ir. Lieven Clement

Statistical validation and spatio-temporal modelling of river monitoring networks

Thesis submitted in fulfilment of the requirements for the degree of Doctor (Ph.D.) in Applied Biological Sciences

Summary

The European Water Framework Directive (WFD)(EC, 2000) is one of the driving forces in environmental policy in the European Union. The WFD's overall environmental objective is the achievement of 'good status' for all of Europe's surface- and ground waters within a 15-year period. Its implementation is a big challenge for the European environmental managers. One of the key actions of the WFD is the design of operational monitoring programmes. Thus, large amounts of water quality data are being collected, processed and stored throughout Europe. Due to the large amount of the data and their complex nature, statistical modelling has become an essential tool to extract reliable information from these observations. Because high quality data is essential for an adequate management of the water resources, data validation procedures are required to build consistent databases. Once the environmental agencies have a consistent database at their disposal, the data should be used to assess the evolution of the water status and to evaluate the impact of their management strategies. Such an assessment should be possible at the level of individual sampling locations as well as on a more regional scale. Due to the spatio-temporal dependence structure of monitoring network data, spatio-temporal models are needed for a correct statistical assessment. For river monitoring networks the development of spatio-temporal models has just begun. The data validation problem and the development of spatio-temporal models for river network data are the two major themes of this dissertation.

In the first part the data validation problem is addressed. Like other environmental data, water quality data have a complex nature. They contain a considerable amount of noise due to their natural variability and the measurement error. They may contain missing values, are often non-normally distributed and are commonly

gathered at irregular time instants. Moreover, they possess cyclic variations and contain nonlinear trends. With this respect, additive models are explored to model water quality data. These models are then used to design an automatic validation procedure for new observations that are acquired with a river monitoring network. Based on historical data, additive models are fitted to predict new observations and to construct prediction intervals (PI's). A new observation is declared valid if it is located within the interval. Several methods were developed to derive such PI's and the PI that was based on bootstrapping studentised prediction errors was shown to be most accurate and most robust to deviations from normality. The coverage of these prediction intervals and their power to detect anomalous data are successfully established in a simulation study. The method is illustrated on two case studies in which the method detected abnormal nitrate concentrations in the water body provoked by a dry summer which was followed by an extremely wet winter period. Currently, the Flemish environmental agency is also using our method for the validation of new observations from their physico-chemical monitoring network.

In the second part a spatio-temporal model is developed for river monitoring network data. The aim was to enable valid statistical inference based on the data that is observed at the sampling locations. Therefore the observations of the monitoring network at a certain time instant can be considered as the realisation of a finite-dimensional multivariate random variable with each dimension corresponding to each of the p sampling locations. This enables us to write the model as a p-dimensional state-space model. The state variable is defined by a Directed Acyclic Graph (DAG) that is derived from the river network topology. In reality the dependence structure based on the DAG may be obscured by environmental factors such as rainfall and climatological conditions in general. This is taken into account by embedding the state variable into an observation model. Initially, the observation model was extended with a linear model for the mean. The specification of the mean model allows the assessment of different research questions. An efficient expectation-conditional-maximisation (ECM) algorithm is proposed for parameter estimation, using the Kalman filter and smoother in both E- and CM-steps. However, many environmental processes are characterised by a nonlinear trend. To allow the estimation of such a nonlinear trend, we replaced the parametric mean model by a semiparametric model that used a smoother for the estimation of the trend component. This procedure also allows to test for trends on a smaller time scale. To detect if the local trend is significant, tests on the first derivative of the nonlinear trend are performed at each time step. This results, however, in a large number of simultaneous tests. Multiplicity is thus another problem which had to be addressed. Many environmental processes are also non-Gaussian. To

handle such data, a generalisation of our spatio-temporal model is needed. A first attempt is presented that can handle binary data. Environmental compliance is often based on threshold levels, providing a binary response to the decision maker. We made use of generalised linear mixed models (GLMM) to model such a binary response. Again, the spatio-temporal dependence structure is introduced by using a latent state variable. In a GLMM the parameters of the mean model have a conditional interpretation. In an environmental context, however, we want to infer on the marginal mean. Therefore the marginalised version of the GLMM of Heagerty and Zeger (2000) is used. They introduced a mapping function between the conditional and marginal model components to identify a conditional model structure that allows immediate estimation of the marginal mean parameters.

The spatio-temporal models are applied on a case study of five sampling locations of the river Yzer. In the sampling period, a first manure action plan (MAP) was introduced in 1996 (Vlaams Parlement, 1995) and a second and more restrictive MAP was established in 2000 (Vlaams Parlement, 1999). Both MAP's aim to reduce the nutrient pollution originating form agricultural activities. Our modelling procedure was shown to be very flexible. Depending on the formulation of the mean model, inference is possible on a regional scale, on the level of a river reach or on the level of individual sampling locations. In a first case study the annual average of the nitrate concentration in 2003 is shown to be very significantly lower than the general mean ($p < 0.01$). Moreover, in the main river, the mean nitrate concentration of 2003 was also significantly lower than the mean of the two most recent years ($p = 0.03$). In the second case study, the spatio-temporal model was used to estimate a nonlinear trend. A significant decrease in the nitrate concentration is established in the study region between the introduction of the first MAP and the second MAP ($\alpha = 0.05$). The trend remains significant until January 2002. Both case studies indicated a strong seasonal variation with lower nitrate values in summer and higher contributions in winter. Finally an assessment was done of the violation of the nitrate standard of 11.3 mg-N/l. In the study region a strong seasonal pattern was present in the violation probability. The probability to violate the standard was larger during the wet winter period than in the dry summer period. There was a strong evidence in favour of the presence of a trend change after the introduction of the second manure action plan. The trend change was large enough to establish a decreasing trend in the violation probability of the standard in the study region. Although the data analysis has no causal interpretation, the results of the case studies give a strong indication that the introduction of the manure action plans had a beneficial effect on the nitrate status in the study region.