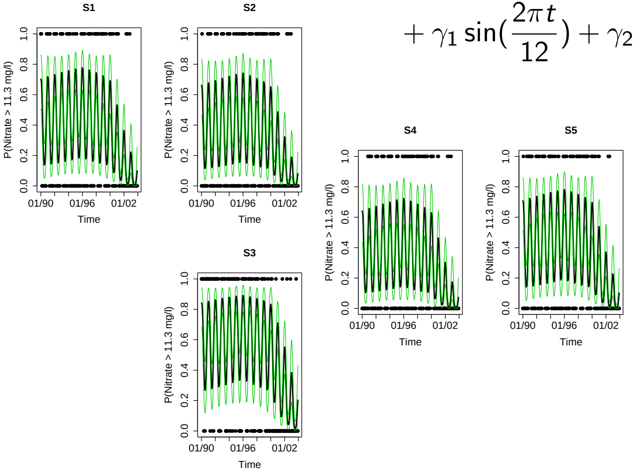




- Nitrate standard: <11.3 mg NO3-N/l
- Binary response for regulator
  1. Above standard: not good (1)
  2. below standard: good (0)
- Trend in the probability to violate the standard?
- Modelling of that probability $E[y_{i,t}|\mathbf{x}_{it}] = \mu^m_{i,t}$

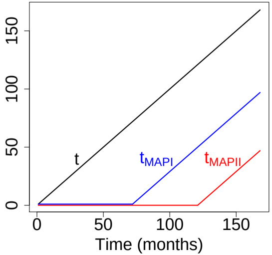
Model for probability of exeedance ($\mu_{i,t}^m$)

$$g(\mu_{i,t}^m) = \alpha_0 + \alpha_i + \beta_0 t + \beta_1 t_{MAPI} + \beta_2 t_{MAPII} + \gamma_1 \sin(\frac{2\pi t}{12}) + \gamma_2 \cos(\frac{2\pi t}{12})$$



Model for probability of exeedance ($\mu^m_{i,t}$)

$$g(\mu^m_{i,t}) = \alpha_0 + \alpha_i + \beta_0 t + \boldsymbol{\beta_1 t_{MAPI}} + \boldsymbol{\beta_2 t_{MAPII}} + \gamma_1 \sin(\frac{2\pi t}{12}) + \gamma_2 \cos(\frac{2\pi t}{12})$$



- Parameter MAP I ($\boldsymbol{\beta_1}$): $[-0.03, 0.01]$
- Parameter MAP II ($\boldsymbol{\beta_2}$): $[-0.06, -0.004]$
- Trend after MAP II ($\boldsymbol{\beta_0} + \boldsymbol{\beta_1} + \boldsymbol{\beta_2}$): $[-0.057, -0.017]$

Outline

- Conclusions
  - Development of spatio-temporal model for river networks
  - Parameter estimation procedure: New algorithm
  - Nonlinear trends: location of trends on a shorter time scale
  - Extension towards marginalised GLMM
  - Case studies: Evidence for beneficial impact of MAPI & MAPII in study region
- Perspectives
  - More complex temporal dependence structures
  - Tidal zones
  - Parameterisation of covariance matrix of observation model
  - Missing data
  - Censored data

1. Clement, L. and O. Thas (2008). Nonparametric trend detection in river monitoring network data: a spatio temporal approach. Environmetrics, DOI: 10.1002/env.929.
2. Clement, L. and O. Thas (2008). Testing for trends in the violation frequency of an environmental threshold in rivers. Environmetrics, DOI: 10.1002/env.911.
3. Clement, L. and Thas, O. (2007). Estimating and modelling spatio-temporal correlation structures for river monitoring networks. Journal of Agricultural, Biological, and Environmental Statistics, 12(2), 161-176.
4. Hastie, T. J. and R. J. Tibshirani (1990). Generalized additive models (First ed.). Monographs on Statistics and Applied Probability. New York: Chapman & Hall.
5. Westfall, P. H. and S. S. Young (1993). Resampling-based multiple testing: Examples and methods for p-value adjustment. New York: John Wiley and Sons.