Nonparametric trend detection in spatio-temporal river monitoring networks

Lieven Clement[1] and Olivier Thas[1]

[1] Ghent University, Department of applied mathematics, biometrics and process control, Coupure links 653, 9000 Ghent, Belgium

Abstract: A spatio-temporal approach is presented for the assessment of trends in river monitoring network (RMN) data. In contrast with most existing methods used for analyzing river data, our model incorporates the spatio-temporal dependence structure explicitly. The temporal dependence structure is assumed to follow an AR(1) process. The spatial dependence structure accounts for the flow direction and is implied by the river topology. The dependence structure is represented by a latent state variable which is further embedded into an observation model to make the correlation structure less rigid. A semi-parametric model is used for the marginal mean, modelling the seasonal trend (ST) and a nonparametric long term trend (NLT). The model is applied to a case study of the river Ijzer (Belgium), where the NLT is shown to be decreasing over the last couple of years.

Keywords: Spatio temporal model; water quality, GAM.

1 Introduction

Data of RMN are often used to asses the evolution of the water quality over time. These networks typically generate data with a strong spatial and temporal dependence structure. Since the water flows only in one direction, an interpretation can be given to the spatial correlations. Many researchers have avoided the estimation of the spatio-temporal dependence in RMN data by using ad hoc methods or, even worse, by simply ignoring it. However, to control the type I error of statistical tests, the spatio-temporal correlation structure has to be modelled. Here we present a spatio-temporal model for trend detection (TD) of RMNs based on local linear regression smoothers. First the spatio-temporal model is presented in Section 2. The parameter estimation is briefly introduced in Section 3, and, Section 4 deals with the TD procedure and concludes with a small case study.

2 The Spatio-temporal model

At each time $t = 1 \dots N$, let $\mathbf{S}_t = (S_{t1} \dots S_{tp})^T$ represent the response variable at time t and sampling locations $j = 1 \dots p$. The correlation struc-

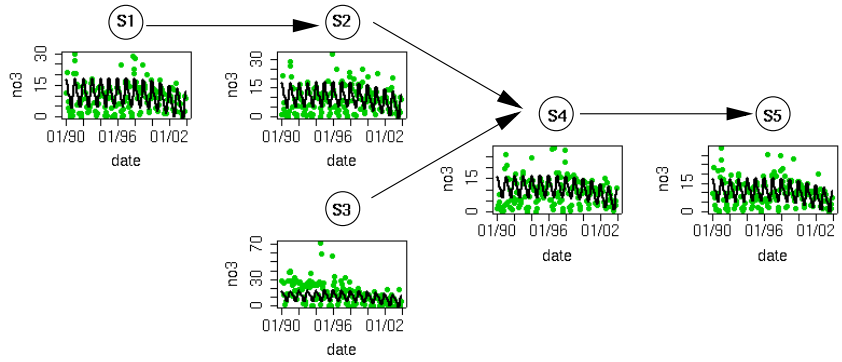

FIGURE 1. Evolution of the water quality at five sampling locations of the river Ijzer. Sampling locations S1, S2, S4, S5 are located on the main river, sampling location S3 is located on a tributary which drains into the Ijzer between S2 and S4.

ture of $\mathbf{S}_t$ is completely defined by the river topology. This is illustrated in Figure 1, which shows the data and the river topology of 5 sampling locations. The same figure can also be interpreted as a Directed Acyclic Graph (DAG) (see e.g. Whittaker, 1990) in which the circles represent S_{tj}'s and arrows immediately determine the conditional independence structure. The DAG can be represented by,

$$\mathbf{S}_t = \mathbf{A}\mathbf{S}_t + \boldsymbol{\gamma}_t, \tag{1}$$

where $\mathbf{A}$ can be written as a lower triangular square matrix with zeroes at the diagonal, and $\boldsymbol{\gamma}_t$ is multivariate normally distributed (MVN): $\boldsymbol{\gamma}_t \sim MVN(0, \boldsymbol{\Sigma}_\gamma)$ with a diagonal variance-covariance matrix $\boldsymbol{\Sigma}_\gamma$.

For the temporal dependence structure we assume an AR(1) process. Extending and rearranging Equation (1) gives,

$$\mathbf{S}_t = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{S}_{t-1} + (\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{\eta}_t. \tag{2}$$

where $\mathbf{B}$ is a diagonal matrix containing the autoregression coefficients, $\mathbf{I}$ is the identity matrix, and $\boldsymbol{\eta}_t \sim MVN(0, \boldsymbol{\Sigma}_\eta)$ with a diagonal variance covariance matrix $\boldsymbol{\Sigma}_\eta$.

In reality, however, the dependence structure might be obscured by common environmental confounders. Therefore, the model is embedded into an observation model

$$\mathbf{Y}_t = \mathbf{S}_t + \boldsymbol{\epsilon}_t, \tag{3}$$

where $\mathbf{Y}_t$ is the observation vector corresponding to $\mathbf{S}_t$, and $\boldsymbol{\epsilon}_t \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$. Equation (3) is extended with a semiparametric model for the mean: $\mathrm{E}[Y_{tj}] =$

$\mathbf{X}_{tj}\boldsymbol{\beta}_j + f_j(t)$, where $\boldsymbol{\beta}_j$ is the parameter vector of the ST and $\mathbf{X}_{tj}$ is the $1 \times q$ design vector with the proper Fourier basis functions at S_{tj} and where $f_j(t)$ is a local linear regression smoother for the estimation of NLT at this location. After embedding the mean model into model (3), we obtain,

$$\mathbf{Y}_t = \mathbf{X}_t\boldsymbol{\beta} + \mathbf{f}(t) + \mathbf{S}_t + \boldsymbol{\epsilon}_t, \tag{4}$$

which specifies together with model (2) the complete stationary spatio-temporal state space model.

3 Parameter estimation

Since the model can be written as a state space model, an EM-algorithm can be used for parameter estimation (Harvey, 1989). A detailed derivation for RMN can be found in Clement and Thas (2006a). The mean model can be estimated using least squares or generalized least squares. For notational simplicity, only the former is considered here. The estimating equations of semi-parametric models with one linear smoother have an analytical solution (Hasti and Tibshirani, 1990);

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T(\mathbf{I} - \mathbf{H}_f)\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{H}_f)\mathbf{Y} \tag{5}$$

$$\hat{\mathbf{f}} = \mathbf{H}_f(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{6}$$

where $\mathbf{H}_f$ is the smoother matrix and $\mathbf{I}$ is the identity matrix. In the case study we used the Epanechnikov Kernel for the local linear regression. The bandwidth was selected by a grid search using generalized cross validation. The results of the model fit are given in Figure 1, the ST is depending on the sampling location and the NLT is common over all sampling locations. Hence, the NLT is estimated over a more regional scale.

4 Assessing the nonparametric long term trend

To know if the NLT is locally varying, an analysis of its derivative, $\mathbf{f}^{(1)}(t)$ is proposed. From local linear regression smoothers, the derivative can be easily calculated and is linear in the response. From the state space model the complete variance-covariance matrix $\boldsymbol{\Sigma}_Y$ of $\mathbf{Y}$ can be directly calculated (Clement *et al.*, 2006b). The variance-covariance matrix of the NLT ($\boldsymbol{\Sigma}_f$) and of the derivative ($\boldsymbol{\Sigma}_{\mathbf{f}}^{(\mathbf{1})}$) are linear combinations of $\boldsymbol{\Sigma}_Y$. Hence, simple test statistics can be used for asymptotic pointwise inference on the derivative, e.g. $\frac{f^{(1)}(t)}{s_{f^{(1)}(t)}}$ has asymptotic standard normal null distribution. A correction for multiplicity is applied as discussed by Langsrud (2005). The trend and its derivative are represented in Figure 2. In this graph a significant decrease ($\alpha = 5\%$) is indicated with red dots. Figure 2 clearly shows that the NLT is decreasing during the last couple of years.

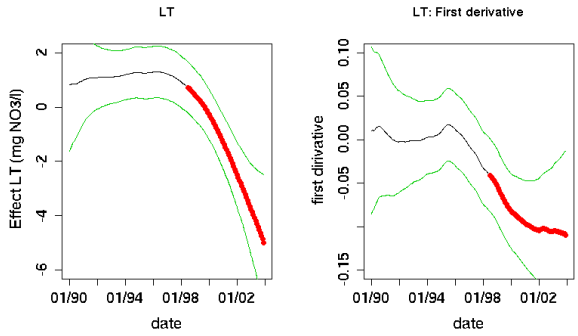

FIGURE 2. Evaluation of the nonlinear trend at five sampling locations of the river Ijzer. The trend is given in the left panel, the first derivative of the trend is given in the right panel. In both graphs, 95% pointwise confidence bands are given. A global significant decrease is indicated with a red dot, no significantly changes are indicated with a thin black line.

When linear smoothers are used, the approach can easily be extended to more general additive models, as they still enable straightforward calculation of the variance-covariance matrices of the additive member functions.

References

Clement, L. and Thas, O. (2006a). Estimating and modelling spatio-temporal correlation structures for river monitoring networks. *Journal of Agricultural, Biological, and Environmental Statistics*, submitted.

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

Hasti, T. J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge: Cambridge University Press.

Langsrud, O. 2005. Rotation tests. *Statics and Computing*, **15**, 53-60.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Chichester: John Wiley & Sons.