Ann. Limnol. - Int. J. Lim., 40 (3), 181-191

Application of Bayesian Belief Networks for the prediction of macroinvertebrate taxa in rivers

V. Adriaenssens*, P.L.M. Goethals, J. Charles, N. De Pauw

Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University , Jozef Plateaustraat 22, B-9000 Gent, Belgium.

Integrated ecological models are of great potential as predictive tools in decision support of river management. Such models need to be transparent and consistent with the existing expert knowledge, and give the river manager adequate information regarding their inherent uncertainty. One way to fulfil these needs is through the use of Bayesian Belief Networks (BBNs). Such networks represent cause-and-effect assumptions between system variables in a graphical structure. To establish the potential of Bayesian Belief Networks in river management, a small-scale study was performed with the aim of assessing the success of prediction of macroinvertebrate taxa in rivers by means of a selected number of environmental variables. Gammaridae and Asellidae were chosen because of their high relative abundances in small and large brooks in contrast to other macroinvertebrate taxa. Based on one-layered BBN networks, the predictive capacity of the models was assessed by means of the number of Correctly Classified Instances (CCI) and Cohen's Kappa (*K*). The performance of these models was moderate to good for presence/absence classifications but showed a low to moderate performance when predicting abundance classes. When extending the former BBN network to a two-layered one, enhancing the number of links and variables, no obvious improvement in model performance was detected. The results indicate that thoughtful input variable selection as well as sensitivity analysis will improve the models for practical use in river restoration management.

Keywords : decision support, river management, uncertainty, ecological modelling.

Introduction

Biological monitoring as a tool for river quality assessment has a long history in Europe and beyond. The implementation of the EU Water Framework Directive (WFD) (EU 2000), the aim of which is to reach a good ecological status in all water bodies by 2015, will require large investments of the members states in river management. Particularly in Flanders, Belgium, the river quality is still far from satisfactory and a lot of investment programmes still need to be carried out in the coming years. In 2002, 71% of the measuring points in Flanders did not reach the basic biological quality standard according to the Belgian Biotic Index (BBI) with a minimum of 7 on a scale of 10 (De Pauw & Vanhooren 1983). Assuming that this basic water quality standard corresponds with a «good» ecological status as defined by the WFD (EU 2000), obviously a great effort will be needed to reach this standard in all rivers in Flanders by 2015. Likely, such efforts should not be restricted to further emission reductions of point sources, but should also address the importance of diffuse pollution originating mainly from agricultural activities. Moreover, attention will have to be paid to physical restoration actions to create favourable biotopes for the biological communities (Maeckelberghe 2003).

Because of the large-scale investments required, governments need reliable scientific decision support to set priorities for their environmental investments in order to reach a cost-effective restoration of the rivers and a drastic improvement of the river water quality. Nowadays, this decision support in river ecology is mainly based on biological assessment methods such as the Belgian Biotic Index, which is based on ma-

* Corresponding author :
E-mail : Veronique.Adriaenssens@UGent.be

Article available at http://www.limnology-journal.org or http://dx.doi.org/10.1051/limn/2004016

croinvertebrates (De Pauw & Vanhooren 1983). Macroinvertebrates have been monitored for more than a decade by the Flemish Environment Agency (VMM) and used as a bio-indicator system for the assessment of running waters in Flanders. Generally, little use is made of the information hidden in the relative abundance levels of the macroinvertebrate taxa, and no use is made explicitly of the information provided by the absence of commonly occurring macroinvertebrate taxa itself. Advanced techniques of analysis can be used to maximise the information value gained from biological monitoring data by taking into account the difference in relative abundance levels between specific sites (O'Connor & Walley 2002). The derived information can be either used for diagnostic or predictive purposes in water management. In particular, ecological models are of great interest as predictive tools in decision support and can be used for setting up monitoring networks, defining and interpreting river quality data and selecting appropriate restoration management actions (Goethals & De Pauw 2001).

Predictive ecological models that are capable of establishing a scientifically sound link between the abiotic and biotic river components and meet the objectives of managers do however require some specific features (Reichert & Omlin 1999, Reckhow 2002, Borsuk et al. 2004). First, these models should be meaningful to the broad range of persons involved in the decision making process and therefore a clear presentation of the model structure and the inference process is required. Second, the existing knowledge and data should be easy to integrate. Third, the models should explicitly incorporate uncertainties in their structure and in their predicted outcomes. Finally, such models should be able to reflect evolving scientific knowledge and policy needs.

A method that comes close at integrating all the above features is the application of Bayesian Belief Networks (BBNs) (Pearl 1988). BBNs are models with a network structure that can focus on the explicit representation of cause-and-effect relationships between variables, representing in this case ecosystem components. The network architecture is linked to probability distributions that allow for dealing with variability and uncertainty in the models. This is particularly useful for the description of ecological systems (Regan 2002). Despite some controversy (Dennis 1996), Bayesian statistics have proven useful in ecology for evaluating and managing wildlife species and forests (Cohen 1988, Haas et al. 1994, Crome et al. 1996, Lee & Riemann 1997, Marcot et al. 2001, Raphael et al. 2001), and for other areas of environmental research and management (Olson et al. 1990, Dixon & Ellison 1996, Ellison 1996, Wolfson et al. 1996, Borsuk et al. 2002, Tattari et al. 2003, Borsuk et al. 2004). More recently, a computer-based BBN system with potential of operational use in river management to diagnose river health has been developed in the United Kingdom, under the authority of the Environment Agency (Trigg et al. 2000, Walley et al. 2002).

Bayesian Belief Networks (Pearl 1988) are probabilistic models that may be based on expert knowledge, empirical data, or a combination of both. A BBN consists of a network of assumed causal relationships between variables and a set of conditional probability matrices that relates each variable to its assumed causal variables (Trigg et al. 2000). Bayes' theorem lies at the heart of Bayesian inference and it is based on the use of probability to express knowledge and combine probabilities to characterise the advancement of knowledge. The simple logical expression of Bayes' theorem stipulates that, when combining information, the resultant (or posterior) probability is proportional to the product of the probability reflecting a priori knowledge (the prior probability) and the probability representing newly acquired knowledge (the sample information or likelihood) (Reckhow 2002). Expressed more formally, Bayes' theorem states that the conditional probability y over the probabilistic outcome χ of the experiment (written $p(y|\chi)$) is proportional to the probability of y before the experiment (written $p(y|\chi)$ times the probabilistic outcome of the experiment (written $p(y|\chi)$) divided by the unconditional probability (written $p(\chi)$).

The Bayesian analysis uses prior knowledge derived from data analysis, expert judgement, or a combination of both (Bernardo & Smith 1994, Gelman et al. 1995). Conditional probability distributions (CPD) for each node need to be specified and if the variables are discrete, these can be represented as a table (CPT) that lists the probabilities of the child nodes for each combination of values of its parent nodes. A new posterior likelihood distribution is then calculated from the new data. This new posterior distribution is intermediate to the prior likelihood and becomes zero where either the prior or the likelihood becomes zero. The flows connecting the nodes, indicated by the arrows in a graphical model, can represent causal relationships and represent conditional dependency (Reckhow 2002). Conditional probability relationships can either be based on (1) experimental investigation, (2) collected field data, (3) process-based models, or (4) elicited expert judgement. When no appropriate and sufficient data exist, the elicited judgement of scientific experts

may be required to quantify some probabilistic relationships (Borsuk et al. 2002).

This paper aims at evaluating the use of BBNs for the prediction of two crustacean macroinvertebrate taxa, namely the Asellidae and Gammaridae, in rivers. Personal expert judgement was used to construct the causal network, and field data were used to calculate the conditional probability relationships. The predictive success of different BBN network architectures (one- and two-layered structures) and the dependence on specific criteria are compared. Finally, the possibilities of the further use of BBNs as decision support techniques in river management is discussed.

Material and Methods

Monitoring

The Zwalm river basin (Fig. 1) which is part of the Upper-Scheldt basin and located in a hilly landscape, is characterized by the presence of numerous small brooks. Since 1999, the water quality in the Zwalm river basin has considerably improved due to investments in sewerage and the waste water treatment plants during the preceding years. The river basin is however still significantly disturbed, mainly because of diffuse pollution originating from agricultural activities (Goethals & De Pauw 2001). The fauna in the upper parts of the Zwalm basin is of high natural value, and characterized, amongst others, by the presence of the Bullhead *(Cottus gobio)* and the Brook Lamprey *(Lampetra planeri)* and several vulnerable macroinvertebrates (e.g. mayflies of the family Heptageniidae).

During August and September 2000, 2001 and 2002, once each year 60 sites within the Zwalm river basin were sampled (Fig. 1). In 2002, 51 extra sites were monitored in three specific brooks in the Zwalm river basin (D'heygere et al. 2004). Biological sampling consisted of collecting macroinvertebrates by means of a standard handnet (IBN 1984). During five minutes, all major habitats within a stretch of 10 m are kick-sampled in a representative way (De Pauw & Vanhooren 1983). In non-wadable sites (8 in total), artificial substrates, consisting of a plastic netting filled with pieces of brick, were used for macroinvertebrate sampling (De Pauw et al. 1994). In addition, a number of environmental variables and habitat characteristics were measured (Table 1).

Selected macroinvertebrate taxa and environmental variables

In a first step, a number of models has been developed based on expert judgement implementing a selected number of variables. Gammaridae and Asellidae

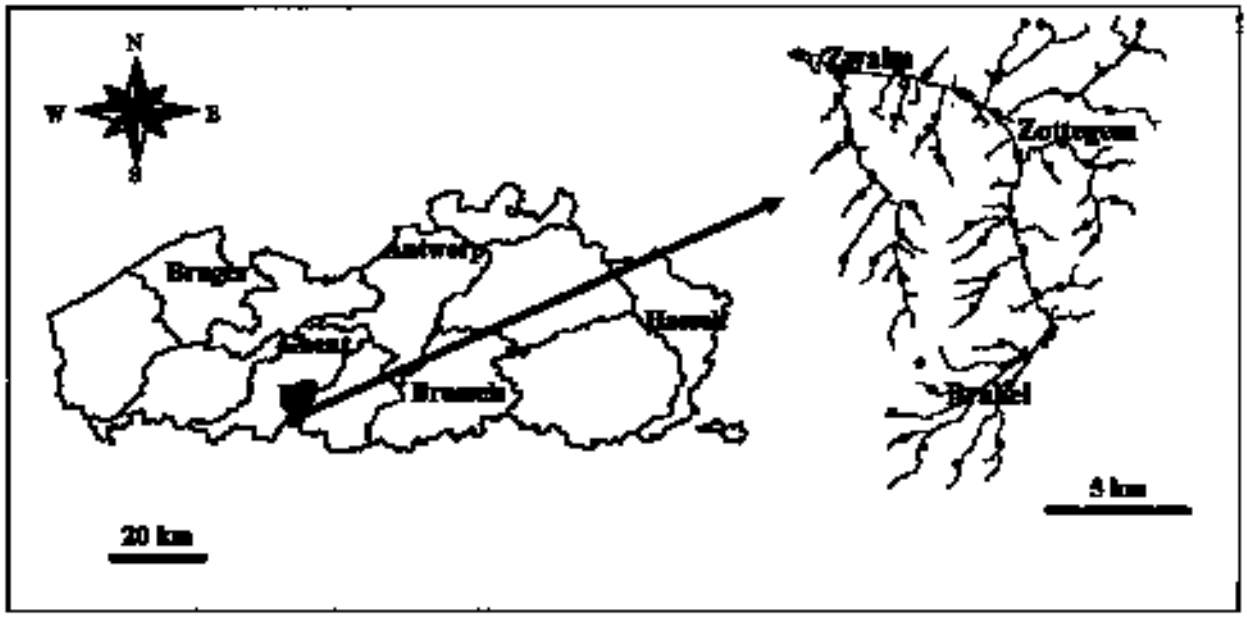

Fig. 1. Sampling points in the Zwalm river basin, located in the Upper-Scheldt basin in Flanders, Belgium.

Table 1. Environmental variables, habitat characteristics measured in the Zwalm river basin and measuring units.

Variables	Measuring units
Temperature	°C
pH	- log [H+]
Conductivity	µS/cm
Suspended solids	mg/l
Dissolved oxygen	mg/l
Water level	cm
Fraction pebbles	% surface
Fraction sand	% surface
Shadow	% surface
Aquatic macrophytes	presence / absence
Width	cm
Stream velocity	m/s
Embankment	2 categories (0 (absent), 1 (partial), 2 (total))
Meandering	6 categories (1 (well developed) to 6 (absent))
Hollow banks	6 categories (1 (well developed) to 6 (absent))
Pools / riffles	6 categories (1 (well developed) to 6 (absent))

were chosen because of their high relative abundances in small as well as large brooks in contrast to other macroinvertebrate taxa (Gammaridae and Asellidae were prevalent in respectively 79.20% and 48.49% of all sites examined). Moreover these taxa are used as bioindicators (e.g. De Pauw & Vanhooren 1983). Besides, MacNeil et al. (2002) revealed that the *Gammarus/Asellus* ratio sometimes responded to changes in parameters linked to organic pollution, but also appeared to be correlated with variables such as conductivity and distance from the source. The main influencing environmental variables selected were based on the ecological preferences of the macroinvertebrate taxa with respect to channel morphology and river type (described by either width or stream velocity), and their tolerance to nutrient and organic load (described by conductivity, measuring the total ionic concentration in a watercourse, and dissolved oxygen concentration). Because of the high pressure of organic and nutrient enrichment in the study area and more general in Flanders, Belgium, conductivity can be used as a comprehensive parameter of river pollution. Dissolved oxygen concentration was selected for implementation in the BBN despite its strong fluctuations in time. This parameter however is part of the water quality assessment of rivers in Flanders by means of the Prati index for dissolved oxygen (PIO, Prati et al. 1971, Maeckelberghe 2003).

For each of the variables, a relevant number of classes has been defined. Environmental variables and macroinvertebrate abundances were divided into four categories and the boundaries were chosen to allow a uniform distribution of the number of data points of

Table 2. Class boundaries of each variable in the Bayesian Belief Network allowing a uniform distribution of data points over the different classes.

	Class		
Variable	1	2	3
Conductivity (µS/cm)	0 - 632	632 - 790	> 790
Stream velocity (m/s)	0 - 0.17	0.17 - 0.36	> 0.36
Width (cm)	0 - 100	100 - 200	> 200
Chemical Oxygen Demand (mg O_2/l)	0 - 14.5	14.5 - 19.5	> 19.5
Temperature (T, °C)	0 - 13.9	13.9 - 16.2	> 16.2
Dissolved oxygen concentration (mg O_2/l)	0 - 5.62	5.62 - 7.78	> 7.78
Gammaridae	0 - 1	2 - 60	> 60
Gammaridae	0 - 1	> 1	
Asellidae	0 - 1	> 1	

the environmental variables regarding the discrete categories. The classes used for the BBNs are as given in Table 2. Macroinvertebrate sampling data were either divided into three discrete abundance classes (<2, 2-60, >60) or brought back to two classes (<2, ≥ 2), the latter corresponding with presence or absence of a specific taxon.

Data analysis

All BBNs were implemented using the Bayes Nets Toolbox (BNT) for MATLAB (Murphy 2001). Based on this toolbox, prior and conditional probabilities were determined for respectively child and parent nodes in the network based on the Zwalm data set. Then, the developed network was used to calculate the conditional probabilities of newly added data (validation data set). This allowed the evaluation of the developed BBNs.

First, links between the different variables were established, allowing the development of a network, based on expert knowledge. Then, for each of the child nodes X_i and parent nodes Pa_i the conditional probability distribution (CPD) $P(X_i|Pa_i)$ is specified. The most simple form of a CPD is a table, the conditional probability table (CPT), that can be used assuming all nodes are containing discrete values. All variables in this study are discrete values, as such, use could be made from the CPT to specify the conditional probabilities of the child nodes. This CPT is based on a training data set of the Zwalm river basin as defined by the ten-fold cross-validation (see later on 'model evaluation'). Assuming each parent node has the same number of classes, a CPT can be defined as a $(n^{AP}\times m)$-matrix, with m the number of classes of a child node X_i and AP the number of parents Pa_i of X_i and n the number of classes of Pa_i.

In the one-layered Bayesian Belief Networks, only Gammaridae and Asellidae were determined as being child nodes, while in the two-layered network, dissolved oxygen concentration was added as a child node besides these two macroinvertebrate taxa. The CPT for these three variables was calculated (see Murphy 2001) based on the data from the Zwalm river basin. The effect of using the conditional or either the prior probability distribution for DO on the models predictive success was investigated to examine the effect of d-separation. The criterion of d-separation determines that the nodes connected to each other by a serial connection are dependent on each other when the condition of the intermediate node is unknown. For the parent nodes, the prior probability distribution was used. Prior probability distributions that were zero for a specific class were added a very small value (0.001) to allow further calculation of the conditional probability distributions. Once the BBN was developed, it could be used for inference. For BBN inference, several algorithms exist which differ in complexity, computational efficiency, generality and accuracy. In this study, the «junction-tree» algorithm (Lauritzen & Spiegelhalter 1988) was used. This algorithm works in two steps: in the first step a junction tree is constructed (only one path is possible from one node to another during calculation), and in the second step, messages are propagated through the junction tree. More information can be found in Murphy (2001).

Model evaluation

To determine the predictive power of the models, ten-fold cross validation (Witten & Frank 2000) was used, which means that 9/10 of the data set was used to develop the BBN (training data) and 1/10 was used to evaluate the model by calculating its predictive performance (validation data). The predictive capacity of these networks was evaluated by means of (1) the CCI = the number of correctly classified instances (= matching coefficient, Buckland & Elston 1993, Fielding & Bell 1997); the outcome of the class with the highest probability value was taken and was compared with the measured class (either an abundance class or presence/absence class); (2) Cohen's Kappa = an evaluation measure which takes account of chance factors depending on the prevalence of the taxon (Cohen 1960); (3) the average misclassification score for abundance class prediction based on a cost matrix (Table 4), allowing an enhanced misclassification score the further the predicted outcome differs from the measured outcome; (4) entropy = a term used by scientists to measure the amount of randomness, disorder, or uncertainty in a population (Shannon & Weaver 1963). The performance results of a presence/absence model are normally summarised in a confusion matrix (Table 3) (Fielding & Bell 1997), however it is also applicable to evaluate prediction success of abundance class classification.

Table 3. The confusion matrix as a basis for the performance measures with true positive values (TP), false positive values (FP), false negatives (FN) and true negative values (TN).

Predicted	Actual	
	+	-
+	TP	FP
-	FN	TN

Cohen's Kappa *(K)* and the number of correctly classified instances (CCI) are based on the confusion matrix and are calculated as follows

$$K = \frac{(TP+TN) - \left[((TP+FN)(TP+FP)+(FP+TN)(FN+TN))/n \right]}{n - \left[((TP+FN)(TP+FP)+(FP+TN)(FN+TN))/n \right]}$$

$$CCI = \frac{TP+TN}{TP+FP+FN+TN}$$

CCI and Cohen's Kappa are expressed on a scale between 0 and 1. For Cohen's Kappa *(K)*, in medical applications (Landis & Koch 1977), values of $K < 0.40$ are considered to indicate slight to fair model performance, values of $0.40 < K < 0.60$ moderate, $0.60 < K < 0.80$ substantial and $K > 0.80$ excellent. However, these values are quite arbitrary and depend on the application (Manel et al. 2001).

For the abundance classification of predictive models, the evaluation was based on the CCI, K and the average misclassification score for each site (see cost matrix in Table 4).

The entropy gives an indication of the way in which values are distributed over the considered classes:

$$entropy = -\frac{1}{\log_2 n}\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

with $p_i.\log_2 p_i = 0$ if $p_i = 0$

where n is the number of classes and p_i is the proportion of data points belonging to a certain class. For example a prediction which gives a probability of 0.50 for absence and 0.50 for presence gives an entropy value of 0, whereas a prediction of probability of 0 to class 1 and 1 to class 1 gives an entropy value of 0.

Table 4. Cost matrix applied for the calculation of the misclassification score of each predicted data point as assessed by the measured data.

Measured/predicted	Class 1	Class 2	Class 3	Class 4
Class 1	0	1	2	3
Class 2	1	0	1	2
Class 3	2	1	0	1
Class 4	3	2	1	0

Results

Development of a one-layered Bayesian Belief Network

A one-layered BBN network for Gammaridae and Asellidae was developed including width or stream velocity, conductivity and dissolved oxygen concentration (Fig. 2). The networks were developed to predict presence/absence as well as the abundance class for each of the taxa. Prediction performance of the developed BBN networks, evaluated by CCI, Cohen's Kappa and the average misclassification score was calculated, and these performance measures were plotted for each model in Fig. 3. As shown in Fig. 3 the predictive success of the one-layered BBN was moderate to high for presence/absence models and low to moderate for the abundance classification scored by Cohen's Kappa. BBNs including width instead of stream velocity showed a better performance, mainly for Asellidae. The amount of uncertainty linked to each prediction with regard to both the presence/absence models and the models predicting abundance classes as measured by the entropy value, is given by box-whisker plots in Fig. 4. Entropy scores, based on the posterior probability of each class at a classified site, showed a high trend towards very uncertain classifications (entropy up to 1), mainly for abundance class prediction, and when implementing stream velocity into the model.

Development of a two-layered BBN network

In a second part of the study, a two-layered network has been constructed for the least performing one-layered network (VCDO Gammaridae, Fig. 3). The primary aim was to analyse how much the model can be improved by including an extra layer into the network representing dissolved oxygen concentration by means of its causal variables chemical oxygen demand (COD), temperature (T) and stream velocity (V). The following networks were developed (Fig. 5): (1) ignoring the effect of stream velocity (Gammaridae 1) only analysing the effect of COD, T and conductivity (C), (2) analysing a two-layered network considering COD and T influencing DO and a direct influence of stream velocity on Gammaridae (Gammaridae 2), (3) analysing a two-layered network considering stream velocity, COD and T influencing DO (Gammaridae 3), (4) analysing a two-layered network considering COD and T influencing DO combined with a direct influence of stream velocity on Gammaridae (Gammaridae 4). For these four networks, probabilities were calculated through the network based on two different approaches. The condition of DO concentration was first assumed to be known, based on the monitoring data (DO data-based, Fig. 5) and in a second phase assumed unknown and calculated through the BBN network based on COD and T (DO predicted, Fig. 5). The predictive success evaluated by their CCI score for the four networks based on these two approaches is shown in

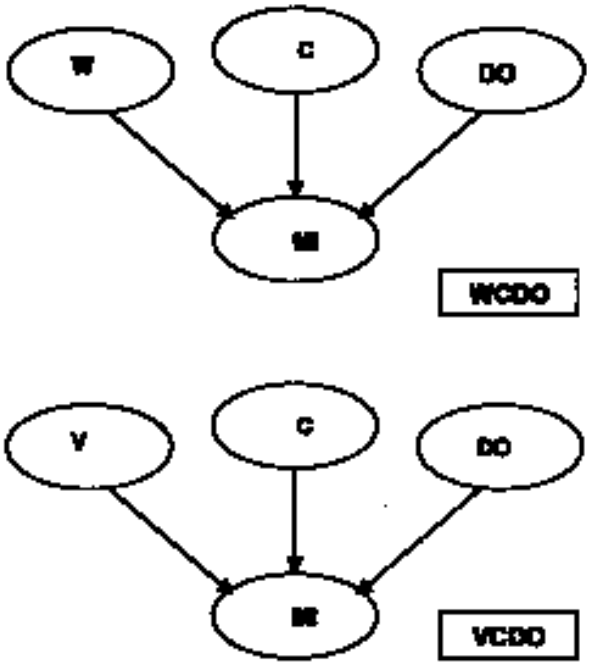

Fig. 2. BBN networks with width (W) or stream velocity (V), conductivity (C), dissolved oxygen concentration (DO) as child nodes and the macroinvertebrate taxon (MI) (Asellidae, Gammaridae) as parent nodes.

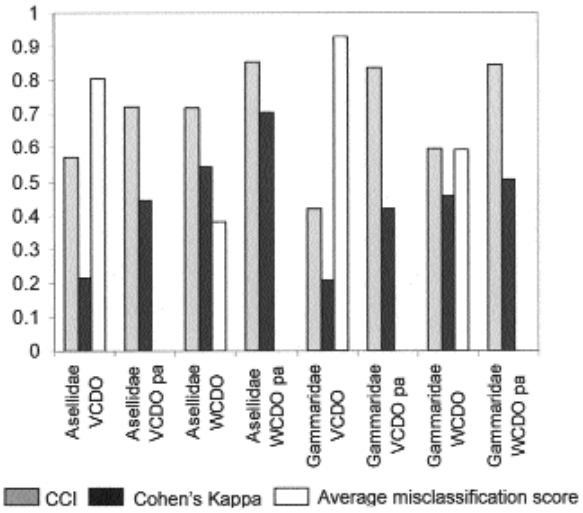

Fig. 3. CCI (correctly classified instances), Cohen's Kappa values, average misclassification score of the predictive presence/absence models for the macroinvertebrate taxa Asellidae and Gammaridae based on different input variables (WCDO = width, conductivity, dissolved oxygen concentration; VCDO = stream velocity, conductivity, dissolved oxygen concentration). pa = predictions of the presence/absence of the macroinvertebrate taxa (in contrast to abundance predictions of the other models).

Fig. 5. The criterion of d-separation determines that the nodes connected to each other by a serial connection are dependent on each other when the condition of the intermediate node is unknown. If this condition is known (DO data-based), then Gammaridae are regarded independent from COD and T and only the DO condition is of importance for the prediction. As such, when the DO condition is data-based, CCI is the same for Gammaridae 1 and 2, both relying on conductivity and DO concentration, and also for Gammaridae 3 and 4, both relying on conductivity, stream velocity and DO (Fig. 5). Unexpectedly, differences in performance of the one- and two-layered networks were very small. One could observe that the BBNs performed better when only relying on the effect of conductivity and dissolved oxygen concentration, as is shown by Gammaridae 1 (DO data-based), in contrast to Gammaridae 3 (DO data-based), actually resembling the former WCDO model (Fig. 3). However, a slightly better performance was obtained when adding a direct link between stream velocity and Gammaridae, but only when DO values were predicted from COD and T.

Discussion

In this study the aim was to evaluate the abilities of BBN modelling networks based on simple networks for presence/abundance predictions of two bio-indicator macroinvertebrate taxa (Asellidae and Gammaridae), in relation to their main influencing environmental variables. The performances of one- and two-layered networks were compared, taking into account the uncertainty coupled to the classification, the input variable selection and their conditional dependence. In comparison with other predictive modelling techniques previously applied on data from the Zwalm river basin, such as ANNs (Dedecker et al. 2004a, 2004b) and fuzzy logic (Adriaenssens et al. 2003), BBNs show a relatively good predictive success based on only three input variables. It has to be mentioned however that a large inherent uncertainty is present in these predictions. Other studies on the contrary have found BBNs to perform well as predictive models (Walley & Dzeroski 1995, Trigg et al. 2000, Fleishman et al. 2001). However, in many of these studies, the general evaluation was based on comparing the class with the highest predicted probability value to the measured class, ignoring the information inherent to the prediction outcome as expressed by probabilities or no rigorous validation was done at all (Fleishman et al.

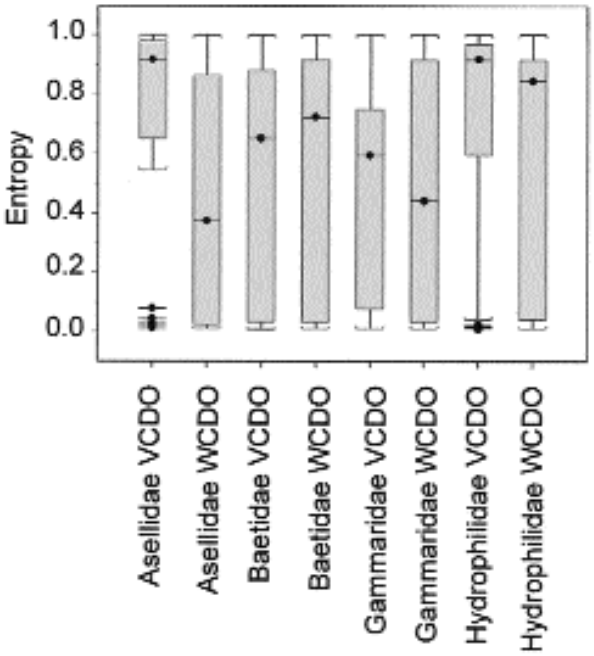

Fig. 4. Box and whisker plots of entropy values for the different models based on presence/absence classes and abundance classes, and with different input variables (WCDO = width, conductivity, dissolved oxygen concentration; VCDO = stream velocity, conductivity, dissolved oxygen concentration) for Asellidae and Gammaridae in the Zwalm river basin. pa = predictions of the presence/absence of the macroinvertebrate taxa (in contrast to abundance predictions of the other models). The box stretches from the 25th to the 75th percentile. The median is shown as a line across the box. Any individual observation that is more than 1.5 * interquartile range from the box is identified separately with a horizontal line. The whiskers extend to the maximal and minimal observations that are not potential outliers.

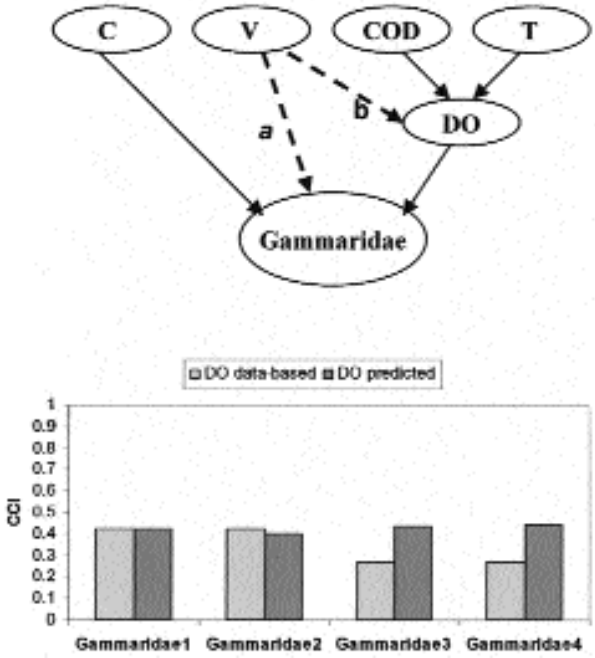

Fig. 5. Top : BBN network for the prediction of abundance classes based on the input variables conductivity (C), stream velocity (V), chemical oxygen demand (COD), temperature (T) and dissolved oxygen concentration (DO). Bottom : performance based on CCI of Gammaridae 1 = network without a and b link, Gammaridae 2 = network without link a, Gammaridae 3 = network without link b, Gammaridae 4 = network with all links.

2002). In the present study, validation of the model outcomes was done by means of the performance measures CCI, Cohen's Kappa and the average misclassification score. The uncertainty regarding the prediction by means of probabilities for each class was measured using the entropy. As expected, based on these performance measures, presence/absence models have a better performance than predictions based on abundance classes. However, in the case of Asellidae, a good performance was given by CCI and Cohen's Kappa as well as the average misclassification score when integrating width, conductivity and DO, revealing a good response of this macroinvertebrate taxon to these environmental variables with regard to abundance classes. This may indicate that Asellidae form a promising taxon for use as a bio-indicator. Considering the effect of prevalence by means of Cohen's Kappa gives a more objective evaluation of model prediction success. This is demonstrated by the relatively low Kappa coefficients for the Gammaridae models including width, although a high predictive success in terms of correctly classified instances was obtained (Fig. 3). In this case the Kappa lowers the performance, giving more weight to the falsely positive classified instances. The predicted outcome is presented as a «degree of belief» to a certain class under specific environmental conditions, and as such a specific evaluation measure is needed for evaluating the probability attached to this classification. This was based on the entropy values of each classification site, that gives a measure of how crisp the prediction into a certain class could be. As such, entropy gives an expression of the uncertainty involved in the prediction. As can be seen from Fig. 4, the entropy values for each of the models are spread over a large range and are relatively high for most of the models. Hence, the outcomes of the predictive models can be regarded as highly uncertain. Obviously, when prediction outcomes are coupled to decisions, this un-

certainty should be considered as important information.

Predictive models allow to assess the role of specific direct and indirect variables with regard to the presence/abundance of a macroinvertebrate taxon. Width and conductivity are both indirect causal variables. Indirect predictors often score the best in the predictive models by embracing different types of pollution to which the organism is subjected. However, these variables should be linked to variables causing this pollution, because measurable variables naturally have an effect on the organism in a more direct way. As such, more variables and causal links are needed. In this study, the relatively small size of the data set restricted the number of variables and their discrete states as well as the number of causal links that could be established between these variables. This lack of training data distributed over different environmental conditions prevented the establishment of a prior probability distribution. Hence, the initial network was extended by incorporating the cause-effect relations from stream velocity, T and COD on DO, with COD aiming to assess the effect of enhanced organic enrichment in the river. At the same time, COD and T are pretended to be subjected to a smaller measurement variability than DO. This extended BBN using two-layers of variables show very small differences in predictive success (Fig. 5). It is clear that the network relies mainly on conductivity and DO. Moreover, inclusion of stream velocity even has a negative effect on the predictive success of Gammaridae when used to calculate the probabilities of DO. A slightly better performance was obtained when adding a direct link between stream velocity and Gammaridae, hence only when DO values were predicted from COD and T. This means that in the prediction of abundance classes of Gammaridae, direct measurements of DO concentration combined with stream velocity as input parameter gave a less clear relation with the abundance of Gammaridae compared to the DO probability data calculated from COD and T measurements in combination with stream velocity. Furthermore, these results show that d-separation is a criterion for deciding, from a given causal graph, whether their should be «dependence» or «independence», or «connected-» or «unconnected-ness», between certain variables within the network (Pearl 2000).

Besides the use of expert knowledge to construct the causal links between the variables as was done in this study, it is recommended to use expert knowledge for setting up prior probability curves in those cases where the data set does not give sufficient information concerning these conditions. Although the BBN technique is suitable for knowledge incorporation, the transformation of this knowledge into prior probabilities is considered to be difficult (Anderson 1998). Therefore, one may recommend to put the BBN through a sensitivity analysis to find out what probabilities are being calculated trusting only on a few training data points, and to recognise for which variables expert knowledge should be included into the prior probabilities (Tattari et al. 2003). Besides the determination of the influence of specific values of prior probabilities, input variable selection is needed to reveal the importance of each variable in the network. The exclusion of a variable after all means this variable is considered to have no impact on the results. This is not such a problem if the network is completely created by experts, since the elicited probability values can only be used to reflect the effects of the variables in the model. However, when prior probabilities are based on data, the causal effect of an excluded variable may have an impact on the conditional probability distributions, hence causing inconsistencies. Another, perhaps as important reason for thoughtful input variable selection in BBNs developed for river management is the considerably effort, and therefore cost, related to an intensive monitoring campaign (D'heygere et al. 2003). It can be useful to start from simple models and then test if more complex ones would give significantly better predictions based on a selected number of data. However, a commonly applied technique is to use a kind of node absorption in a carefully structured way to create a hierarchy of abstractions of the network (Peot & Shachter 1991, Delcher et al. 1996). Several techniques exist to perform node abstraction and are explained more in detail in Vehtari & Lampinen (2002).

Conclusion

A transparent ecological model such as a BBN can improve insight into cause-effect relationships between macroinvertebrates and their main influencing environmental variables and thus be of strong additional value for river management. However, the results indicate that despite the relatively good predictive success based on only a few input variables, the outcome of the BBN was accompanied by a high degree of uncertainty. This fact demonstrates the importance of objective model evaluation, taking into account the completeness of the data, the prevalence of the organism, the input variable selection and the propagation of uncertainty through the model. It may thus be expected that thoughtful input variable selection as well as sensitivity analysis will further improve the models which can be of use in river management.

Acknowledgements

The first author holds a grant from the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT). The authors like to thank the anonymous reviewers for their interesting comments.

References

Adriaenssens V., Goethals P.L.M. & De Pauw N. 2003. - Fuzzy knowledge-based models for *Gammarus* and *Asellus* in watercourses in Flanders (Belgium). *Ecol. Model.*, in press.

Anderson J. L. 1998. - Embracing uncertainty: the interface of Bayesian statistics and cognitive psychology. *Conserv. Ecol. Online*. http://www.ecologyandsociety.org/vol2/iss1/art2/

Bernardo J. & Smith A. 1994. - Bayesian Theory. John Wiley and Sons, Chichester.

Borsuk M.E., Reichert P. & Burkhardt-Holm P. 2002. - A Bayesian network for investigating the decline in fish catch in Switzerland. Pages 108-113 in Rizzoli A.E., Jakeman A.J. (Eds.) *Integrated Assessment and Decision Support*. Proceedings of the 1st biennial meeting of the International Environmental Modelling and Software Society, Lugano, Switzerland.

Borsuk M., Stow C.A. & Reckhow K.H. 2004. - A Bayesian network of eutrophication models for synthesis, prediction and uncertainty analysis. *Ecol. Model.*, 173, 219-239.

Buckland S.T. & Elston D.A. 1993. - Empirical models for the spatial distribution of wildlife. *J. Appl. Ecol.*, 30, 478-95.

Cohen J. 1960. - A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20(1), 37-46.

Cohen Y. 1988. - Bayesian estimation of clutch size for scientific and management purposes. *J. Wildl. Manage.*, 52 (4), 787-793.

Crome F.H.J., Thomas M.R. & Moore L.A. 1996. - A novel Bayesian approach to assessing impacts of rain forest logging. *Ecol. Appl.*, 6, 1095-1103.

Dedecker A., Goethals P.L.M., Gabriels W. & De Pauw N. 2004a. - Optimization of Artificial Neural Network (ANN) design for prediction of macroinvertebrates in the Zwalm river basin. *Ecol. Model.*, 174, 161-173.

Dedecker A., Goethals P.L.M., D'heygere T., Gevrey M., Lek S. & De Pauw N. 2004b - Habitat preference study of *Asellus* (Crustacea, Isopoda) by applying input variable contribution methods to Artificial Neural Network models. *Water Res.*, submitted.

Delcher A.L., Grove A.J., Kasif S. & Pearl J. 1996. - Logarithmic-time updates and querys in probabilistic networks. *J. Artif. Intell. Res.*, 4, 37-59.

Dennis B. 1996. - Discussion: should ecologists become Bayesians? *Ecol. Appl.*, 6, 1104-1123.

De Pauw N. & Vanhooren G. 1983. - Method for biological quality assessment of watercourses in Belgium. *Hydrobiologia*, 100, 153-183.

De Pauw N., Lambert V., Van Kenhove V. & bij de Vaate A. 1994. - Performance of two artificial substrate samplers for macroinvertebrates in biological monitoring of large and deep rivers and canals in Belgium and the Netherlands. *Environ. Monit. Assess.*, 30, 25-47.

D'heygere T., Goethals P.L.M. & De Pauw N. 2003. - Use of genetic algorithms to select input variables in artificial neural network models for the prediction of benthic macroinvertebrates. *Ecol. Model.*, 160, 291-300.

D'heygere T., Goethals P.L.M., Dedecker A., Adriaenssens V & De Pauw N. 2004. -Development of a monitoring network to model the habitat suitability of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). *Aquat. Ecol.*, submitted.

Dixon P. & Ellison A.M. 1996. - Introduction: ecological applications of Bayesian inference. *Ecol. Appl.*, 6, 1034-1035.

Ellison A.M. 1996. - An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecol. Appl.*, 6, 1036-1046.

EU 2000. - Directive of the European Parliament and of the Council 2000/60/EC establishing a framework for community action in the field of water policy. European Union, The European Parliament, The Council, PE-CONS 3639/1/00 REV 1 EN, 62 p. + annexes.

Fielding A.H. & Bell J.F. 1997. - A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.*, 24, 38-49.

Fleishman E., Mac Nally R., Fay J.P. & Murphy D.D. 2001. - Modeling and predicting species occurrence using broad-scale environmental variables: an example with butterflies of the Great Basin. *Conserv. Biol.*, 15, 1674-1685.

Fleishman E., Mac Nally R. & Fay J.P. 2002. - Validation tests of predictive models of butterfly occurrence based on environmental variables. *Conserv. Biol.*, 17, 806-817.

Gelman A., Carlin J., Stern H. & Rubin D. 1995. - Bayesian data analysis. Chapman and Hall, London.

Goethals P.L.M. & De Pauw N. 2001. - Development of a concept for integrated ecological river assessment in Flanders (Belgium). *J. Limnol.*, 60 (1), 7-16.

Haas T.C., Mowrer H.T. & Shepperd W.D. 1994. - Modelling aspen stand growth with a temporal Bayes Network. *Artif. Insemen. Appl.*, 8, 15-28.

IBN 1984. - Norme Belge T 92-402. Biological water quality. Determination of a biotic index based on aquatic macroinvertebrates. Institut Belge de Normalisation, Brussels, Belgium, 11 p (in French).

Landis J.R. & Koch G.G. 1977. - The measurements of observer agreement for categorical data. *Biometrics*, 33, 159-174.

Lee D.C. & Riemann B.E. 1997. Population viability assessment of salmonids by using probabilistic networks. *N. Am. J. Fish. Manage.*, 17, 1144-1157.

MacNeil C., Dick J.T.A., Bigsby E., Elwood R.W., Montgomery W.I., Gibbins C.N. & Kelly, D.W. 2002. - The validity of the Gammarus:Asellus ratio as an index of organic pollution: abiotic and biotic influences. *Wat. Res.*, 36, 75-84.

Maeckelberghe H. 2003. - The quality of the Flemish surface water and wastewater discharges in 2002. A commentary explanation of the results of monitoring networks of the Flemish Environment Agency in 2002. *Water*, 10, 1-6. (in Dutch).

Manel S., Williams H.C. and Ormerod S.J. 2001. - Evaluating presence-absence models in ecology: the need to account for prevalence. *J. Appl. Ecol.*, 38, 921-931.

Marcot B.G., Holthausen R.S., Raphael M.G., Rowland M. & Wisdom M. 2001. - Using Bayesian Belief Networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *Forest Ecol. Manag.*, 153, 29-42.

Murphy K.P. 2001. - The Bayes Nets Toolbox for MATLAB. Computing science and Statistics, volume 33.

O'Connor M.A. & Walley W.J. 2002. - A River Pollution Diagnostic System (RPDS) for England and Wales. *Water, Sci. Technol.*, 46, 17-23.

Olson R.L., Willers J.L. & Wagner T.L. 1990. - A framework for modelling uncertain reasoning in ecosystem management II. Bayesian Belief Networks. *AI Appl. Nat. Res. Man.*, 4, 11-24.

Pearl J. 1988. - Probabilistic reasoning in intelligent systems. Morgan Kauffman Publishers, San Francisco, CA, USA.

Pearl J. 2000. - Causality. Models, Reasoning and Inference. Cambridge University Press, 500 p.

Peot M.A. & Shachter R.D. 1991. - Fusion and propagation with multiple observations in belief networks. *Artif. Intell.*, 48, 299-318.

Prati L., Pavanello R. & Pesarin F. 1971. - Assessment of surface water quality by a single index of pollution. *Wat. Res.*, 5, 741-751.

Raphael M.G., Wisdom M.J., Rowland M.M., Holthausen R.S., Wales B.C., Marcot G. & Rich T.D. 2001. Status and trends of habitats of terrestrial vertebrates in relation to land management in the interior Columbia River Basin. *Forest Ecol. Manag.*, 153, 63-78.

Reckhow K.H. 2002. - Bayesian Approaches in Ecological Analysis and Modeling. in Canham, C. D., Cole, J. J. & Lauenroth, W. K. (Eds.) *The Role of Models in Ecosystem Science*. Princeton University Press, in press.

Regan H.M. 2002. - A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecol. Appl.*, 12, 618-628.

Reichert P. & Omlin M. 1999. - A comparison of techniques for the estimation of model prediction uncertainty. *Ecol. Model.*, 115, 45-59.

Shannon C.E. & Weaver W. 1963 - Mathematical theory of communication. Urbana, IL: University of Illinois Press.

Tattari S., Schultz T. & Kuussaari M. 2003. - Use of belief network modelling to assess the impact of buffer zones on water protection and biodiversity. *Agr., Ecosyst., Environ.*, 96, 119-132.

Trigg D.J., Walley W.J. & Ormerod S.J. 2000. - A prototype Bayesian belief network for the diagnosis of acidification in Welsh rivers. ENVIROSOFT 2000, Bilbao, Spain. In: Brebbia C.A., Ibarra-Berastegui G., Zannetti P. (Eds.), Development and Application of Computer Techniques to Environmental Studies.

Vehtari A. & Lampinen J. 2002. - Bayesian input variable selection using posterior probabilities and expected utilities. Helsinki University of Technology, Laboratory of Computational Engineering publications.

Walley W.J. & Dzeroski S. 1995. - Biological monitoring: a comparison between bayesian, neural and machine learning methods of water quality classification. Pages 229-240 in R. Denzer, G. Schimak & D. Russell, Eds. *Environmental Software Systems*. Chapman & Hall, London.

Walley W.J., O'Connor M.A., Trigg D.J. & Martin R.W. 2002. - Diagnosing and predicting river health from biological survey data using pattern recognition and plausible reasoning. R&D Technical Report E1-056/TR. Environment Agency, Swindon, UK.

Witten L.H. & Frank E. 2000. - Data mining: Practical machine learning tools and techniques with Java implementations. Morgan Kaufmann Publishers, San Francisco.

Wolfson L.J., Kadane J.B. & Small M.J. 1996. - Bayesian environmental policy decisions: two case studies. *Ecol. Appl.*, 6, 1056-1066.