Spatio-temporal statistical models for river monitoring networks

L. Clement, O. Thas, P.A. Vanrolleghem and J.P. Ottoy

Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure Links 653, B-9000 Gent, Belgium (E-mail: *Lieven.Clement@ugent.be*)

Abstract When introducing new wastewater treatment plants (WWTP), investors and policy makers often want to know if there indeed is a beneficial effect of the installation of a WWTP on the river water quality. Such an effect can be established in time as well as in space. Since both temporal and spatial components affect the output of a monitoring network, their dependence structure has to be modelled. River water quality data typically come from a river monitoring network for which the spatial dependence structure is unidirectional. Thus the traditional spatio-temporal models are not appropriate, as they cannot take advantage of this directional information. In this paper, a state-space model is presented in which the spatial dependence of the state variable is represented by a directed acyclic graph, and the temporal dependence by a first-order autoregressive process. The state-space model is extended with a linear model for the mean to estimate the effect of the activation of a WWTP on the dissolved oxygen concentration downstream.

Keywords Theory intervention analysis; parameter estimation; water quality; space-time model

Water Science & Technology **Vol 53 No 1 pp 9–15** © IWA Publishing 2006

Introduction

The increasing interest in environmental issues has currently been translated into legislation. The European Water Framework Directive is an example of these efforts. One of the major goals of this directive is to maintain and improve the aquatic environment. In order to reach this goal one of the possible actions is to built wastewater treatment plants (WWTPs). Obviously, investors and policy makers want to know if there indeed is a beneficial effect of the installation of a WWTP on the river water quality. Such an effect can be established in time, after as compared to before the installation, and in space, downstream as compared to upstream of the WWTP.

In the time-series literature, this question is referred to as intervention analysis. Given a known intervention, the analysis assesses the evidence that an expected change in the time-series actually occurred and if so, the nature and magnitude of the change is also investigated (Box and Tiao, 1975). To investigate the effect of the WWTP, the mean difference of the water quality before and after the event has to be estimated at sampling locations down- and upstream from the WWTP. To be significant, the estimated change of the constituent concentration at the sampling location downstream has to be significantly different from zero. This shift in the mean concentration can only be attributed to the WWTP if no similar shift occurs at the sampling locations upstream from the WWTP. Since a river monitoring network is used to collect the data, measurements show spatial and temporal correlation. Hence the statistical model must incorporate this dependence structure. When the spatio-temporal correlation structure is ignored or modelled incorrectly, all statistical inference on the effect of the WWTP is not guaranteed to be valid. Given the spatio-temporal dependence and the fact that the effect of the WWTP can be established both in time and space, it is clear that spatio-temporal models are needed for intervention analysis (Thas and Ottoy, 1999).

doi: 10.2166/wst.2006.002

Traditional approaches to this problem have focused on the geostatistical paradigm (Wikle and Cressie, 1999; Bilonick, 1983) and on multivariate time-series methods, which specify dynamic models that are linked spatially (Rouhani and Wackernagel, 1990). Huang and Cressie (1996) classify time-series as dynamic since the temporal dependence arises from a unidirectional correlation. This unidirectional structure often is utilised in time-series techniques. A first-order autoregressive model is a clear example. Geostatistical methods, on the other hand, are classified as descriptive because of the non-directional correlation. There is no causative interpretation associated with the observed spatial correlation. Based on these considerations Huang and Cressie (1996) derived a temporal dynamic and spatial descriptive model to predict the snow water equivalent during the snow season.

In this paper a spatio-temporal model is presented for the intervention analysis of data obtained from a river monitoring network. With respect to the spatial dependence structure an important distinction has to be made with the classical spatial structures. Since the water flows only in one direction within the river reaches, a causal interpretation can be given to the correlations. However, as opposed to time, rivers can join or split. This implies a more general branched unidirectional structure. Therefore, according to Cressie's terminology, a spatio-temporal model is required that is dynamic with respect to both the spatial and the temporal dependence structure.

For the case study presented in this paper, only a small part of the river monitoring network of the Region of Flanders in Belgium is considered. The network consists of three sampling locations upstream of a WWTP and one sampling location downstream. The dissolved oxygen concentration (DO) is measured monthly. To derive a valid statistical procedure for the evaluation of the effect of the WWTP, a linear model for the mean DO is embedded into a spatio-temporal model.

First the statistical model is formulated. The model consists of two parts: a model for the covariance structure, which is uniquely determined by the spatio-temporal dependence structure, and a model for the mean, which is needed to answer the substantive research question raised in the intervention analysis. Then the case study is presented and is followed by the results and discussion and a conclusion.

The spatio-temporal model

Dependency structure

At each time $t = 1, K, N$, let $S_t = (S_t^1 \ldots S_t^p)^T$ denote the stationary spatial process, where S_t^j represents the DO concentration at time t and sampling location j. The correlation structure of S_t is defined by the conditional independence structure, which is easily derived for a river monitoring network of branched unidirectional river reaches. This is illustrated in Figure 1. In this figure five sampling locations along two joining river reaches are schematically represented; the direction of the water flow is also indicated. The same figure can also be interpreted as a directed acyclic graph (DAG) (Whittaker, 1990) in which the circles represent the vertices associated with the corresponding S_t^j. Missing edges or arrows

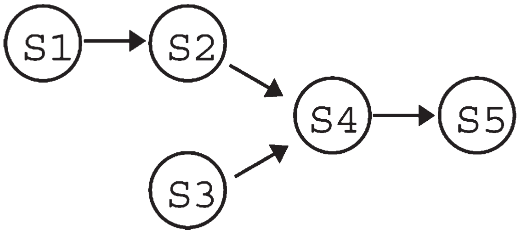

Figure 1 DAG of the five sampling locations along two joining river reaches

immediately determine the conditional independencies: S_t^1 is independent of S_t^3; S_t^1 is independent of S_t^4 given S_t^2; S_t^1 is independent of S_t^5 given S_t^2; S_t^1 is independent of S_t^5 given S_t^4; S_t^2 is independent of S_t^3, S_t^2 is independent of S_t^5 given S_t^4; S_t^3 is independent of S_t^5 given S_t^4. This conditional independence structure is assumed to hold at each time *t*.

Since the covariance structure can be formulated as a DAG, it can also be equivalently represented by a recursive system of equations (Wermuth, 1980).

$$S_t = AS_t + \gamma_t, \tag{1}$$

where the order of the elements of S_t can be arranged such that A is a lower triangular square matrix with zeroes on the diagonal, and γ_t is a multivariate zero-mean random vector with a diagonal variance-covariance matrix $\Sigma_{\gamma.}$ It is further assumed that γ is multivariate normally distributed (*MVN*), denoted as $\tilde{\gamma}MVN(0, \Sigma\gamma)$.

In reality, however, the dependence structure may possibly be obscured by common environmental influences such as rainfall or climatological conditions in general. The rather strict structure implied by the model in Equation (1) is assumed to hold only for an isolated river system. Therefore this model is seen as the model for the unobservable state variable S_t, and the model is embedded into an observation model

$$Y_t = S_t + \eta_t, \tag{2}$$

where Y_t is the observation vector corresponding to S_t, and η_t is the zero-mean error term. Here, $\tilde{\eta}MVN(0, \Sigma\eta)$. Equations (1) and (2) define the spatial model, which is a state-space model.

Temporal dependence structure

For the temporal dependence structure a first-order autoregressive model for the state variable is assumed,

$$S_t = BS_{t-1} + \delta_t, \tag{3}$$

where B is a diagonal matrix containing the autoregressive parameters and $\tilde{\delta}MVN(0, \Sigma_\delta)$ with a diagonal variance-covariance matrix Σ_δ.

Model for the mean

Up to here it has been assumed that the mean of Y_t is zero, i.e. $\mathrm{E}[Y_t] = 0$ for all t. Only the covariance structure of the stationary process Y_t has been modelled. Throughout this paper a linear model for the mean is used,

$$\mathrm{E}[Y_t] = X_t\beta, \tag{4}$$

where $\beta = (\beta_1, \ldots, \beta_q)^T$ is the parameter vector and X_t is the $p \times q$ design matrix which may contain time-dependent covariates. After embedding the mean model in Equation (2), the following equation is obtained

$$Y_t = X_t\beta + S_t + \eta_t. \tag{5}$$

Spatio-temporal model formulated as a structural equation model

Another equivalent formulation of the spatio-temporal model is accomplished by recognising that the model (Equations (1), (3) and (5)) can be written as a structural equation model (SEM) (Maruyama, 1997)

$$\begin{cases} \quad CS = \zeta \\ Y = X\beta + S + \varepsilon, \end{cases} \tag{6}$$

where $S = (S_1^T \dots S_N^T)^T$, $Y = (Y_1^T \dots Y_N^T)^T$, $X = (X_1^T \dots X_N^T)^T$, C is a $pN \times pN$ square matrix constructed from the elements of the matrices A and B, $\tilde{\zeta}MVN(0, \Sigma_\zeta)$, where Σ_ζ is diagonal, and $\tilde{\varepsilon}MVN(0, \Sigma_\varepsilon)$ where Σ_ε is block-diagonal with blocks Σ_η.

From this SEM formulation it is obvious that a zero-mean spatio-temporal model ($\beta = 0$) only specifies the covariance structure of the observation vector Y

$$\Sigma_y = \operatorname{var}(Y) = C^{-1}\Sigma_\zeta C^{-T} + \Sigma\varepsilon. \tag{7}$$

Parameter estimation

The spatio-temporal model determines the covariance structure of the observation vector Y. The parameter vector β can be estimated by weighted least squares (WLS)

$$\hat{\beta} = (X^T \hat{\Sigma}_Y^{-1} X)^{-1} X^T \hat{\Sigma}_Y^{-1} Y \tag{8}$$

for which a consistent estimator of var(Y) is needed (Equation (7)),

$$\hat{\Sigma}_Y = \hat{C}^{-1}\hat{\Sigma}_\zeta \hat{C}^{-T} + \hat{\Sigma}_\varepsilon. \tag{9}$$

Inference on β is then based on the estimated covariance matrix $\hat{\Sigma}_\beta$

$$\hat{\Sigma}_\beta = (X^T \hat{\Sigma}_Y^{-1} X)^{-1}, \tag{10}$$

where $\hat{C}$, $\hat{\Sigma}_\zeta$ and $\hat{\Sigma}_\varepsilon$ are all consistent estimators based on the model represented by Equation (6). The estimation of the parameters is based on the factorisation of the likelihood according to the recursive nature of the DAG. Since the DAG is only applicable to S, the likelihood of Y cannot be factorised accordingly. An efficient algorithm has been developed to overcome this computational problem. Details are given elsewhere (Clement *et al.*, submitted).

Case study

The data used in this case study are part of a public database of the Flemish Environmental Agency (http://www.vmm.be); more details can be obtained from the first author.

Figure 2 shows schematically the location of four sampling locations along three river reaches in the neighbourhood of the city of Ertvelde in Belgium. Actually the three river reaches join just before sampling location S1, but this is not indicated to maintain the DAG-interpretation of Figure 2.

Monthly observations were available at each sampling location between January 1990 and November 2002. In August 1997 a WWTP, located just downstream from the junction of the river reaches coming from locations S2, S3 and S4 and just upstream from location S1, was activated. The question addressed in this paper concerns the possible effect of the WWTP on the DO concentration. The DO concentration shows a highly seasonal pattern, which is modelled by a factor with 12 levels (one for each month). The model of the mean DO at time t and location j is given by

$$\mathrm{E}\left[DO_t^j\right] = \beta_{1j} + \beta_{2j}t + \beta_{3j}I(t \geq t_c) + X_m\alpha, \tag{11}$$

where I (.) denotes the indicator variable, t_c is the date of the activation of the WWTP (August 1997); X_m is a matrix with dummy variables used to model the seasonal effect (it is equal to 1 when the data belongs to the current month and 0 elsewhere), and α is the vector with the effect parameters for the 12 months, with the restriction $\Sigma_{k=1}^{12} \alpha_k = 0$. Note that the model in Equation (11) is a linear model which can be written in matrix notation as in Equation (4).

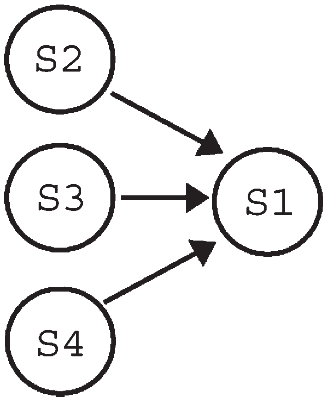

Figure 2 DAG of the four sampling locations along three river reaches. The WWTP is located between locations S2, S3, S4 and S1

The effect of the activation of the WWTP at location S1, quantified by β_{31}, has to be assessed by means of statistical tests. If a beneficial effect exists due to the wastewater treatment plant, the effect should be detected over time as well as over space. The former is equivalent to $\beta_{31} \neq 0$ and the latter effect can be tested by comparing β_{31} to the effects β_{32}, β_{33}, β_{34} in the sampling locations upstream. The hypothesis of interest for the first test is formulated as

$$H_0 : \beta_{31} = 0, \tag{12}$$

which expresses that there is no change in the mean DO at the downstream location (location S1).

The second hypothesis of interest is formulated as

$$H_0 : L = \begin{bmatrix} L_1 & L_2 & L_3 \end{bmatrix}^T = H\beta = \begin{bmatrix} \beta_{31} - \beta_{32} & \beta_{31} - \beta_{33} & \beta_{31} - \beta_{34} \end{bmatrix}^T = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T, \tag{13}$$

where H is the appropriate contrast matrix. Equation (13) expresses that, if there is a change in mean at the downstream location (location S1), this shift in mean is the same at the upstream locations, and hence it cannot be attributed to the installation of the WWTP. The null hypothesis has to be tested against the alternative hypothesis that the effect at the downstream location is different from all three effects at the upstream locations. (As soon as one of the upstream effects is similar to the downstream effect, the effect at the downstream location cannot be attributed to the WWTP). The vector L will be estimated by

$$\hat{L} = \begin{bmatrix} \hat{L}_1 & \hat{L}_2 & \hat{L}_3 \end{bmatrix}^T = \begin{bmatrix} \hat{\beta}_{31} - \hat{\beta}_{32} & \hat{\beta}_{31} - \hat{\beta}_{33} & \hat{\beta}_{31} - \hat{\beta}_{34} \end{bmatrix}^T, \tag{14}$$

which is $MVN(0, H\Sigma_\beta H^T)$. This multivariate normal distribution will be used to calculate the p-value.

All the statistical tests will be performed on a 5% significance level.

Results and discussion

Figure 3 shows monthly observations between January 1990 and November 2002, and the resulting model fit for DO at each location. This figure clearly indicates a sudden increase in DO at location S1 at the time the WWTP was activated (August 1997). The results of the parameter estimation are presented in Table 1.

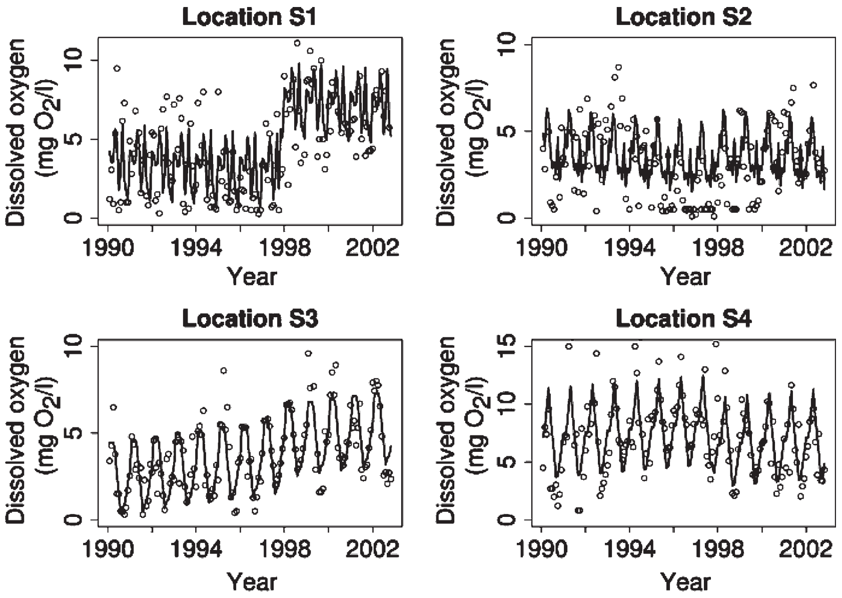

Figure 3 Time-series (o) and model fit (–) of the DO (mg/l) at the different sampling locations (location 1 downstream from the WWTP, locations 2–4 upstream from the WWTP)

The parameter estimates and the *p*-values for the intercept, the long-term trend and the intervention effect are given in Table 1 (estimates for the seasonal trend are not shown). They indicate a very significant increase of the mean DO ($p < 0.001$) after the start-up of the WWTP at the downstream location (S1), whereas at the upstream locations S2, S3 and S4 no significant changes are seen ($p = 0.63$, $p = 0.08$ and $p = 0.06$). Since a significant positive shift in mean DO is only established in the location S1, there seems to be a positive effect of the WWTP over time.

The presence of a spatial effect of the WWTP is tested using Equations (13) and (14). The *p*-value is $p < 0.001$, and, therefore, the null hypothesis is very strongly rejected. Thus, it may be concluded that the shift in mean at location S1 is different form the upstream locations, indicating that there is indeed a spatial effect of the WWTP. Although this data analysis methodology has no causal interpretation, the results give a strong indication that the WWTP has a positive effect on the downstream DO concentration.

The current spatial and temporal dependence structure implies some strong assumptions on the data, such as multivariate normally distributed residuals and a second-order stationarity in the temporal and spatial covariance structure. For many space-time processes there is little reason to expect spatial (and sometimes temporal) stationarity of the covariance structure (Chunsheng, 2003; Wikle and Cressie, 1999). It is a challenge to develop new, less-restrictive spatio-temporal covariance structures for river network modelling.

Table 1 Estimates of βij, their standard deviations (sd) and the *p*-values corresponding to H_o:$\beta_{ij} = 0$, *p*-values are not given for the intercepts

	Intercept ($i = 1$)		Slope ($i = 2$)			Intervention effect ($i = 3$)		
J	**Estimate**	**sd**	**Estimate**	**sd**	***p***	**Estimate**	**sd**	***p***
Location S1	3.27	0.55	−0.007	0.01	0.5	4.53	0.92	<0.001
Location S2	3.65	1.04	−0.009	0.018	0.61	0.79	1.64	0.63
Location S3	2.20	0.29	0.014	0.005	0.01	0.87	0.5	0.08
Location S4	6.41	0.62	0.013	0.012	0.26	−2.00	1.07	0.06

Conclusions

The proposed spatio-temporal model is able to quantify the spatial and temporal dependence in river water quality networks. This dependence structure is necessary to perform correct statistical inference on the model parameters. In the case study the model was used to perform an intervention analysis on the activation of a WWTP. The results gave a strong indication of a positive effect of the WWTP on the downstream DO concentration.

Acknowledgements

The results presented in this publication, have been elaborated within the framework of a research project of the Flemish Environmental Agency (VMM) and the EU project Harmoni-CA, contract no EVK1-CT-2002-20003. The programme is organised within the Energy, Environment and Sustainable Development Programme in the 5th Framework Programme for Science Research and Technological Development of the European Commission. The VMM provided the data for the case study.

References

Bilonick, R.A. (1983). Risk qualified maps of hydrogen ion concentration for the New York state area 1966–1978. *Atmospheric Environment*, **17**, 2513–2524.

Box, G.E.P. and Tiao, G.C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, **70**(349), 70–79.

Chunsheng, M. (2003). Nonstationary covariance functions that model space–time interactions. *Statistics and Probability Letters*, **61**(4), 411–419.

Clement, L., Thas, O., Vanrolleghem, P.A. and Ottoy, J.P. (Submitted). Estimating and modelling a spatio-temporal correlation structure for river monitoring networks. *Journal of Agriculture, Biological and Environmental Statistics*.

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

Maruyama, G.M. (1997). *Basics of structural equation modeling*, Sage Publications, Thousand Oaks, CA, USA.

Rouhani, S. and Wackernagel, H. (1990). Multivariate geostatistical approach to space-time data analysis. *Water Resources Research*, **26**, 585–591.

Thas, O. and Ottoy, J.P. (1999). An approach to intervention analysis in spatio-temporal modelling of river quality. In *Statistics for the Environment 4: Statistical Aspects of Health and the Environment*, Barnett, V. (ed.), Wiley, Chichester, UK, pp. 221–234.

Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis. *Journal of the American Statistical Association*, **75**(372), 963–972.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley, Chichester, UK.

Wikle, C. and Cressie, N (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, **86**(4), 815–829.