18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Analysing biological, chemical and geomorphological interactions in rivers using Structural Equation Modelling

S. Bizzi[1], B. Surridge[1], D.N. Lerner[1]

[1] *Catchment Science Centre, The University of Sheffield, UK*
Email: s.bizzi@sheffield.ac.uk

Abstract: Effective management of river ecosystems requires knowledge of the interrelationships between biological, chemical and geomorphological processes and patterns. This is a complex challenge, and there are significant gaps in our understanding of these interrelationships. For example, the response of biological communities to geomorphological changes in rivers is particularly poorly understood. These knowledge gaps are compounded by the lack of coherent biological, chemical and geomorphological datasets for many rivers, limiting the extent to which traditional data analysis and modelling techniques can be applied.

Here we describe the application of a new technique, Structural Equation Modelling (SEM), to the investigation of biological, chemical and geomorphological data collected from rivers across England and Wales. The SEM approach is a multivariate statistical technique enabling simultaneous examination of direct and indirect relationships across a network of variables. Further, SEM allows a-priori conceptual or theoretical models to be tested against available data. For example, a-priori models can be developed in collaboration with river managers and then evaluated using SEM as part of participatory modelling projects. This is a significant departure from the solely exploratory analyses which characterise other multivariate techniques. Bayesian approaches can also be applied within the SEM framework, offering the opportunity to address challenges such as incomplete datasets and non-normal data distributions. Such challenges are common in the analysis of spatial patterns associated with riverine ecosystems.

We took biological, chemical and geomorphological data collected by the Environment Agency for 700 sites in rivers across England and Wales, and created a single, coherent dataset suitable for SEM analyses. Biological data cover benthic macroinvertebrates, chemical data relate to a range of standard parameters (e.g. BOD, dissolved oxygen and phosphate concentration), and geomorphological data cover factors such as river typology, substrate material and degree of physical modification. We developed a number of a-priori theoretical models based on existing understanding of river ecosystems. These models were able to explain correctly the variance and covariance shown by the datasets, proving to be a relevant representation of the processes involved. The models explained around 80% of the variance in indices describing benthic macroinvertebrate communities. Dissolved oxygen was of primary importance, but geomorphological factors, including river habitat type and degree of habitat degradation, also had significant explanatory power. The model produced new insights into the relative importance of chemical and geomorphological factors for river macroinvertebrate communities. The SEM technique proved powerful, for example able to deal with the co-correlations that are common in rivers due to multiple feedback mechanisms.

In this paper we also examine how SEM could be used to guide data collection and support decision-making (DM) for river ecosystems. We highlight the benefits of a Bayesian approach to solving SEM, especially in the context of supporting DM. We demonstrate how both simple and more complex a-priori conceptual models can be used in SEM. We explore whether greater complexity, which may add credibility to a model, increases explanatory power compared to relatively simple models. We examine how subjective judgements that are inherent to the development of a-priori models, for example relating to the separation between individual habitat types, influence the outcomes and interpretations of SEM analyses. Our experience highlights the importance of close collaboration with potential users throughout each step of the SEM framework, and we examine how this collaboration might be put into practice.

Keywords: *Fluvial ecology, geomorphology, fluvial habitat, structural equation modelling, river management, decision making, modelling.*

1. INTRODUCTION

Biologists and geomorphologists are increasingly combining their research efforts to improve our theoretical understanding of the complex interactions between geomorphological, chemical and biological processes within fluvial ecosystems (Vaughan et al., 2007). Research at both local and catchment/regional scale has contributed to our knowledge of key drivers, processes, temporal and spatial scale interactions affecting these ecosystems. Despite these advances, our understanding of geomorphological, chemical and biological interactions in rivers remains limited. Compared to the physically-based models that simulate hydrological and hydraulic conditions in rivers, or to water quality models, in fluvial ecology models are often data driven, black box, or expert based. Exploratory multivariate statistical analyses are often used to analyse fluvial ecosystems (Borcard et al., 1992; Turak and Koop, 2008) primarily because of the large number of variables involved and our partial theoretical understanding of these systems. Such analyses do not always lend themselves to hypothesis testing, and do not always allow clear process interpretations of the outcomes to be made. These are two key challenges if our understanding of fluvial ecosystems is to continue to advance.

Although partial and uncertain, our knowledge of fluvial ecosystems remains crucial for management that increasingly seeks to deliver integrated, cost-effective policies and actions that produce multiple benefits. However, there is a clear need to create better vehicles for knowledge exchange amongst scientists and decision makers. Significant improvements in the field of decision support systems (DSS), especially related to water resource management, have been made in the last thirty years (Castelletti and Soncini-Sessa, 2007; Loucks and Van Beek, 2005). Nevertheless, the uptake and application of DSS by decision makers remains limited. Amongst a number of reasons for this is the complexity of many DSS, which necessitates training and expert knowledge to run the DSS and to interpret the results (Newman et al., 1999). An efficient trade-off between complexity – for example a more complete representation of spatial and temporal process interactions - and simplicity – towards the need for synthesis and clarity in representing our understanding of any system in a decision-making context – is a key challenge. A framework for model development that enables such trade-offs to be made could significantly advance the development of useful models.

This paper aims to use Structural Equation Modelling (SEM) to test a number of theoretical models regarding geomorphological, chemical and biological process interactions in rivers, and how these interactions give rise to spatial patterns in these ecosystems. Benthic macroinvertebrates have been chosen as a specific group used in assessments of rivers that are expected to respond to both chemical and geomorphological conditions, and as a link between primary producers and higher organisms. Our application of SEM adopts a confirmatory approach, developing and testing process-based conceptualizations of the system. The paper also aims to discuss the strengths of the SEM framework that we believe make it well suited as a tool to stimulate interaction between scientists and decision makers in the context of fluvial ecosystems.

2. THE SEM FRAMEWORK AND MODEL RESULTS

SEM is a multivariate statistical technique that encompasses path and factor analysis (Grace, 2006). Within the SEM framework, a-priori conceptual or theoretical models are evaluated against data. These a-priori models create an expected covariance structure, which is tested against the covariance matrix from observed data. Note that in contrast to conventional statistical models, where rejection of a null hypothesis is sought, one objective of SEM is *acceptance* of the null hypothesis (i.e. where $p \geq 0.05$ the model provides an acceptable fit to the data when testing at the 0.05 level of significance). An optimization algorithm, the Maximum Likelihood method was used in this paper, fits the parameters of the model to minimize the difference between the observed and model-predicted covariances. SEM uses the covariance structure to infer process interactions within a system, and as such is limited mainly by the completeness and quality of the data available to describe a system. Because of this, the models developed with SEM must be based on robust theoretical understanding of process interactions within a system, rather than simply using potentially spurious covariances to identify additional process interactions. An a-priori model can be accepted, rejected, or modified based on the outcomes of the analyses. The particular advantages of SEM for our application include: the ability to test direct and indirect effects of explanatory variables on dependent variables; the incorporation of latent and composite variables which are variables that are of conceptual or theoretical interest yet are not measured directly; and the use of a Bayesian approach to deal with non-normality and incomplete datasets.

Three national datasets developed by the Environment Agency of England and Wales (EA) have been analysed in this work: the biological and the chemical General Quality Assessment (GQA) databases, and the river habitat survey (RHS) database. RHS is a methodology developed in UK to assess the physical characteristics of rivers (Raven et al., 1998). GQA chemistry and RHS sites were chosen where they were within 500 m of GQA biology sites. We developed a complete database of 370 sites and a second database of 750 sites where some chemical data were missing. Several a-priori models were developed and tested using SEM: a chemical model, a

fluvial habitat model and a unified model (see Figure 1b-d). We also developed a multiple regression model (Figure 1a) to compare with the process based models. Indices were used to describe macroinvertebrate community composition. We chose Average Score Per Taxon (ASPT) as an index sensitive to organic pollution (Armitage et al., 1983), and Lotic-invertebrate Index for Flow Evaluation (LIFE) as an index potentially sensitive to geomorphological condition (Extence et al., 1999). Our SEM analyses sought to explain spatial variation in these indices, using "benthic macroinvertebrates" as a latent variable with observed ASPT and LIFE as indicators.

The chemical model (Figure 1b) uses observed 90th percentile BOD concentration (BOD90), annual average orthophosphate concentration (Orth avg), and 10th percentile oxygen saturation (Ox10) as predictor variables. These data are taken from 3 years of monthly sampling prior to the collection of biological data at the relevant GQA biology site. Biological data comes from an average over one year where each site is sampled twice, in autumn and spring. The latent variable "effective oxygen" was created as a biologically-relevant representation of dissolved oxygen concentration. Although Ox10 is an indicator variable for this latent, effective oxygen conceptually includes additional parts of the temporal distribution of dissolved oxygen concentration that are not captured by the observed data, for example sags in dissolved oxygen concentration during the night. BOD90 and Orth avg, represented by a composite variable 'Chemicals', are assumed to directly influence ASPT, but also to indirectly influence both ASPT and LIFE through their control on dissolved oxygen concentration. The influence of BOD90 on dissolved oxygen is self-explanatory. Orth avg is assumed to influence dissolved oxygen concentration through control on rates of primary productivity and thereby rates of respiration and decomposition of organic material. This model provided a satisfactory fit to the observed data ($p \geq 0.05$), and the relative importance of individual paths is shown by the path coefficients in Figure 1b.

UNIFIED MODEL						
Total effects	Chemicals	Orth avg	BOD90	Effective Oxygen	substrate	HMScore
Benthic macroinv.	-0.275	-0.089	-0.226	0.624	0.598	-0.188
ASPT	-0.453	-0.146	-0.373	0.435	0.418	-0.131
LIFE(F)	-0.262	-0.085	-0.215	0.594	0.57	-0.179
Direct effects	Chemicals	Orth avg	BOD90	Effective Oxygen	substrate	HMScore
Benthic macroinv.	0	0	0	0.624	0.28	-0.188
ASPT	-0.262	0	0	0	0	0
LIFE(F)	0	0	0	0	0	0
Indirect effects	Chemicals	Orth avg	BOD90	Effective Oxygen	substrate	HMScore
Benthic macroinv.	-0.275	-0.089	-0.226	0	0.319	0
ASPT	-0.192	-0.146	-0.373	0.435	0.418	-0.131
LIFE(F)	-0.262	-0.085	-0.215	0.594	0.57	-0.179

Table1. Total, direct, and indirect standardized effects on ASPT, LIFE and the latent benthic macroinvertebrates for the Unified model (Figure 1d). The effects reflect the change in standard deviation units of the dependent variable that is induced by a change of one standard deviation unit in the explanatory variable. These effects provide a means to assess the relative importance of different direct, indirect and total paths within the model.

The Fluvial Habitat model uses only variables from the RHS as predictor variables. We constructed the latent variable "substrate" using the observed occurrence of cobbles and boulders. This latent effectively represents a gradient of river habitat typology, moving from upland to lowland reaches with gradually decreasing bed sediment size. The composite variable "morphological features" represents a number of key habitat features within fluvial ecosystems. Finally, habitat modification score (HMScore), an index developed by the EA to account for the occurrence of engineering modifications with the surveyed site, is used as a representation of habitat degradation. Again, this model provided a satisfactory fit to the observed data ($p \geq 0.05$), and the relative importance of individual paths is shown by the path coefficients in Figure 1c.

In the unified model (Figure 1d), the chemical and fluvial habitat models have been joined together. The composite "morphological features" was removed because the influence of this variable and its components was shown to be not statistically different from zero ($p \geq 0.05$). The unified model provided a satisfactory fit to the observed data ($p=0.92$). The unified model explains 91% of the variance in LIFE and 75% of the variance in ASPT. In comparison, the multiple regression model (Figure 1a) was able to explain 58% of the variance in both indices. In contrast to the outcomes of multiple regression, our SEM analyses enabled us to identify both direct and indirect influences of explanatory variables on LIFE and ASPT (see Table 1). By creating the network structure in Figure 1d a-priori, and subsequently evaluating the network against observed data, we were also able to test our conceptual understanding of the system far more effectively than could be achieved through the multiple regression analyses in Figure 1a.

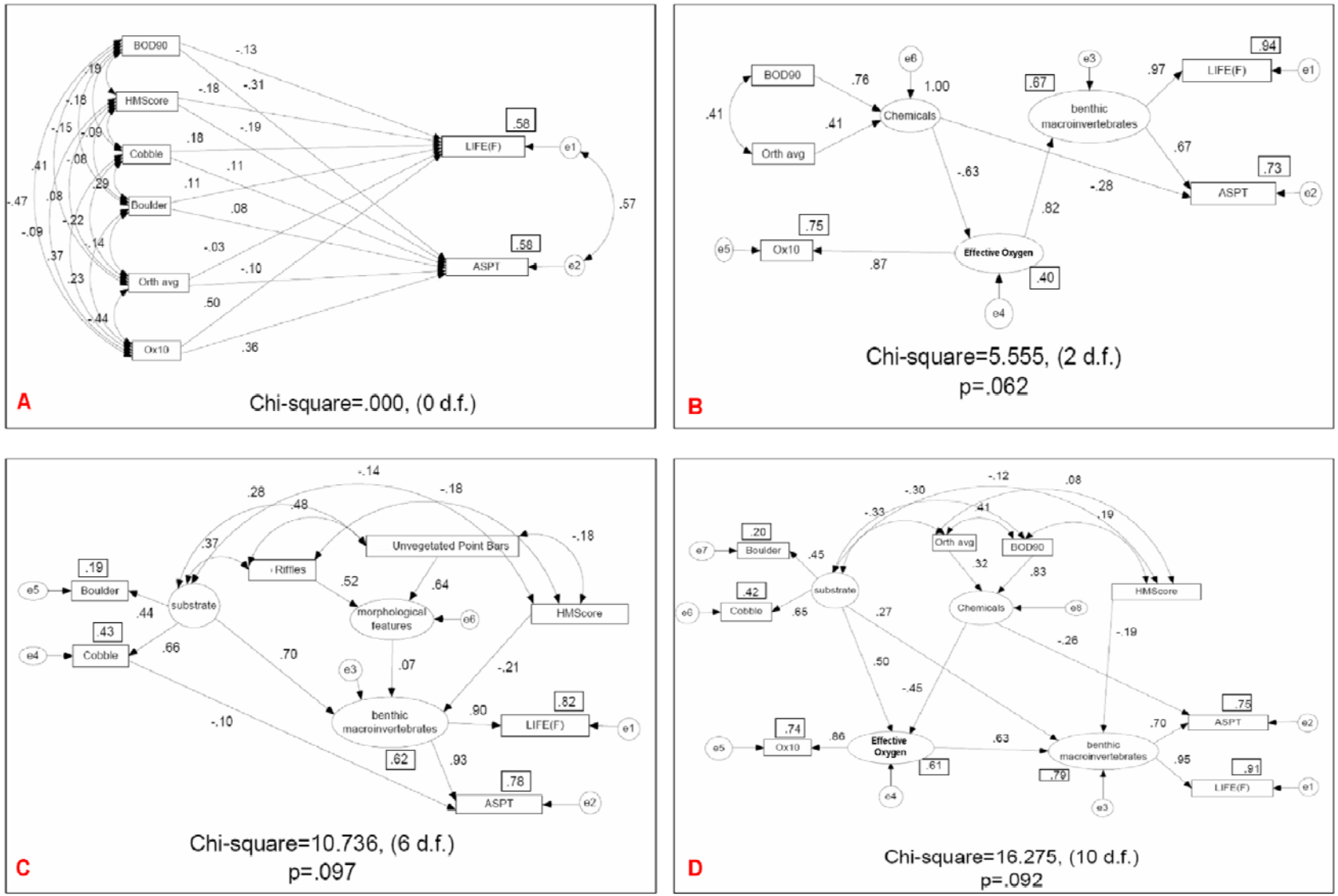


Figure 1. The four models developed: a) Regression model; b) Chemical model; c) Fluvial Habitat model; d) The Unified model. Numbers next to single-headed arrows are standardized path coefficients, next to double-headed arrows are co-variances, and in boxes are R^2 values. Standardised path coefficients reflect the change in standard deviation units of the dependent variable due to a change of one standard deviation in the explanatory variable. They provide a means of assessing the relative importance of different paths in the model. Chi-square, degrees of freedom, and p values are also shown. Where p≥0.05 the model provides an acceptable fit to the data when testing at the 0.05 level of significance. Legend for the chemical variables: 90th percentile BOD concentration (BOD90), annual average orthophosphate concentration (Orth avg), and 10th percentile oxygen saturation (Ox10)

3. THE ROLE OF LATENT VARIABLES

Latent variables within SEM enable components of theoretical or conceptual interest to be included within the network, even if they are not directly measured. These latent variables are usually associated with a number of indicator (observed) variables. These indicator variables are assumed to be reasonably, although not perfectly, correlated with the latent variable, and to be able to provide information about the latent variable (Grace, 2006). Three latent variables have been created in the work described in this paper:

- The variable "benthic macroinvertebrates" draws on the two biological indices LIFE and ASPT. Although the indicators showed different degrees of sensitivity to individual predictor variables, for example LIFE was particularly sensitive to oxygen concentration and ASPT to BOD concentration, they generally showed similar response trends to the same predictor variables. Conceptually, this latent variable represents an integrated assessment of benthic macroinvertebrate community composition, and from the decision-making perspective provides a simplified interpretation of the biological consequences of chemical and geomorphological processes. From an academic perspective, the use of this type of biological latent variable offers the potential to create and test integrated biological indices to evaluate the impact of chemical and geomorphological processes on multiple groups of organisms.
- The latent variable "effective oxygen" is a biologically-relevant representation of dissolved oxygen concentration. Conceptually it includes parts of the temporal distribution of dissolved oxygen concentration, such as night-time dissolved oxygen sags, that are not captured by the observed dissolved oxygen data. Using this latent variable within our models provided both an increase in the variance of LIFE and ASPT explained (approximately 91% for LIFE and 75% for ASPT compared to 58% for both indices in the regression model). The strength of the path coefficients describing the effect of oxygen on benthic macroinvertebrates are also increased by the introduction of the latent variable effective oxygen. In the regression model the coefficients were 0.36 for ASPT and 0.5 for LIFE (Figure 1a), in Figure 1d the coefficients were 0.44 for ASPT and 0.60 for LIFE. A further attractive feature of latent variables within the SEM framework is the ability to incorporate measurement error related to the indicator variables within the analysis, for example e5 in Figures 1b and 1d. Although we have not included measurement error in our analyses to date, because of the lack of realistic estimates for the observed data, including such estimates would enable us to correct for the downward bias that is likely to be present in both the path coefficients and estimate of variance explained because of measurement error.
- The latent variable "substrate" in the fluvial habitat model has two indicator variables, the occurrence of boulders and of cobbles. Our initial analysis of the RHS database suggested approximately 40 variables describing the physical characteristics of a river could be useful for our analyses. However, the covariance between many of these variables was high, and inclusion of each individual variable within our SEM analysis was not useful. By creating the latent variable substrate we have simplified the network of variables to be understood without losing explanatory power – the variance in LIFE and ASPT explained by the fluvial habitat model in Figure 1c was slightly higher (55%) that in a multiple regression based on all 40 potentially useful variables in the RHS. The latent substrate is likely to represent a gradient of river habitat typologies, moving from coarse bed material in upstream reaches to finer bed material in lowland reaches. Other latent variables describing a similar gradient, but composed of different indicator variables, could have been created. The particular choice of latent and associated indicator variables illustrates one example of subjective choice affecting the SEM analyses. The conceptual and theoretical differences between individual latent variables and their specific indicators, and the impacts of these choices for the SEM analyses, have not been investigated in our work, but deserve further study.

4. SIMPLICITY VERSUS COMPLEXITY IN MODEL DEVELOPMENT

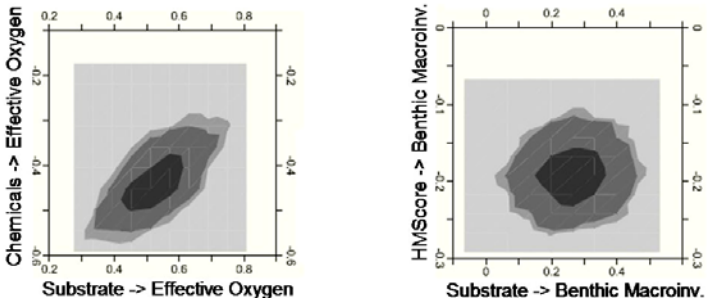
The SEM framework provides a number of structural ways of balancing complexity and simplicity within the model network. The use of latent variables enables theoretical concepts to be included, supporting aggregation and increased simplicity within a network. Composite variables (Grace, 2006), such as the variable "Chemicals" in Figure 1b, also enable the combined impact of multiple observed variables to be aggregated and simplified. Both latent and composite variables enable the user to focus on more general processes and concepts that frequently become 'lost' in complex networks of observed variables.

Complexity and simplicity are also relevant elements of the approach taken to the overall modelling framework. In our work we chose to develop two sub-models independently (the chemical and the fluvial habitat models), and only later to combine them in a unified model. This allowed us to interpret the outcomes of analyses of the individual models, as well as to recognise and explain changes to the individual models when they were combined. Two examples serve to illustrate these points:

- In Figure 1c, a direct link exists between the variables "Cobble" and "ASPT". This link was added to improve model fitting, and represents a covariance between the presence of cobbles and ASPT. One possible explanation for this link is that the variable Cobble is acting as a proxy for other aspects of the system that affect ASPT, in particular chemical conditions. However, the path coefficient is negative, suggesting that higher presence of cobbles is associated with lower ASPT. This would seem counter-intuitive – greater presence of cobbles would be expected in upstream reaches where chemical pollution is generally less severe than in lowland reaches, and this would be expected to result in a higher value of ASPT. However, in the unified model the direct link between Cobbles and ASPT is no longer present, and is replaced by a negative covariance between "Substrate" and Orth avg and BOD90.
- The fluvial habitat model included the influence of the composite variable "Morphological features". However, within the unified model the influence of this composite variable was found to be insignificant, and the network could be simplified through the removal of "Morphological features" without decreasing the fit of the model to the observed ASPT and LIFE data. Such simplification of the model network would have been a far more complicated task if all possible variables had been introduced within a single network from the outset, primarily because of the high degree of covariance between many of the variables.

5. THE POTENTIAL OF A BAYESIAN APPROACH

Adopting a Bayesian approach to solving SEM can bring a number of benefits. Bayesian approaches are particularly well suited to the challenges of incomplete datasets and non-normal data distributions. Non-normality was a particularly relevant challenge for data obtained from the RHS which are often based on simple occurrence of specific features within a river. The Bayesian approach also allows the incorporation of prior knowledge regarding the probability distribution of values for individual variables within the network. This could be of particular benefit for future applications where prior distributions may be known or extrapolated from previous studies and/or data collection. In contrast to many other multivariate approaches, this offers the opportunity to explicitly incorporate existing knowledge within a statistical analysis of fluvial ecosystems, and is consistent with the confirmatory rather than purely explanatory nature of SEM.



(a) *(b)*

Figure 2. The marginal posterior distributions for the standardised direct effects of: a) the variables "Substrate" and "Chemicals" on the variable "Effective oxygen", and b) the variables "Substrate" and "HMScore" on the variable "benthic macroinvertebrates".

The Bayesian approach also allows uncertainty in the parameters of the model to be represented, and the impact of this uncertainty to be evaluated. It is possible to treat stochastically the metrics of latent variables and to produce intuitively a representation of uncertainty in the parameters. It is also possible to examine the reciprocal influences of parameter uncertainty within the model network. Figure 2a shows the marginal posterior distributions for the standardised direct effects of the variables "Substrate" and "Chemicals" on the variable "Effective oxygen", and Figure 2b the marginal posterior distributions for the standardised direct effects of the variables "Substrate" and "HMScore" on the variable "benthic macroinvertebrates". Figure 2b shows a symmetrical distribution of probabilities centred on the means of the two parameters, whereas Figure 2a shows a far less symmetrical distribution. The statistical interpretation is that if we assume a change in the value on one axis of either Figure 2a or 2b, the probability that the value on the second axis will also change is higher in

Figure 2a than in 2b. The process-based interpretation is that the variables Substrate and HMScore in Figure 2b each describe independent impacts upon the variable benthic macroinvertebrates. In contrast, Figure 2a suggests that the predictors Substrate and Chemicals are more strongly related, and that a change in our knowledge of one parameter is likely to result in a change to the other parameter. This is valuable and relevant information to take into account when we try to assess the relative importance of individual drivers on fluvial ecosystems, both from a scientific perspective in terms of our understanding of the system, and a management perspective, for example when trying to manage trade-offs between possible consequences of policy when they are deeply affected by uncertainty.

6. CONCLUSIONS

We have briefly presented the results of a SEM-based analysis of the relationships between biological, chemical and geomorphological components of fluvial ecosystems, drawing on data collected from several hundred sites in rivers across England and Wales. Different models based on current conceptual and theoretical understanding of fluvial ecosystems were developed and tested within the SEM framework. Although separate chemical and fluvial habitat models provided satisfactory fits to observed data, the most powerful model was developed through the combination of chemical and fluvial habitat variables within a unified model network. The SEM framework provides a range of attractive features for modelling complex systems such as rivers, including the potential of latent and composite variables for aggregating and simplifying networks towards more intuitive conceptual or theoretical versions, and the power of using Bayesian approaches to solve SEM. We believe there are significant opportunities for future application of SEM to the challenges of understanding complex process interactions within fluvial ecosystems. We believe the SEM framework could offer a vehicle for improved collaboration between scientists and stakeholders in the future, for example through the inclusion of knowledge from both groups in the development of a-priori conceptual models, and the subsequent testing of these models against available data.

ACKNOWLEDGMENTS

The data for this research was released to our office from Environment Agency. Thanks to Mark Diamond and all his staff for their useful lessons about the RHS database. Thanks to David Triggs from the Centre for Intelligent Environmental Systems at the University of Staffordshire for providing data on biological and chemical quality. Thanks to the CatSci project (Early Stage Training in Catchment Science, funded by the European Commission, Marie Curie Actions Project No.: 21149).

REFERENCES

Armitage, P.D., Moss, D., Wright, J.F. and Furse, M.T. (1983), The performance of a new water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. *Water Research*, 17, 333-347.

Borcard, D., Legendre, P. and Drapeau, P. (1992), Partialling out the Spatial Component of Ecological Variation. *Ecology*, 73(3), 1045-1055.

Castelletti, A. and Soncini-Sessa, R. (2007), Coupling real-time control and socio-economic issues in partecipatory river basin planning. *Environmental Modelling & Software*, 22, 1114-1128.

Extence, C.A., Blabu, D.M. and Chadd, R.P. (1999), River flow indexing using British benthic macroinvertebrates: a frameowrk for setting hydroecological objectives. *Regulated River-Research & Management*, 15(6), 543-574.

Grace, J.B. (2006), Structural Equation Modeling and Natural Systems. Cambridge University Press, 365 pp.

Loucks, D.P. and Van Beek, E. (2005), Water resource systems planning and management. UNESCO, 680 pp.

Newman, S., Lynch, T. and Plummer, A.A. (1999), Success and failure of decision support system: learning as we go. Proceedings of the Amrican Society of Animal Science.

Raven, P.J., Holmes, N.T.H., Dawson, F.H. and Everard, M. (1998), Quality assessment using River Habitat Survey data. *Aquatic conservation: marine and freshwater ecosystems*, 8, 477-499.

Turak, E. and Koop, K. (2008), Multi-attribute ecological river typology for assessing ecological condition and conservation planning. *Hydrobiologia*, 603, 83-104.

Vaughan, I.P. et al. (2007), Integrating ecology with hydromorphology: a priority for river science and management. *Aquatic conservation: marine and freshwater ecosystems*, DOI: 10.1002/aqc.895.