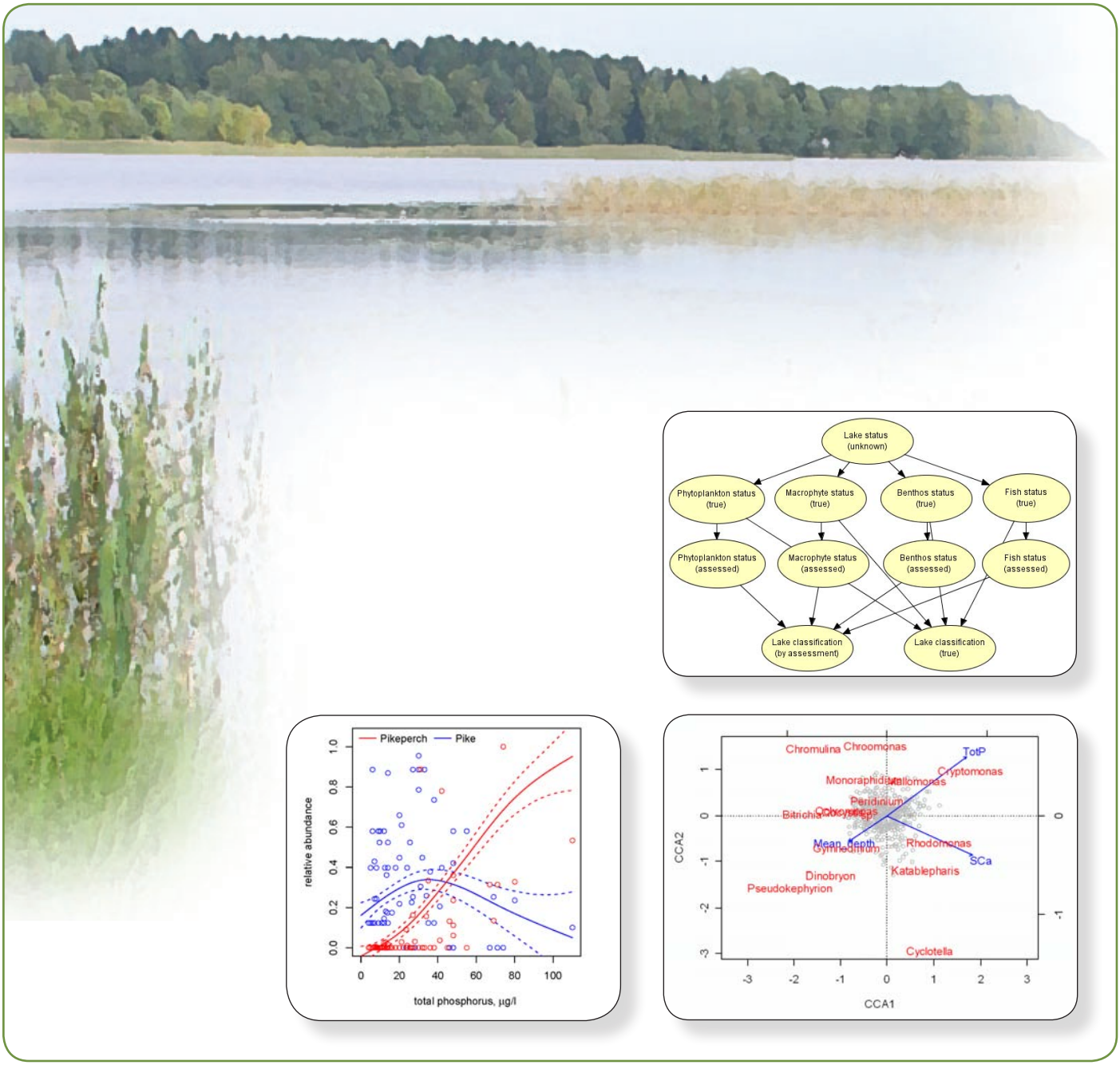
REBECCA Deliverable 12

REPORT SNO 5459-2007

Statistical and modelling methods for assessing the relationships between ecological and chemical status in different lake types and different geographical regions



Norwegian Institute for Water Research
– an institute in the Environmental Research Alliance of Norway

REPORT

Main Office
Gaustadaléen 21
N-0349 Oslo, Norway
Phone (47) 22 18 51 00
Telefax (47) 22 18 52 00
Internet: www.niva.no

Regional Office, Sørlandet
Televeien 3
N-4879 Grimstad, Norway
Phone (47) 37 29 50 55
Telefax (47) 37 04 45 13

Regional Office, Østlandet
Sandvikaveien 41
N-2312 Ottestad, Norway
Phone (47) 62 57 64 00
Telefax (47) 62 57 66 53

Regional Office, Vestlandet
P.O.Box 2026
N-5817 Bergen, Norway
Phone (47) 55 30 22 50
Telefax (47) 55 30 22 51

Akvaplan-NIVA A/S
N-9005 Tromsø, Norway
Phone (47) 77 68 52 80
Telefax (47) 77 68 05 09

Title	Serial No.	Date
Statistical and modelling methods for assessing the relationships between ecological and chemical status in lakes. REBECCA Deliverable 12	5459-2007	28.06.2007
	Project No. Sub-No. O-23035-3	Pages Price 38
Author(s) S. Jannicke Moe, Robert Ptacnik, Ellis Penning, Sakari Kuikka, Olli Malve	Topic group Ecological modelling	Distribution
	Geographical area Europe	Printed NIVA

Client(s)	Client ref.
The European Commission	SSPI-CT-2003-502158

Abstract

This report is deliverable no. 12 of the EU FP 6 project REBECCA (Relationships between ecological and chemical status of surface waters). It describes the statistical approaches and modelling methods used within Workpackage 3 (Lakes). An important purpose of the project has been to develop indicators for ecological status, and to describe and quantify the relationships between chemical pressure gradients and the biological indica-tors. Within Workpackage 3 (Lakes) we have developed indicators and analysed relationships for four main taxonomic groups: phytoplankton, macrophytes, macroinvertebrates and fish. This report describes the statistical methods used for both indicator development and relationships analyses, divided into three types of approaches. (1) Univariate methods: mainly used for describing relationships, after the response variable has been developed. (2) Multivariate methods: mainly used for data exploration and for deriving univariate response variables from multivariate community data. (3) Bayesian methods: particularly useful for analysing complex ecological systems, because hierarchical models and network models can be more easily constructed and analysed within a Bayesian framework.

4 keywords, Norwegian	4 keywords, English
1. Statistiske modeller	1. Statistical models
2. Ikke-lineære sammenhenger	2. Non-linear relationships
3. Biologiske indikatorer	3. Biological indicators
4. Bayesiske modeller	4. Bayesian models

Jannicke Moe
Project manager

Tone Jøran Oredalen
Research manager

Harsha Ratnaweera
Project and Innovation Director

ISBN 82-577-5194-4

REBECCA Deliverable 12

Statistical and modelling methods for assessing the relationships between ecological and chemical status in different lake types and different geographical regions

Edited by S. Jannicke Moe and Robert Ptacnik

Preface

This report is deliverable no. 12 of the EU FP 6 project REBECCA (Relationships between ecological and chemical status of surface waters). It describes the statistical approaches and modelling methods used within Workpackage 3 (Lakes).

REBECCA is jointly funded by the EC 6th Framework Programme (Contract number SSPI-CT-2003-502158) and the project partners.

Jannicke Moe (NIVA, Norway) has been responsible for Sections 1 and 3, and Robert Ptacnik (NIVA) has been responsible for Section 2. The following authors have contributed to the different sections:

- Tom Andersen, NIVA (Section 1.4)
- Sakari Kuikka, University of Helsinki, Finland (Section 3.2)
- Olli Malve, SYKE, Finland (Section 3.1)
- Ellis Penning, WL|Delft, Netherlands (Section 2)
- Geoff Phillips, EA, UK (Section 1.1)
- Olli-Pekka Pietiläinen, SYKE (Section 1.1)

The WP leader Anne Lyche Solheim (JRC, Italy/NIVA) has provided comments to the draft report.

Oslo, 11.01.2007

Jannicke Moe and Robert Ptacnik (editors)

Contents

Introduction

An important purpose of the project has been to develop indicators for ecological status, and to describe and quantify the relationships between chemical pressure gradients and the biological indicators. Within Workpackage 3 (Lakes) we have developed indicators and analysed relationships for four main taxonomic groups: phytoplankton, macrophytes, macroinvertebrates and fish. Examples of analyses from all groups will be given in the following chapters

The development and testing of ecological indicators for main organism groups in lakes and analyses of relationships are reported by Lyche-Solheim (2006; REBECCA Deliverable 11). Here we describe the statistical methods used for both indicator development and relationships analyses in more detail.

This report is organised according to the complexity of the methods, not according to the steps of the analyses for particular cases. Hence, we first present applications of univariate methods (where the response variable is a single variable). These methods are mainly used for describing relationships, after the response variable has been developed. Next, we present applications of multivariate methods (where the response consists of several variables). These methods are mainly used for data exploration and for deriving univariate response variables from multivariate community data. Finally, we present applications of Bayesian methods, which represent an alternative to the classical or "frequentist" statistical framework. These methods are particularly useful for analysing complex ecological systems, because hierarchical models and network models can be more easily constructed and analysed within a Bayesian framework than within the classical framework.

1. Univariate methods

Jannicke Moe

This chapter describes various statistical methods where the response is a single variable. The chapter starts with the linear methods (simple/multiple linear regression and analyses of variance). These methods are easy to apply and to interpret, and appropriate where the relationship between pressure and ecological response can be expected to be approximately linear. However, an important purpose of the project is to identify non-linear responses and threshold values in the relationships, for which more complicated non-linear methods are required. We have applied a set of different non-linear regression models (generalised linear models, generalised additive models, and quantile regression models). This combination of approaches allows for analysing different properties of the response variable (such as mean abundance, presence/absence or quantiles of abundance). Moreover, we have attempted to use non-parametric versions of these models, which require minimum of assumptions about functional forms. This means that it is possible to explore the structure of the relationships, and to discover non-linearities and thresholds, in a more flexible way than is possible for parametric models.

Non-parametric regression models generally require larger data sets than ordinary linear methods for giving reliable estimates. In WP 3 Lakes, large databases with contributions from several partners have been compiled for most of the biological elements (chlorophyll, phytoplankton, macrophytes and macroinvertebrates). It has therefore been possible to apply these data-demanding methods to analysing relationships for most of the elements.

1.1 Linear methods

Linear methods have been used most extensively for analysing relationships between TP (total phosphorous) or TN (total nitrogen) as a pressure and Chl-a (chlorophyll a) as an ecological response and indicator of phytoplankton biomass. The linear relationship between these variables is already well documented (Vollenweider 1976). The aims of the analyses within REBECCA have been to characterise the linear relationships within each of the Intercalibration lake types, and to identify the most important covariables (alkalinity, depths etc.)

To determine relationships between Chl-a and the supporting elements TP and TN, linear regression analysis was used (software: SPSS). All data were log-transformed for analysis to ensure that error terms conformed to normal distributions and to minimise heteroscedasity. To provide type-specific relationships, lakes were grouped by both GIG types and by the core intercalibration typology of alkalinity and depth. Significance of differences between type-specific relationships and the effects of other categorical variables, such as lake colour, were tested using the general linear model routine. Prior to analysis scatter plots of data by lake type were examined to determine obvious outliers, which were removed from the analysis.

As for analysis of the core lake types, regression analysis of variance was used to investigate differences between GIG types. In particular the effect of humic substances was investigated in the shallow low and moderate alkalinity lakes. Further results and interpretations are reported by Lyche-Solheim (2006).

Figure 1 shows an example of a set of linear regression fits, for different core types of lakes.

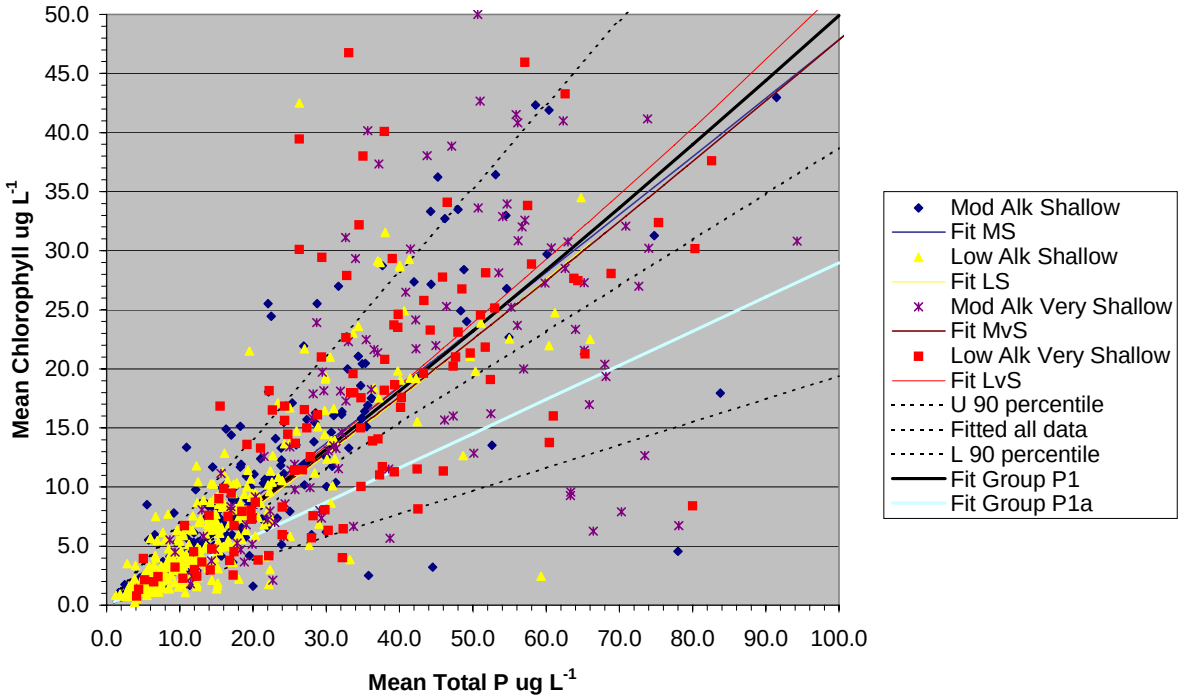


Figure 1. Scatter plot of mean growing season total P and chlorophyll a concentration for low and moderate alkalinity shallow and very shallow lakes (Group P1). Solid lines represent log-log regressions for each lake type, the bold line the log-log regression for all lakes within the group. The lower (light blue) solid line represents the regression for the significantly different sub-type of moderate altitude low and moderate alkalinity shallow and very shallow lakes (Group P1a). The dotted lines show the regression line and upper and lower 90th and 10th percentiles of points for all lakes. Figure from G. Phillips and O.-P. Pietiläinen (in Lyche-Solheim 2006).

1.2 Generalised linear models (GLM)

Simple linear models are often not suitable for describing the response of an indicator to a pressure gradient For example, the response variable may be not normally distributed but binary (e.g. presence/absence), proportions, or Poisson-distributed (e.g. counts). The framework for analysing linear models can be generalised to other distributions of the response variable y, using a link function g().

Linear models: $y = b0 + b1*x1 + b2*x2 + ...$

Generalised linear models: $g(y) = b0 + b1*x1 + b2*x2 + ...$

The link function transforms the y variable into something that can be modelled as a linear combination of predictor variables. Ordinary linear models are then a "special case" where the link function = identity.

A typical example of a GLM is logistic regression, which can be applied when the response variable is binary distributed (such as presence/absence). Such data typically result in a sigmoid, rather than linear, response to a pressure gradient. The link function "logit" can then be used to obtain a linear relationship between the predictor and transformed response variable y:

$\text{logit}(y) = \log(y / (1-y))$

While y has the range [0,1], the transformed logit(y) has range <-Inf, Inf>, and can therefore be modelled as linear combination of the predictor variables.

$\log(y/(1-y)) = b0 + b1*x$.

Back-transformation gives:

$y = \exp(b0 + b1*x)/(1 + \exp(b0 + b1*x))$,

which produces a sigmoid curve.

In some cases it is useful to transform a numeric response variable into a binary variable and apply logistic regression, e.g. to analyse whether the response is above or below a threshold value. An example is shown in Figure 2.

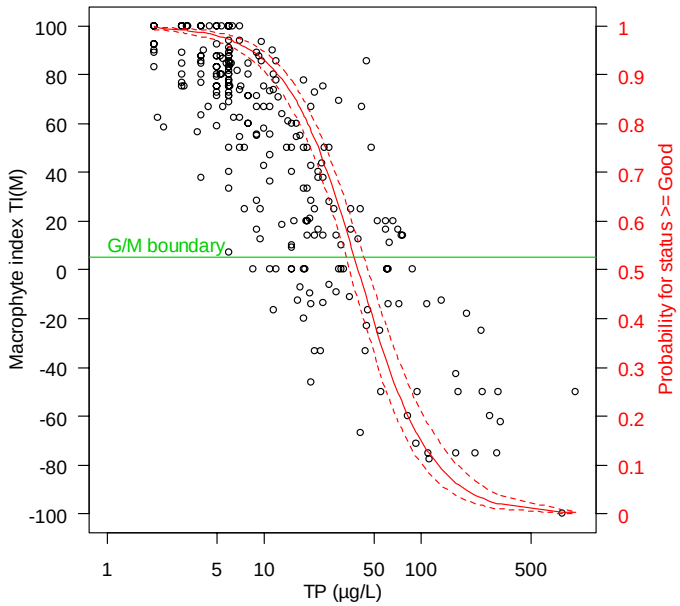


Figure 2. Logistic regression for the macrophyte index TI(M) versus total phosphorus. The response variable is transformed into 1/0, representing above/below a proposed boundary value for good/moderate status.

1.3 Additive models (AM) and Generalised additive models (GAM)

Generalised linear models allow for analysis of non-linear models within a linear framework, but this method still requires that a parametric model is specified. Additive models (AM), on the other hand, are not restricted to linear combinations of predictor variables. Additive models allow for addition of various functions of predictor variables.

Linear models (LM): $y = b_0 + b_1*x_1 + b_2*x_2 + b_3*(x_2^2) + \ldots$

Additive models (AM): $y = b_0 + f_1(x_1) + f_2(x_2) + \ldots$

The functions f() can be non-parametric, e.g. loess (locally weighted regression), or so-called splines. Non-parametric regression models provide flexible curves as estimates, but not conventional parameter estimates.

Just like shown for linear models (section 1.2), additive models can also be generalised, using a link function g():

Generalised additive models (GAM): $g(y) = b_0 + f_1(x_1) + f_2(x_2) + \ldots$

In this way, generalised linear models such as logistic regression models (Section 1.2) can be estimated in a more flexible way (see section 1.3.2).

In practice, additive models are also referred to as GAMs, even if the response variable is not transformed by a link function. The examples of AM and GAM shown in sections 1.3.1 and 1.3.2 are performed with the function "gam" the statistical package "mgcv" (Wood 2000, 2006) in the software R (Development Core Team 2006),.

1.3.1 Application of additive model to fish data

Figure 3 shows application of an additive model to fish abundance versus total phosphorus, for data from 104 Finnish lakes. The models explain 68 %, 12.5 % and 48.4 % of the deviance for silver bream, pike and pikeperch, respectively (although the effects of TP on fish are likely to be indirect, e.g. affecting food availability or visibility for predators). Estimated degrees of freedom (edf) indicate the degree of non-linearity: 1 = approximately linear, higher values = more non-linear.
The smoothed predictor variable was highly significant (p<0.00001) for both the silver bream (edf=5.1) and pikeperch (edf=2.7) models, and almost significant in pike (edf = 2.3, p = 0.062).

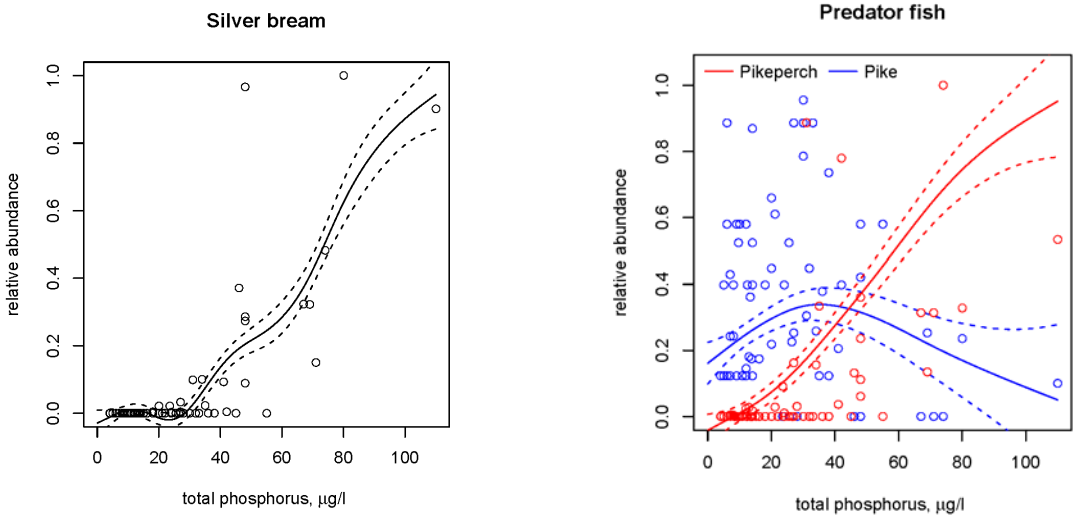


Figure 3. Additive model predictions of relative abundances (relative to maximum abundance) for silver bream (left panel) and two piscivorous species (right panel; pike and pikeperch) as function of total phosphorus. Curves with corresponding colours constitute model predictions (and dashed curves are standard errors) where relative abundance is predicted from total phosphorus concentrations (as smoothed predictor variable).

1.3.2 Application of AM and GAM to macroinvertebrate data

Figure 4 gives an example of application of an additive model for regression of a macroinvertebrate indicator (proportional abundance of taxonomic groups) against pH. A smooth spline function is used for the predictor variable pH. Note that the estimated curves can be "wiggly" (middle panel, left) as well as non-uniform or "hump-shaped" (lower panel, right).

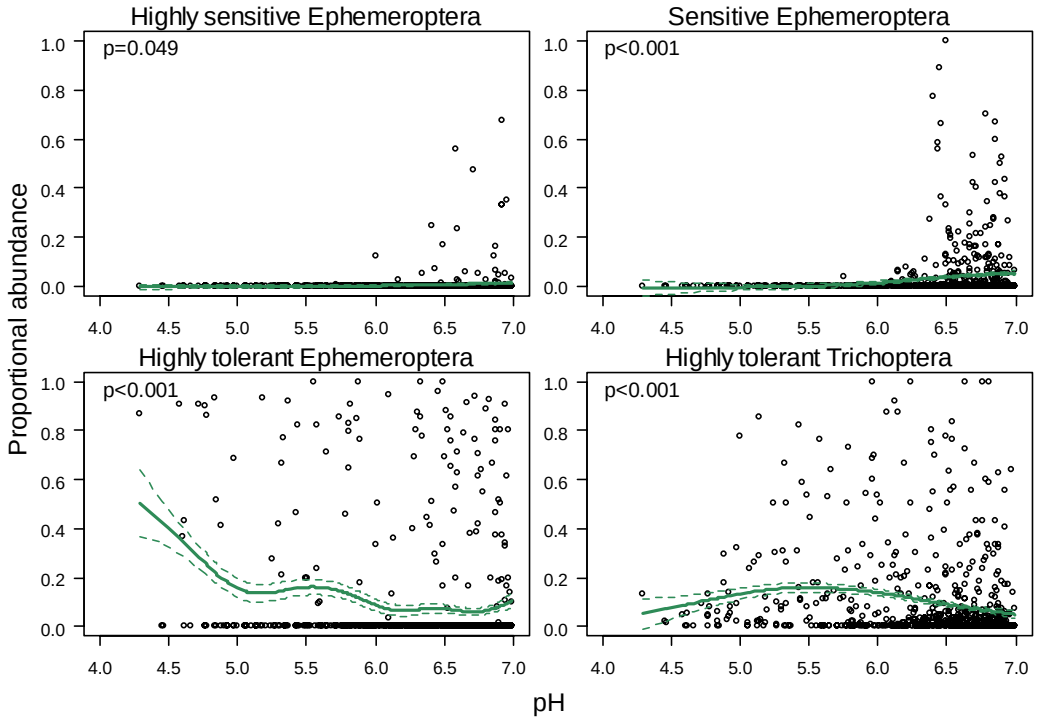


Figure 4. Application of additive regression model to macroinvertebrate indicators (proportional abundance of taxonomic groups) versus pH. The curves represent the estimated mean proportional abundance ± 2 SE.

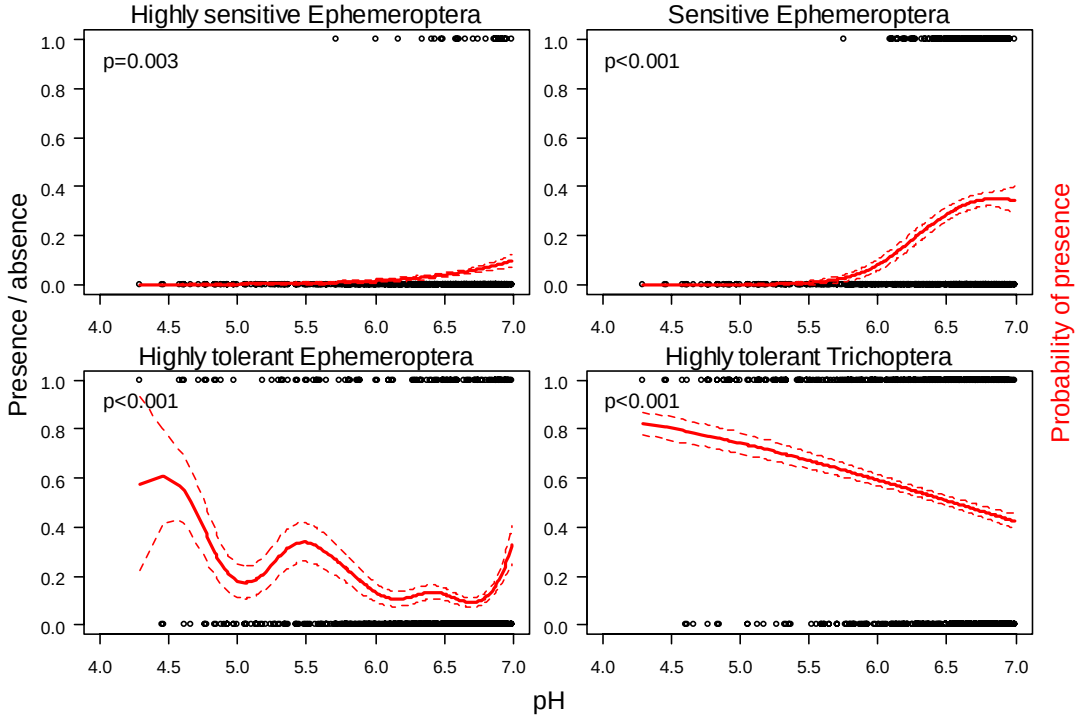


Figure 5. Application of generalised additive regression model to macroinvertebrate indicators (proportional abundance of taxonomic groups) versus pH. The curves represent the estimated probability of presence ± 2 SE.

Figure 5 shows a GAM version of logistic regression, i.e. a smooth spline function is used for the predictor variable. Note that for some of the indicators (upper panel), the transformation of the proportions (which are often low) into 0/1 data have made the response more obvious. However, although this non-parametric version of logistic regression is more flexible than ordinary parametric logistic regression, the logit link function still forces the estimated curve to be more or less monotonous (either increasing or decreasing throughout the range of the pressure gradient). For indicators with a non-monotonous response (lower panel, right), this method is therefore less suitable.

1.4 Quantile regression

Tom Andersen

Statistical models for pressure-response relationships have traditionally been focused on making interferences about central tendencies, such as the mean response conditional to pressure level, while less attention has been given to other distributional aspects of the response variable. This limitation becomes particularly evident for processes with partially unknown or unobserved explanatory variables, where pressure-response relationships become polygons rather than curves (Scharf et al.1998, Cade 2003). One would also expect to see this pattern in ecosystems exhibiting abrupt transitions between alternative stationary states (Scheffer et al. 2001). Thus, bimodality in the conditional distribution of ecosystem response to external pressures is considered a strong indicator of a threshold for ecological regime shift (Scheffer and van Nes 2004). In such situations, the worst-case response may be strikingly evident to visual inspection, even though the fitted model for the conditional mean response may not even be statistically significant. In many cases, such mismatches result from limitations of the statistical methods rather than from erroneous subjective perception of the ecological risk. The introduction of quantile regression methods (Koenker and Bassett 1978) opens new opportunities for bridging this gap by fitting statistical models to arbitrary quantiles of the conditional distribution rather than to the central tendency.

Consider the situation where we have quantities y_i measured as response to changing levels of an explanatory variable x_i, and we want to describe this relationship by some function $g(x,\theta)$ which also depends linearly on parameters contained in a vector θ. Most commonly this estimation is done by finding parameter values $\hat{\theta}$ that minimize a weighted sum of squared residuals between observations and model predictions $\sum w_i \left(y_i - g\left(x_i\right)\right)^2$. The weighted least squares estimate is known to be an unbiased and efficient estimator of the conditional mean of *y* given *x*, under certain distributional assumptions and for appropriate weights (w_i). The τ th conditional quantile regression estimate for *y* given *x* is found by minimizing the superficially similar loss function $\sum \rho_\tau \left(y_i - g\left(x_i\right)\right)$ (Koenker and Bassett 1978). The residual weighting function $\rho_\tau(\cdot)$ is called the tilted absolute value, or Koenker's check function: $\rho_\tau(u)$ is equal to τu if $u > 0$ and $(\tau - 1)u$ if $u \le 0$. Thus, positive residuals will get a higher absolute weight than negative ones for high quantiles (τ close to 1), and vice versa for low quantiles. At the loss function minimum a fraction τ of the residuals will, on the average, be negative while the remaining $1-\tau$ will be positive. This means that the fitted function will be an estimate for the conditional τ th quantile of y given *x*. The biggest technical difference between ordinary least squares (OLS) regression and quantile regression is that while the former is found by solving a system of linear equations, the latter becomes a linear programming problem, although with appropriate choice of methods the computational load is claimed to be similar for OLS and quantile regression (Portnoy and Koenker 1997). As long as *g* is linear in the parameters, both problems will have unique solutions that can be found without iterating from an initial estimate.

A recent review of ecological applications of quantile regression focuses mainly on linear models (Cade 2003), while only few reports have been made on nonlinear or non-parametric applications (e.g. Schröder et al.2005). Since biological indicators that are found useful often have strongly nonlinear responses to changes in external pressures, it would seem advantageous to have model functions for conditional quantiles with more flexibility than straight lines. It is known that any smooth function on a finite interval can be approximated by piecewise polynomials or splines (e.g. de Boor 1978). Moreover, this fit will have the desirable property of being linear in the parameters as long as the number and location of the piecewise polynomials (the knots) are fixed. Increasing the number of knots will enable the spline curve to fit perfectly through an increasing number of data points at the price of producing curves with increasingly irregular fluctuations. The balance between goodness of fit and smoothness can be controlled by a single weighing factor λ for the wiggliness penalty, typically a positive function of the second derivative of the fitted function (Wahba 1990, Hastie and Tibshirani 1986). Thus, fitting a quantile spline regression becomes equivalent to minimizing the tilted absolute value loss function with a wiggliness penalty λ: $\sum \rho_\tau \left(y_i - g\left(x_i\right)\right) + \lambda \int \left| g''\left(x\right)\right| dx$. This smoothness criterion has been shown to have superior numerical properties compared to alternatives such as the squared second derivative (Koenker et al. 1994). We have used function rqss() in the R package "quantreg" (Koenker 2006), which gives an efficient implementation of the Koenker et al. (1994) algorithm. Since quantile regression is still a relatively new method, routines for calculating standard errors etc. are not yet readily available.

Figure 6 shows an example of the application of non-linear quantile regression to the same set of macroinvertebrate indicators as Figure 4 and Figure 5, for estimates of the 90% quantile. Note that the estimated 90% quantile is more sensitive to changes in the pressure gradient than the regression based on the mean (Figure 4) for the metrics with high proportions low values (upper panel). In this case, the quantile regression is somewhat less sensitive than the more logistic regression (Figure 5), which used a stronger transformation. However, the quantile regression works much better than logistic regression when the response in non-monotonous (lower panel, right).

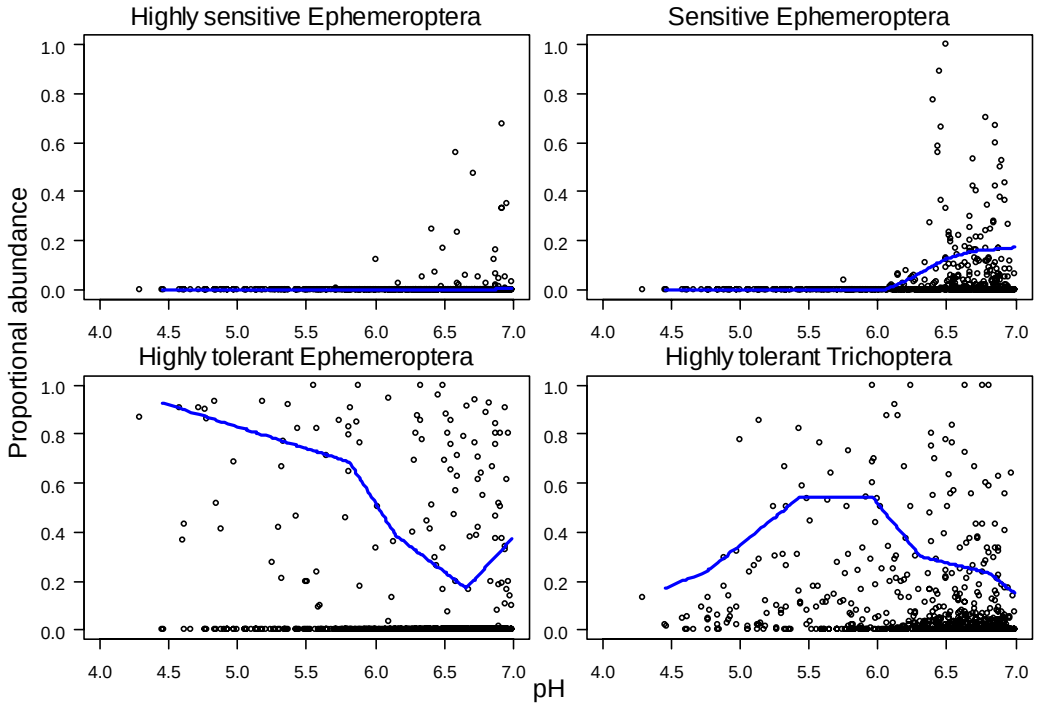


Figure 6. Application of quantile regression model to macroinvertebrate indicators (proportional abundance of taxonomic groups) versus pH. Estimated curves represent the 90% quantile of the abundance.

1.5 Model selection

An important consideration for all types of models is which and how many explanatory variables should be included in the model. If there are too few variables, the model will not be able to explain much of the variation. On the other hand, if there are too many variables, then the model will be too specific for the current data set. A guideline for model selection is: when is a more complicated model significantly better than a simpler model?

Various criteria are developed for model selection, including the following:
- Compare how much variation is explained (R2)
- Compare how much variation is left (Mallow's Cp, deviance,…)
- Compare criteria based on log-likelihood (AIC, BIC)
- Compare ability to predict new data: cross-validation

Below we give examples of model selection procedures by AIC and by cross-validation.

1.5.1 Selection of predictor variables: AIC

Akaike's information criterion (AIC) is calculated as

-2*log-likelihood + k*npar,

where npar = the number of parameters in the fitted model, and k = penalty per parameter. Lower AIC score implies better fit. As a rule of thumb, the difference in AIC value should be >2 for the difference to be significant. The AIC method requires that the log-likelihood function can be easily calculated (which is the case for e.g. GLMs). Non-parametric models like GAMs must also include parameters in the usual sense, for AIC to be applied. The models do not have to be subsets of each other, as for model selection by ANOVA, but they must use the same response variable.

The AIC provides an 'objective' measure on whether a more complex model is to be preferred over a simpler model or not. An example of model selection by AIC for phytoplankton is in given Figure 7, where two generalized additive models (GAM) are fitted for the response of proportion of phytoplankton classes to eutrophication. We compare two models, where the simpler model has only one predictor variable (a non-parametric spline function of Chl-a), and the more complicated model in addition has lake type as a categorical predictor variable (two levels: Lake type L-N1, or all other lake types). Since the second model contains one more parameter, it and gives therefore always a better fit in terms of R^2. However, the AIC criterion allows to judge whether the additional parameter is justified or not (Figure 7).

1.5.2 Degree of non-linearity: cross-validation

For non-parametric regression models such as GAMs (section 1.3), an important consideration is how "smooth" the estimated curves should be. If the curves are too smooth, important details in the data structure may be ignored. On the other hand, if the curves are allowed to be too "wiggly", then the estimated curve may become too specific for your current data set, and be less useful for more general interpretations.

Cross-validation is a flexible model selection approach that is applicable for all types of models. The procedure for cross-validation in general is as follows.

1. Specify a model (e.g. full model)
2. Exclude a subset of data (e.g. 1/10): x_{excl}, y_{excl}
3. Estimate the parameters with the remaining data: x_{incl}, y_{incl}
4. Use the x_{excl} as input in the model parameterised by x_{incl}, y_{incl} and predict y_{pred}
5. Compare the y_{pred} with the real y_{excl} , calculate the squared differences
6. Repeat for each subset of the data
7. Sum the calculated differences. This gives the CV score for this model

8. Repeat CV calculation for each model

The CV scores are then compared among the models, and the lowest CV score represents the best fit. However, this approach has no general rule for defining a significant difference between CV scores.

In the package "mgcv" (Wood 2000, 2006) that we have used for generalised additive model, the degrees of freedom are optimised by Generalized Cross Validation (GCV) criterion. The optimal degree of smoothness versus wiggliness is then solved together with the model fitting. The GCV score is calculated as:

GCV = n*D/(n - DoF)^2 ,

where D is the deviance, n the number of data and DoF the effective degrees of freedom of the model. It is also possible to replace D by the Pearson statistic, but this can lead to over-smoothing. Smoothing parameters are chosen to minimize the GCV score for the model, and the main computational challenge solved by the 'mgcv' package is to do this efficiently and reliably. The effective degrees of freedom estimated by this method give a measure of the optimised degree of non-linearity (as in Section 1.3.1).

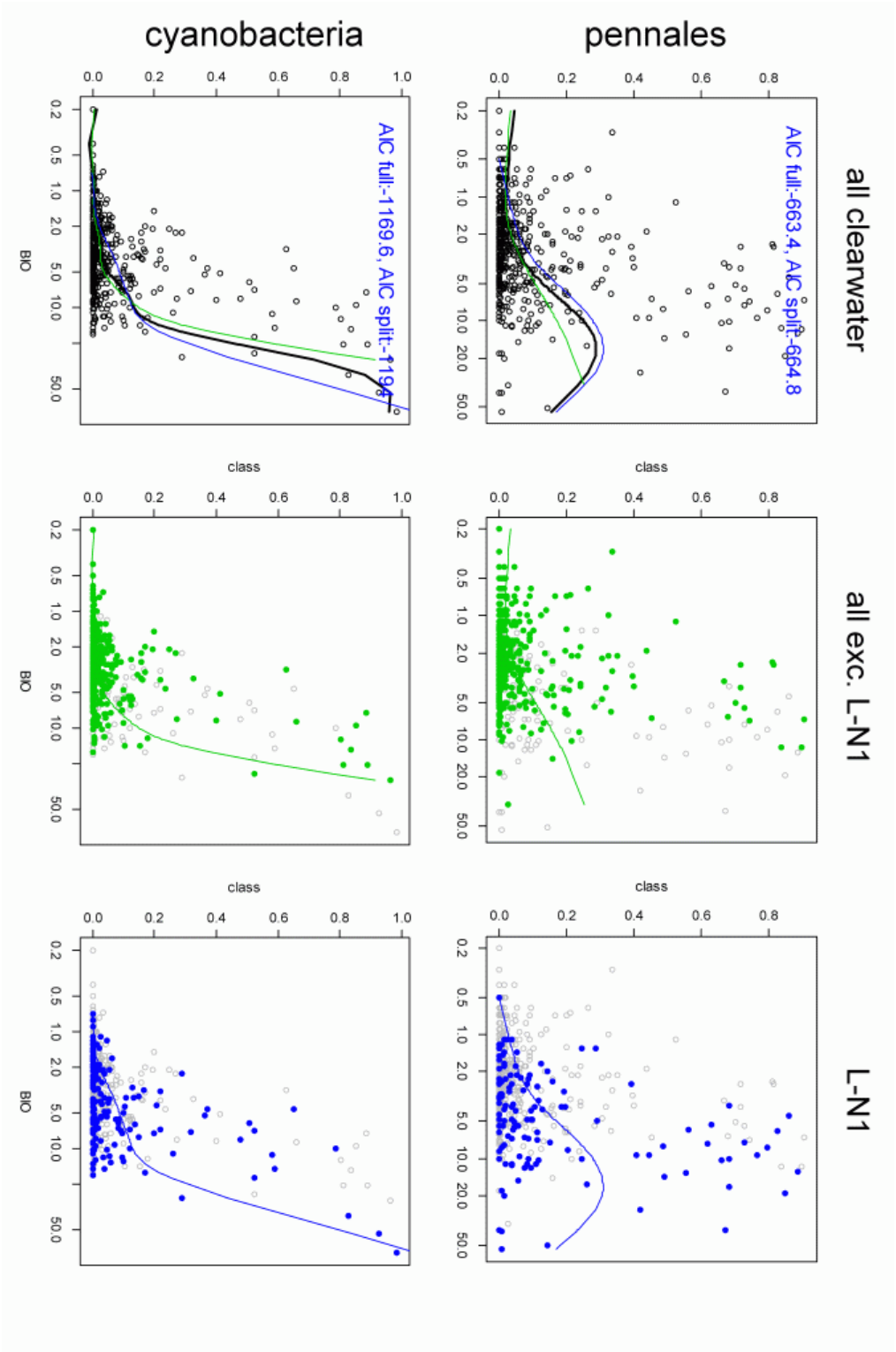


Figure 7. Model selection by AIC. The response of cyanobacteria (left) and pennate diatoms (right) is fitted along the Chlorophyll-a gradient. Blue curve: lake type L-N1; green curve: all other lake types; black curve: all lake types combined. For the cyanobacteria, the model split by lake types gives a better fit (AIC = -1169.6) than the combined model (AIC = -1194). For pennate diatoms, the model that includes lake type (AIC = -664.8) is not significantly better than the combined model (AIC = -663.4; difference <2). In the latter case, the additional co-variable is therefore not justified.

2. Multivariate methods

Robert Ptacnik and Ellis Penning

There are numerous multivariate assessment methods available, all having the following underlying assumption: Based on a dataset clusters of groups can be found by comparing the contents of the individual samples. Samples that are similar to each other are being grouped together, using matrix (dis-) similarity assessment methods. This allows for instance to identify specific groups within biological community data on various levels of detail. Often there is a suite of environmental data available for each data point of a biological dataset as well. The two datasets together give the user a large amount of information on the functioning of the studied system.

Multivariate analyses can be used for analysing the species/samples abundance (or biomass) matrices that are often used in biological monitoring of environmental impact and more fundamental studies in community ecology, together with associated physico-chemical data. The methods normally make few, if any, assumptions about the form of the data ('non-metric' ordination and permutation tests are fundamental to the approach). Most calculation methods are rather robust and this makes them widely applicable, leading to greater confidence in interpretation of community patterns. The methods have been adopted worldwide, particularly in marine science but increasingly in terrestrial, freshwater, palaeontology etc contexts.

The potential causes of the occurrence of various groups can be taken into account by superimposing the available environmental data on the biological dataset, or by interactively assessing these datasets together. Depending on the calculation method and the assumptions underlying these methods the results may be slightly different. A good understanding of the various options available is necessary to correctly interpret the results of a multivariate analysis.

Various famous applications of multivariate analyses exist. Here we mention as examples the RIVPACS methodology (Wright et al. 1997) and the BEAST (benthic assessment of sediments; Reynoldson et al. 1995), which both use benthic communities in reference rivers and lakes to define the respective reference community groups. After these reference groups are defined, communities on other, supposedly impacted sites are compared to these groups to assess the deviation from the reference.

This chapter will give some background information on the most common calculation methods used in ecological studies. More in-depth information can be found in various handbooks (e.g. Gauch 1982, Green 1979) and on websites (e.g. http://ordination.okstate.edu/).

The main use of multivariate analyses lays in the exploration of data of which structure is yet unknown. For example, when large amounts of data are compiled, as is the case in the REBECCA datasets, these techniques allow identifying community groups that are present in the database, and correlations between these groups and main explanatory factors. In addition to environmental data, secondary data on for instance collection method or country can be included, as these have a potential effect on the biological community observed at a specific sampling location.

2.1 (Dis-)Similarity analysis and Non-metric multi-dimensional scaling (NMDS)

In a similarity matrix, the relative similarity between each pair of samples is represented by a value scaling from 0 (no matching elements) to 1 (all species are represented in both samples in equal abundances). Dissimilarity is defined as similarity - 1. Similarity matrices can be used for analysis of biological data with respect to differences in community composition. They underlie popular methods such as *Metric*- as well as *Non-metric Multidimensional Scaling* (MDS and NMDS; see below), and cluster analysis. Similarity can be quantified by a number of algorithms, with the *Bray-Curtis* algorithm being the most popular one in biological sciences.

NMDS – Non-metric multi-dimensional Scaling is a multivariate method for indirect gradient analyses. Other indirect gradient methods are for example Principal Co-ordinates Analysis (PCoA), Correspondence Analysis (CA) and Detrended Correspondence Analysis (DCA). These types of computational techniques are available in various software packages, such as PRIMER, CANOCO and R (the package "vegan"). NMDS was introduced by Shepard (1962) and Kruskal (1964) and is based on the assessment of a (dis)similarity matrix among of a community. In NMDS, the ranking of similarities among samples is used for the analyses, instead of absolute distances among samples, making the method more flexible toward choice of transformation or type of data. Based on the ranking of the similarity of samples, a configuration in a specified number of dimensions is created, which attempts to satisfy all conditions imposed by the rank (dis)similarity matrix.

2.1.1 Example of use of NMDS: macrophytes

The dataset of the macrophyte working group was tested for usability and homogeneity via standard multivariate analyses (non-metric multidimensional scaling of the similarity matrices of species per sample matrices, using the programme PRIMER (Clark & Warwick 1994). Results of the initial data screening for the overall data set (1300 samples, 164 species) using NMDS show a primary clustering based on alkalinity and a second gradient of clustering based on country (stress value 0.22). For the high alkalinity sites we addressed the differences between sorting by country and by GIG-lake typology (Figure 8 and Figure 9) by using a legend that indicated either country (Figure 8) or pre-assigned lake type based on expert judgement (Figure 9).

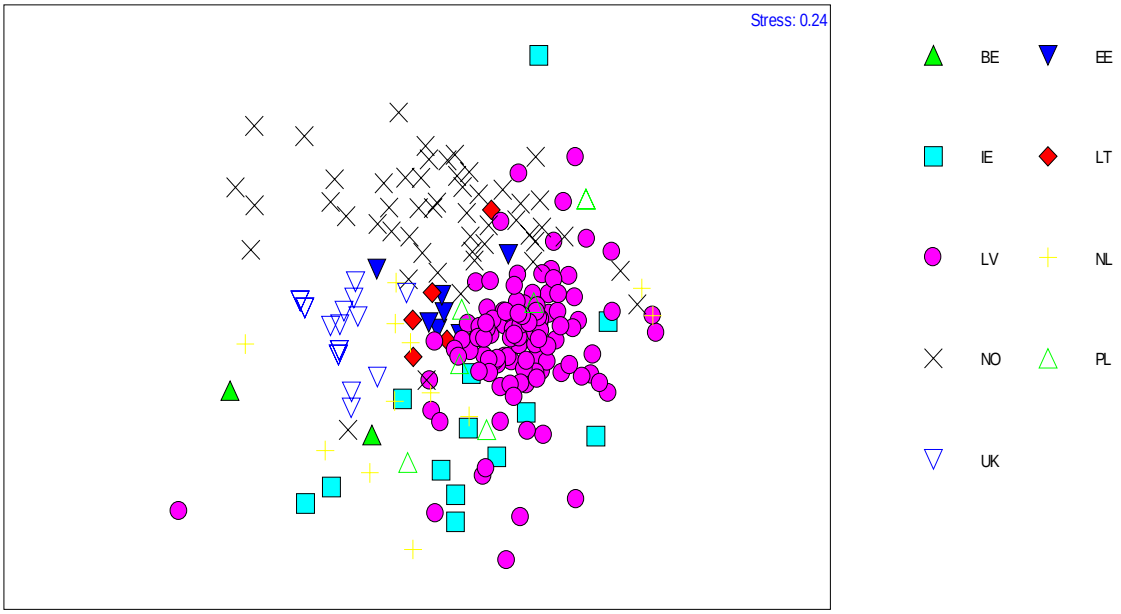


Figure 8. NMDS of species ordination of high alkalinity sites. Symbols represent different countries. Norwegian (NO) and Latvian (LV) sites are dominating the ordination.

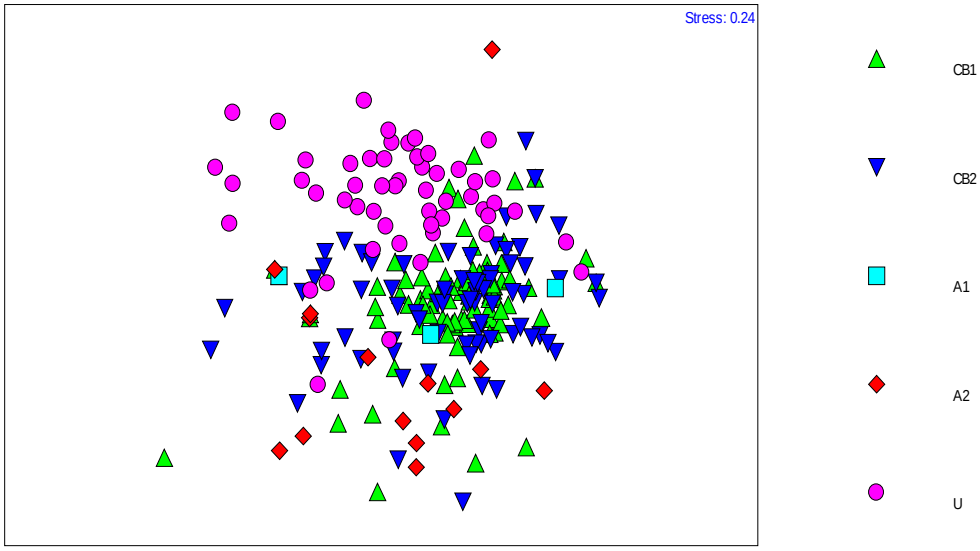


Figure 9. The same NMDS as in Figure 8, but here showing the pre-defined lake types instead of the countries. CB1 and CB2: high alkaline very shallow and shallow central lakes, A1 and A2: Atlantic lake types, U: unclassified lakes (predominantly from Norway).

As NMDS only takes into account the species composition, and superimposes the potential environmental characteristics afterwards, it is also apparent that the country wise sorting of the database can be partly due to species that have only been identified/registered as such in a single country. For example the Charophytes in Latvia are not keyed out to species level, while this is done in the other countries. Charophytes therefore contribute greatly to the separation of the Latvian data as a single group. The only solution to avoid such problems is a strong harmonization of data input. Unfortunately this would result in less in-depth information on the individual charophyte species.

2.1.2 Experiences and recommendations on the usability of the method for this dataset

The NMDS was used for screening a dataset which contained information from different sources. The main purpose of this exercise was to get acquainted with the full dataset and search for potential pitfalls in future analyses of this dataset. It was expected that data from different countries might be influenced by the differences in sampling methodology. The check with NMDS confirmed that data clusters were strongly separated according to country. Moreover, the predefined lake typology designed for European WFD assessments appeared not to be as clearly reflected as might be desired. The separation of data along an alkalinity gradient was present, but the various depth classes in the typology were not reflected as clearly as expected.

The results of this initial dataset assessment show that care needs to be taken when using this dataset for assessments of Europe-wide gradients of macrophyte responses. Harmonisation of macrophyte sampling techniques throughout Europe might improve this situation in the future. For the purpose of this report, we have continued to use these data to show that despite the obvious discrepancies between sub-sets of data, various patterns of responses seem to hold. This encourages further work on these responses using more homogeneous datasets in the future.

2.2 Detrended correspondence analyses (DCA)

Detrended Correspondence Analysis (DCA; Hill and Gauch 1980) is an improved ordination method for Correspondence analyses (CA), which counterbalances the mathematical artefact of distortion of the final resulting ordination pattern (an arch-shaped trend in the data plotting). In DCA this distortion is removed by detrending and rescaling of the ordination axes. The method is thoroughly explained in various handbooks (see also http://ordination.okstate.edu/DCA.htm for a very clear explanation).

DCA is a widely used ordination technique, and has been available for a longer time than NMDS. Like NMDS, DCA is a form of indirect gradient analysis. It also provides an overview of where samples are plotted in a multidimensional space and how clusters are formed along the axes of this space. In contrast to NMDS, it does not require that the user predefines the number of axis. There is no general clear answer to the question of which method is more useful for a given dataset, as both methods have advantages and disadvantages. These have been summarised by Palmer (see Table 1; source http://ordination.okstate.edu/).

Palmer states: "Note that the last two entries in [Table 1] do not indicate which method has the advantage. This is perhaps the biggest difference between the two methods: DCA is based on an underlying model of species distributions, the unimodal model, while NMDS is not. Thus, DCA is closer to a theory of community ecology. However, NMDS may be a method of choice if species composition is determined by factors other than position along a gradient: For example, the species present on islands may have more to do with vicariance biogeography and chance extinction events than with environmental preferences – and for such a system, NMDS would be a better a priori choice. As De'ath (1999) points out, there are two classes of ordination methods - 'species composition representation' (e.g. NMDS) and 'gradient analysis' (e.g. DCA). The choice between the methods should ultimately be governed by this philosophical distinction."

Table 1. Some of the major differences between Non-metric multidimensional scaling (NMDS) and de-trended correspondence analysis (DCA). (Source: http://ordination.okstate.edu/overview.htm)

	NMDS	**DCA**
Computation time	High	Low
Distance metric	Highly sensitive to choice of distance metric	Do not need to specify
Simultaneous ordering of species and samples	No	Yes
Arch effect	Rarely occurs	Artificially and inelegantly removed
Related to direct gradient analysis methods	No	Yes
Need to pre-specify numbers of dimensions prior to interpretation	Yes	No
Need to specify parameters for number of segments, etc.	No	Yes
Solution changes depending upon number of axes viewed	Yes	No
Handles samples with high noise levels	No(?)	Yes
Guaranteed to reach the global solution	No	Yes
Results in measures of beta diversity	No	Yes
Used in other disciplines (e.g. psychometrics)	Widely	No(?)
Axes interpretable as gradients	No	Yes
Derived from a model of species response to gradients	No	Yes

2.3 Canonical correspondence analyses (CCA)

In contrast to the indirect gradient analyses like NMDS and DCA, CCA (canonical correspondence analyses) is a direct gradient analysis, assuming that an underlying gradient is known (e.g. an environmental pressure gradient). CCA is often compared to RDA (redundancy analysis). In both methods, ordination axes are built based on linear combinations of the given environmental factors. While RDA is useful only for shorter environmental gradients with monotonous responses, CCA finds relationships both for monotonous, but also for unimodal relationships, and is therefore more applicable to data with long gradients such as response to eutrophication. CCA is available in many software packages, e.g. CANOCO or the R package "vegan".

2.3.1 Application of CCA to macrophyte data

CCA was used in the REBECCA macrophyte data analyses to indicate species responses along a gradient of totP. By grouping the species along this axis a distinction could be made between species response and change in tot. A full description of this method was presented by Willby et al. 2006. In the resulting image all species ranked on the CCA axis were grouped based on the assumption that species below the y-intercept (0) were positively responding to an increase in totP and species above the y-intercept were negatively responding to an increase in totP (Figure 10).

The CCA analysis that was done to classify macrophytes into groups is a transparent way of classifying species. However, the boundaries between the different classes remain to some extend rather arbitrarily chosen and have to be still drawn by hand. Although it helps, other methods like the suggested percentage approach (see Lyche-Solheim 2006) are also valid and more accessible to people that are not familiar with the interpretation of CCA and multivariate methods in general.

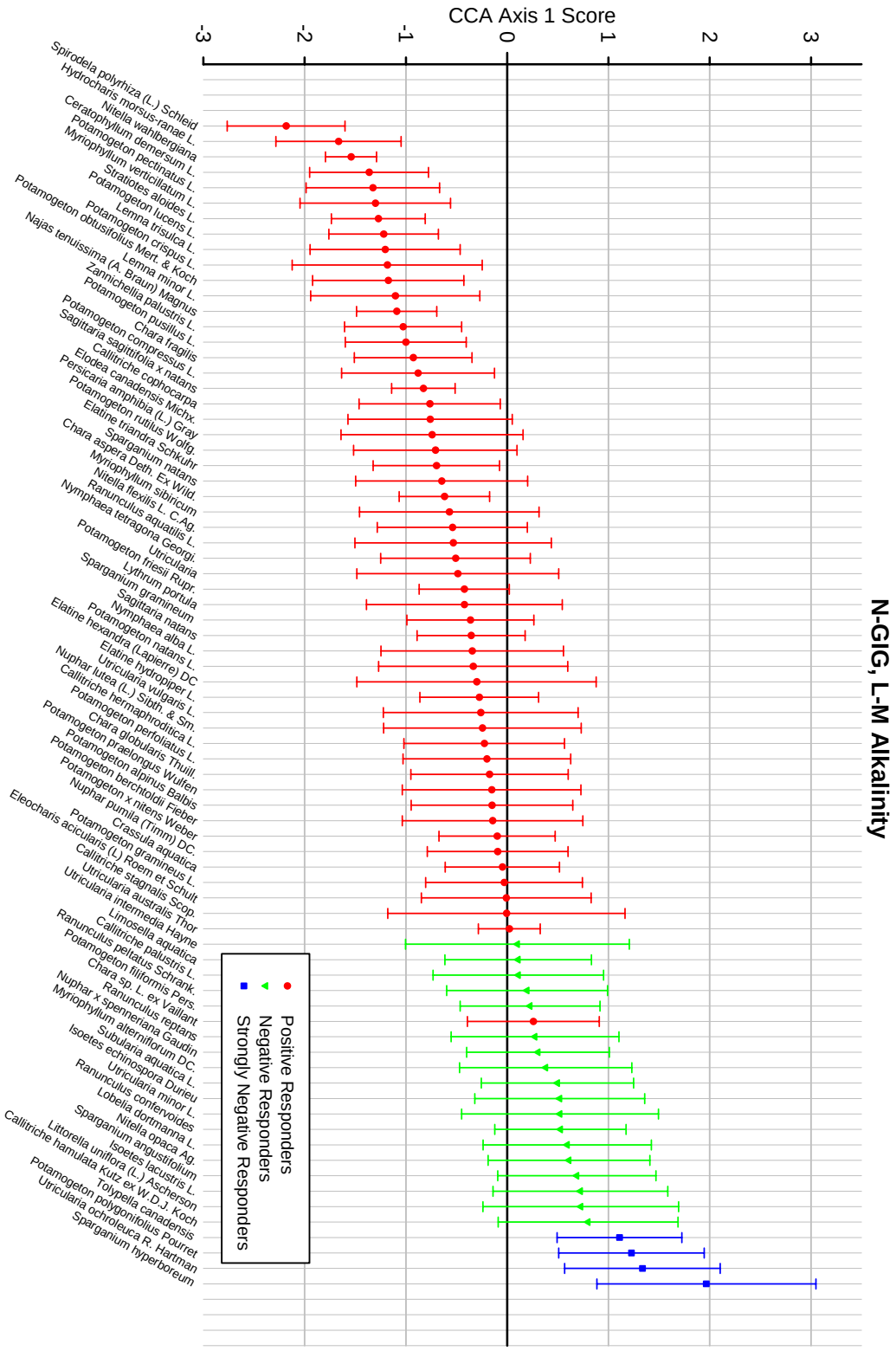


Figure 10. Resulting categories of species responding along a totP gradient.

2.3.2 Application of CCA to phytoplankton data

Once species optima are identified, they may be used for assigning site-scores to new sites. Such site scores may e.g. be used for estimating the ecological status of new sites. This approach is used in the CBAS tool for classification of new sites. Weighted averaging can be applied in an iterative way in order to obtain robust site scores. This approach was used for analysis of change points along pressure gradients.

Change point analysis based on cross validated site scores
An interesting aspect in gradient analysis is non-linearity. Communities do no necessarily change linearly along pressure gradients, but may exhibit sudden changes on short segments of the whole gradient. Above mentioned inferring of site scores from species optima represents a basic tool for the analysis of species composition along the pressure gradient. If species composition does not change gradually, but shows sudden shifts, then the calculated site scores, which are based on the species occurring at given sites, should also show sudden changes along the pressure gradient. In order to avoid circularity in the analysis, site scores and species scores should not be calculated from the same dataset. In this approach, a dataset has been split repeatedly in order to obtain robust site scores.

Resampling – The dataset was repeatedly (100x) split into two subsets, A and B. Species optima were derived by weighted averaging from subset A. The optima derived in each step were used for calculating trophic scores for each site in the corresponding subset B using again weighted averaging of all species at a given site with their square root-transformed abundances as weights. The total of the obtained trophic scores was plotted against the Chlorophyll-a concentrations (Figure 11).

Change point analysis – The relationship between chlorophyll-a concentration and the trophic scores was analysed in two ways. First a generalized additive model (GAM) was fitted to the data (Figure 11), using the *mgcv* package of *R* (see section 1.3). Visual inspection of this relationship gave a first indication whether a change-point exists in the dataset. Change-points were then analysed using the *segmented* package of *R* (Muggeo 2004). In *segmented*, broken linear regressions can be fitted to a two-dimensional dataset with one or more change points in the neighbourhood to one ore more provided initial value(s) (see Figure 11 for an example). Within each analysis, several runs were performed with varying initial values along the pressure gradient. Usually more than one change point were found. In these cases, the optimal change point was estimated by comparing the AIC values (Sakamoto et al. 1986; see Section 1.5.1) of the different segmented regressions, and choosing the model with the lowest AIC value. Segmented fits were always compared to simple linear fits by their AIC values. In the example shown in Figure 11, the data show a clear nonlinear response that is well described by a segmented regression.

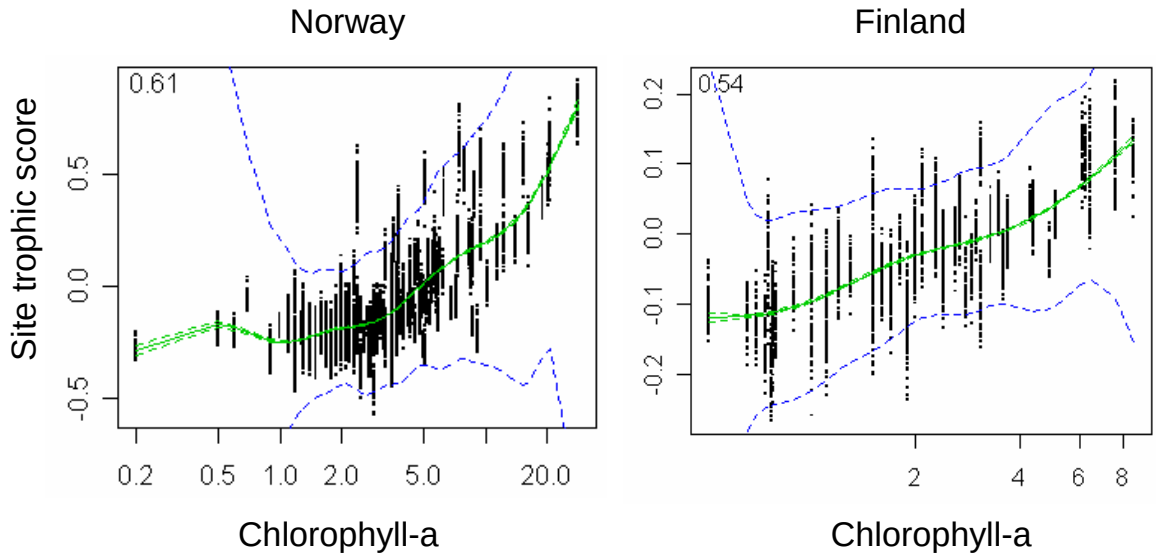


Figure 11. Change-point analysis based on site trophic scores. Site scores were calculated repeatedly for a number of Norwegian (left) and Finnish (right) lakes. Each site is represented by 100 repeatedly calculated site scores. For the Norwegian data, a non-linear regression gave a better fit than a linear (cf. Figure 12), while the linear model had to be preferred for the Finnish data (selection based on AIC criterion). This is likely due to the much shorter gradient in the Finnish dataset. The fit of a generalised additive model (GAM) together with its confidence interval (95%) are given by green and blue lines, respectively. The r^2 of the GAM is given in the upper left corner.

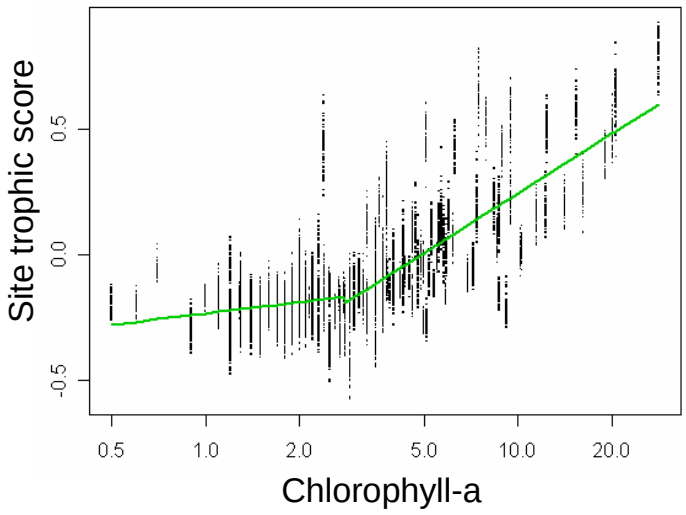


Figure 12. Fit of a segmented regression over the data shown in Figure 11, left panel (Norwegian data).

The CBAS classification tool

The CCA-Based Assessment System (CBAS; Dodkins et al. 2005, Dodkins and Ripley 2006) was adopted for phytoplankton by Carvalho et al. (2006) in order to match the needs of having too few observations for performing lake-type-specific analysis. In the CBAS tool, first species optima and tolerances are calculated with respect to the main pressure gradients from all available data (TotP and Chlorophyll-a). Using weighted averaging, these species optima are used for inferring site scores for the available reference sites. The obtained site scores of the reference sites are then used in a multivariate regression for extracting type-specific reference scores.

For any given monitoring site, a site score is calculated from its phytoplankton community (see Figure 13 for a regression between predicted site score and actual pressure values). Next, an 'expected reference score' is calculated for this site based on above mentioned regression with reference sites. An impact metric is then calculated by subtracting the reference metric from the site's actual score.

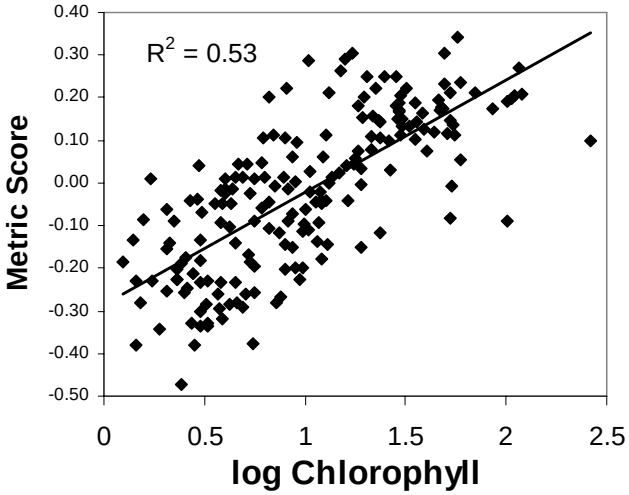


Figure 13. Scatter plot showing relationship between log Chlorophyll and phytoplankton chlorophyll metric weighted by both abundance and Indicator Value (Carvalho et al. 2006).

CCA-derived species optima appear to be useful for analysis of responses to pressure gradients. A remaining drawback (not limited to CCA) is that species optima calculated from a given dataset are valid only for sites within this particular dataset. This is because any dataset represents only a subset of the total pool of lakes. Moreover, as CCA is based on weighted averaging, species optima will always lie within the boundaries of the pressure gradient, though we may expect that the true optimum of many species may lie outside the currently observed range of sites.

3. Bayesian methods

Jannicke Moe

Probabilistic methods are being used increasingly within ecological modelling and analysis. Bayesian models are based on and predict probability distributions, and incorporate uncertainties in a more explicit way the than the more common classical or "frequentist" statistics do. There are several features of Bayesian methods that make them especially suitable for characterising pressure-response relationships for data that are compiled from various sources, as in REBECCA WP3.

1) Information from different sources and of different nature can be combined into a single model (e.g. monitoring data, experimental data, expert judgement, best guesses).
2) A hierarchical Bayesian model can be constructed more easily than an equivalent hierarchical frequentist model. This means that information of hierarchical nature, such as from different spatial scales, can be used more efficiently in a Bayesian model.
3) The model can predict a probability distribution, which provides a more direct interpretation of risk than traditional model predictions (e.g. point estimate with standard error)
4) A Bayesian model may be formulated as a causal network, which is constructed graphically as a flow diagram. A Bayesian network can thus be developed as a conceptual model, and at the same time be analysed as a mathematical model. This can facilitate cooperation on model development, analysis and interpretation with stakeholders and other involved parts.
5) A Bayesian network may consist of several sub-networks, which can be developed and analysed independently of the full network.

The comprehensive accounting of variability and uncertainty is central, and the distinction will provide guidance as to what is "predictable", what is inherently unpredictable, and where additional data can provide the most benefit (Clark 2005). For example, effects of temperature increase cannot be directly predicted because of the many uncertain causal links. However, we may explore different scenarios where uncertainties in climatic variables are incorporated directly into simulation models as probability distributions, and are accounted for in the uncertainty of the predictions. Bayesian methods may therefore seem like a natural choice for risk assessment and decision analysis. Still, the use of Bayesian methods is still growing rather slowly within ecology, due to the strong tradition of so-called frequentist statistics. Bayesian statistics used to be computationally more demanding, but with modern computers and software this is no longer a problem.

In Work package Lakes, two different Bayesian approaches have been applied for two particular purposes:
(1) Estimation of target concentration of TotP for obtaining good ecological status for lakes, according to proposed boundary levels for chlorophyll a. A hierarchical Bayesian regression model was used for this purpose.
(2) Classification of lake status according to the WFD, using information from four different biological elements and the one-out-all-out principle. A Bayesian Network approach was used for this purpose.

3.1 Hierarchical Bayesian regression model (HRM): application to target load estimation for total phosphorous and nitrogen

Olli Malve

Hierarchical or multilevel models (HM, Gelman et al. 2005) can be used widely in statistics of environmental sciences and management, in which data is often gathered in a nested or hierarchical fashion, e.g. lakes within lake types, or lake types within eco-regions, or eco-regions within continents (Malve and Qian 2006, Malve 2007). Nested data arise also in meta-analysis in environmental sciences where the goal is to combine information from a number of studies of essentially the same phenomena, and to produce more accurate inferences and predictions than those available from any single study. For example, if individual lakes are sampled cross-sectionally but studied longitudinally, a hierarchical data structure arises as well. Often environmental data are modelled at a high level of aggregation – for instance, assuming that all observations are sampled homogenously from a single population such as a common lake type. However, heterogeneity is often the rule rather than the exception, and the available predictor variables often do not explain this heterogeneity sufficiently. The heterogeneity can be described using mixture models that employ latent variables in a hierarchical structure. Hierarchical models also provide a natural way to treat issues of model selection and model uncertainty. Several models can be worked with simultaneously and be weighted in proportion to their plausibility given the data.

3.1.1 Management question

In the implementation of EU Water framework directive (WFD), the specification of target nutrient concentrations for lake management can be based on the predictions from a regression model of chlorophyll a versus the nutrient. However, regression analyses are often based on insufficient data from monitoring networks, which results in nutrient and chlorophyll a (Chl-a) observations sampled in a nested fashion. Most of the lakes are sampled a few times only, and observational nutrient concentration ranges do not always cover targeted water quality criteria. In these cases, the lake-specific regression model is not useful in water quality prediction, and lake-type-specific models may be inaccurate and biased when used to predict concentrations in a given lake.

3.1.2 Statistical method

A hierarchical regression model (HRM) that is fitted to hierarchical lake water quality data can "extrapolate" outside lake-specific observational range, and often results in reduced model uncertainty and improved accuracy in estimated model parameters. This is useful e.g. if the sampled lake is within a eutrophic region, while the management target is somewhere in a mesotrophic region.

A hierarchical model for this purpose can be constructed as follows. Individual observations of Chl-a concentration are modelled conditionally on lake-specific model parameter values; the lake-specific model parameter values are modelled conditionally on lake-type-specific parameters; and the lake-type-specific parameters are in turn modelled conditionally on a parameter distribution of all lakes in a given region. The model can be fitted to observations of Chl-a, TP and TN data in lakes. To enable hierarchical model fitting, the lakes and observations have to be classified into geomorphological lake types. Inclusion of the main effects and the interaction of nutrients in the model is normally justified.

The full statistical distribution of parameters and predictions can be estimated using Markov-chain Monte-Carlo sampling methods and Bayesian Inference (Gelman et al. 2005). Freely available OpenBug software (http://mathstat.helsinki.fi/openbugs/) is useful in estimation.

3.1.3 Setting up the nutrient targets for Finnish lakes using HRM

National water quality monitoring of Finnish lakes started in 1965 after the passage of the Water Act in 1962. Sampling strategy and analysis methods have been described by Niemi (2001). Information was required on the status of Finnish water resources, quality and quantity and how the status relates and responds to pressures on the environment. The Geomorphological typology of Finnish Lakes was constructed to aid in the classification of ecological status (Table 2). Lakes are divided into the nine lake types according to their surface area, depth and water colour.

Table 2. Geomorphological typology of Finnish Lakes specified by Finnish Environment Institute. SA=Surface Area, D=Depth. Unit for colour is mg/l Pt.

Lake Type	Lake Type Name	Characteristics
1	Large, non-humic lakes	SA > 4,000 Ha, colour < 30
2	Large, humic lakes	SA > 4,000 Ha, colour > 30
3	Medium and small, non-humic lakes	SA: 50 - 4,000 Ha, colour < 30
4	Medium Area, humic deep lakes	SA: 500 - 4,000 Ha, colour: 30-90, D > 3 m
5	Small, humic, deep lakes	SA: 50 - 500 Ha, colour: 30-90, D > 3 m
6	Deep, very humic lakes	Colour > 90, D > 3 m
7	Shallow, non-humic lakes	Colour < 30, D < 3 m
8	Shallow, humic lakes	Colour: 30-90, D < 3 m
9	Shallow, very humic lakes	Colour > 90, D < 3 m

19,248 July and August observations of TP, TN, Chl-a from 2,289 Finnish lakes from 1988 to 2004 are used in this study. About 42% of the observations are from July and 58% from August. However, observations are unevenly distributed among years, types and lakes (Table 3 and Table 4). Of the total of 2,289 lakes, 900 lakes have only one observation. The average number of observations is eight (s.d. 26) per lake. One lake has 441 observations and there are 12 lakes that have more than 150 observations.

Table 3. The number of observations (N) per year from 1988 to 2004.

Year	N	Year	N	Year	N	Year	N
1988	2	1993	426	1998	1,610	2003	2,220
1989	59	1994	1,478	1999	1,533	2004	774
1990	66	1995	1,621	2000	2,029		
1991	78	1996	1,687	2001	1,972		
1992	71	1997	1,714	2002	2,088		

Table 4. Number of observations (N) within the lake types.

Type	N	Type	N	Type	N
1	485	4	3,949	7	391
2	6,536	5	1,080	8	2,729
3	388	6	1,326	9	2,544

Posterior chlorophyll a response surfaces as a function of nutrients are simulated with a hierarchical regression model. Response surfaces reveal the effects of the nutrients and demonstrate their usage in lake eutrophication management. We therefore simulated posterior probabilities of chlorophyll a exceeding the criteria. The TN and TP plane of interest was divided into 100*100 grid cells, and the predictive chlorophyll a concentration distribution in each grid cell was calculated. The results are presented as the contour lines of the 80th percentiles of these predictive distributions (Figure 14). The 80th percentile was selected number to reflect the potential risk attitude of a lake manager. From these contour lines, we can identify nutrient concentrations that result in an

80th percentile of chlorophyll a distribution at 30 µg L^{-1}, which is the nutrient condition to ensure a <20% probability of chlorophyll a concentration exceeding the criteria.

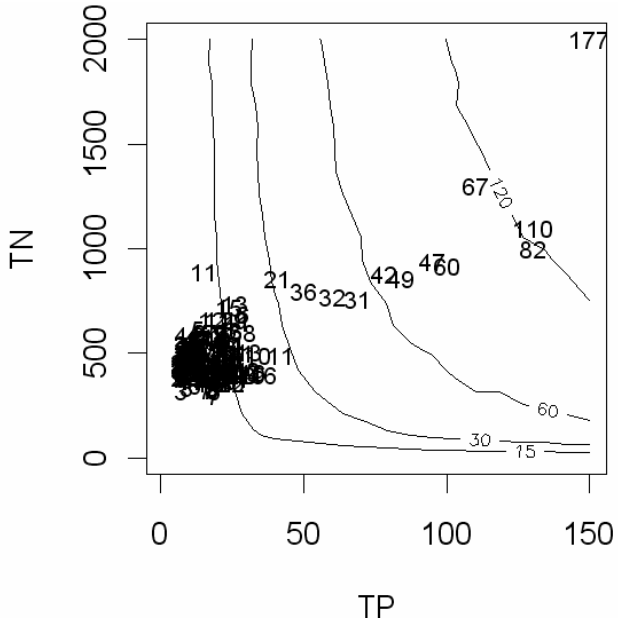


Figure 14. 80 % percentile contour lines of predictive Chl-a concentration in Lake Päijänne (large humic lake, type 2) at 15, 30, 60, 120 [µg L^{-1}] as a function of observed TP and TN concentrations [µg L^{-1}]. Predictions are simulated with the hierarchical linear model. Numbers are observed Chl-a concentrations [µg L^{-1}].

Chlorophyll a concentration as a function of TP and TN concentration [µg L^{-1}] for the Lake Päijänne (large humic lake, type 2) is also simulated with the hierarchical linear model (Figure 16). First, we vary TP concentration within the observational range while TN was held constant (at 50% percentile). Then, we repeat the simulation holding TP constant at the same percentile while varying TN within the observational range.

The hierarchical chlorophyll a model is compared both to a non-hierarchical type-specific dummy variable model and to a linear lake-specific model. Model fits for four selected lakes are computed to illustrate the differences between the models (Figure 15). The lakes are selected to show the effect of the sample size on the model's fit and on the predictive confidence region. The selected lakes are Lake Onkilampi (shallow humic lake, type 8), Lake Nurmijärvi (large non-humic lake, type 1), Lake Kuhajärvi (shallow non-humic lake, type 7) and Lake Päijänne (large humic lake, type 2). The numbers of observations are 3, 7, 22 and 265 respectively. In general, the comparison is overwhelmingly in favour of the hierarchical model compared to the non-hierarchical type-specific model. Median Chl-a concentrations predicted by the hierarchical model were usually closer to the observed Chl-a values than means predicted by non-hierarchical dummy variable model (Figure 15), suggesting that the hierarchical model fits the data far better. This is indicated by the R^2 which is greater for hierarchical model. Also, the deviance and the deviance information criterion (DIC) of the hierarchical model are smaller than those of the non-hierarchical dummy variable model (Table 5). The lower DIC of the HM indicates that the increased number of model parameters of the hierarchical model is more than compensated by the improved model fit.

Table 5. The deviance information criterion (DIC) and deviance (D) for the hierarchical model (HM) and for the non-hierarchical type-specific dummy variable (DM) model.

Model	**DIC**	**D**
HM	25,946	28,358
DM	31,474	31,515

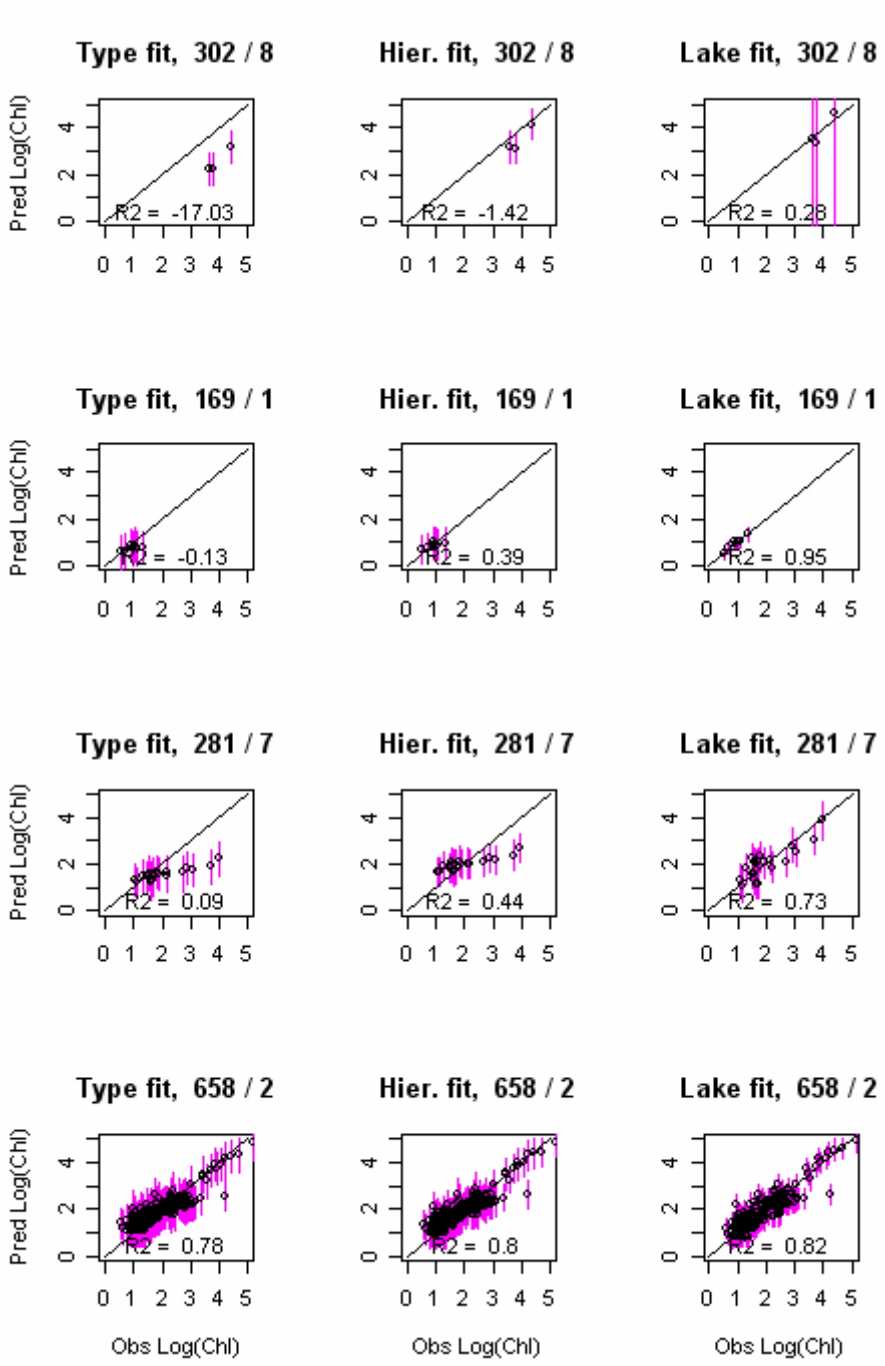


Figure 15. 10 %, 50% and 90% percentiles of predicted Chl-a concentration [µg L^{-1}] as a function of observed value for four selected lakes. 50% percentiles are represented by circles; 10% and 90% percentiles are connected by red vertical lines bars. Upper panel: Lake Onkilampi (shallow humic lake, type 8); Second upper panel: Lake Nurmijärvi (large non-humic lake, type 1); Second lower panel: Lake Kuhajärvi (shallow non-humic lake, type 7); Lower panel: Lake Päijänne (large humic lake, type 2). The line in 45° angle represents the 1-1 line (perfect fit). Percentiles have been calculated with the lake type specific non-hierarchical model ("type", left panel), with the hierarchical linear model ("hier", middle panel) and with the lake-specific non-hierarchical model ("lake", right panel).

When using the non-hierarchical lake-type-specific dummy variable model, we treat all lakes within a type as the same and pool individual observations to form a type-specific model. This model represents a weighted average, with the weights proportional to each lake's sample size. That is, the lake-type-specific model is heavily weighted by lakes with larger sample sizes. Consequently, the resulting model can be grossly biased for lakes with small sample sizes. This feature is clearly illustrated in the four selected lakes (Figure 15). The hierarchical model treats lakes within the same type as exchangeable and fits lake-specific model parameters. But these parameters are assumed to come from the same prior distributions; thereby information from similar lakes can be pooled. This pooling of information reduces the bias at the lake level, and reduces model error variance as well.

Lake-specific non-hierarchical linear models are fitted using only data from a specific lake. Despite of the better fit of the non-hierarchical lake specific model compared to its counterparts, the model error variance tended to be large when sample size is small but decreases heavily as sample size increases (Figure 15).

The lake-specific 80% percentile contour lines for Lake Päijänne (large humic lake, type 2) simulated with the hierarchical model (Figure 14) reveal the usefulness of the posterior simulations in water quality management. Simulations have been confined within the part of the observational ranges of TP and TN in large humic lakes (type 2, TP: 2-160, TN: 31-4400) that is below the lake-specific maximum values (TP: 150, TN: 2000). The simulation in Figure 15 included TN and TP values outside the lake-specific observational ranges (TP: 6-150, TN: 300-2000). However, the extrapolation under the hierarchical setting is reasonable due to the pooling of information within and among lake types. This is a distinct advantage compared to the non-hierarchical lake model, which can predict only within a lake-specific observational range. For lakes with few observations, this range can be limited. The contour lines for Lake Päijänne are parallel to the y-axes in the observational range, showing clear TP limitation of Chl-a within this range. In contrast, near the low TN boundary and in the high TP range, TN limitation seems to prevail. From figures similar to Figure 14, a lake manager can read nutrient concentrations that comply with the given Chl-a standards with a given certainty.

The effects of TP and TN are illustrated also in the predictive plots (Figure 16). Simulated Chl-a increases with TP, but not very much with TN. The 10%-90% percentile predictive intervals look rather wide at first glance. The predictive interval is the predicted credible interval for individual observations, which is always wider than the commonly presented fitted confidence interval for the mean. The predictive distribution is directly related to the process of lake eutrophication assessment, while the fitted mean is not.

The collinearity of TP and TN makes it difficult to determine the effects of each on Chl-a from the estimated slopes alone. Therefore the posterior simulations for the Lake Päijänne (large humic lake, type 2) were calculated. Simulations show very clear TP limitation within the observational range. This indicates accurate separation of the effects despite the high correlation (0.7) between the coefficients $beta_1$ and $beta_2$. The collinearity is not transferred to the predictions. The Monte Carlo Markov-Chain (MCMC) simulation of the linear coefficient parameters $beta_0$, $beta_1$, $beta_2$ and $beta_3$ together with their correlation, enable the separation of the effects by including correlation of the parameters in the predictive simulation of Chl-a.

Despite the collinearity of nutrients, the main effects of their standardized normal deviates ($\frac{\log(TP) - \mu_{\log(TP)}}{\sigma^2_{\log(TP)}}$) on Chl-a were estimated for the nine lake types with the hierarchical model (Table 6). On average the main effects agree with results (not shown) from a CART model (Classification and Regression Tree). The effect of TP is twofold compared to the effect of TN. However

in small humic, deep very humic and shallow humic/very humic lakes (types 5, 6, 8 and 9), TN effects almost equalled TP effects. Here the main effects are inconsistent with the CART model. The main effects represent an average trend within the observational range, whereas the CART model splits the range into discrete domains. Therefore the interpretation of the main effects is different.

Table 6. Main effects of the standardized normal deviates of nutrients for the nine lake types estimated with the hierarchical regression model.

Type	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Mean**
TP	1.4	1.2	1.3	1.2	0.9	0.8	1.4	0.8	0.6	1.1
TN	0.5	0.3	0.6	0.4	0.5	0.5	0.4	0.6	0.4	0.5

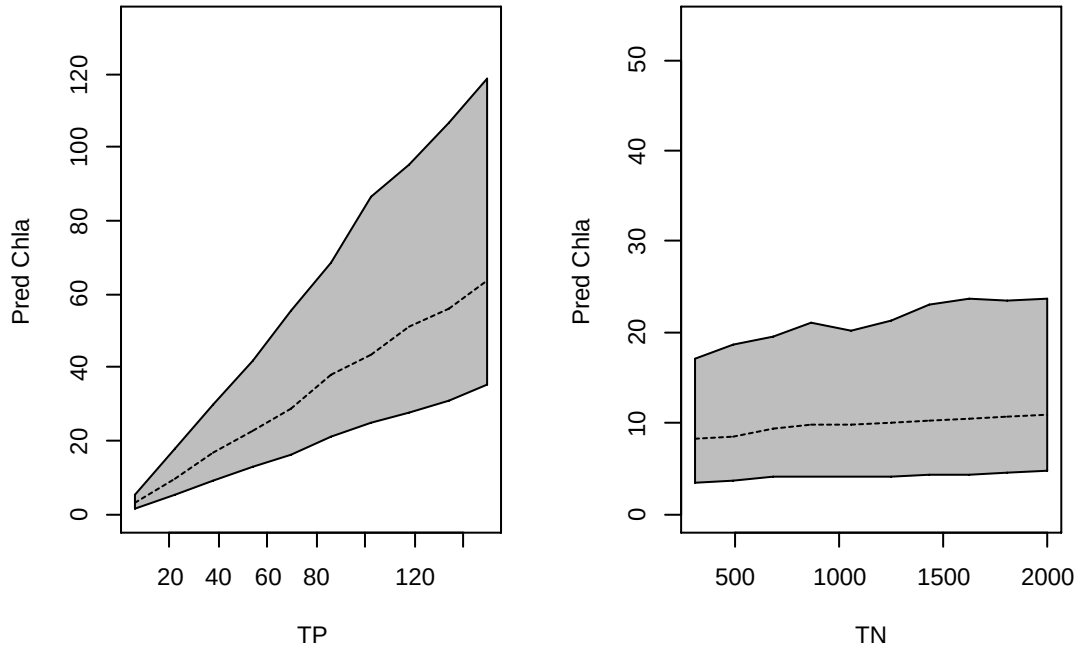


Figure 16. Chl-a concentration [µg L^{-1}] as a function of TP and TN concentration [µg L^{-1}] for Lake Päijänne (large humic lake, type 2) predicted with the hierarchical linear model. (50 % percentile - dotted line and 10 % - 90 % percentile confidence region - solid lines). While TP is varied within the observational range TN is kept constant (50% percentile) and vice versa.

3.1.4 Experience and recommendations for use

To aid in water quality management of Finnish lakes the relationship between chlorophyll a and the nutrients was investigated. The Bayesian hierarchical linear model with three levels (lake, lake type, and all lake types) was fitted to a cross-sectional data set of 19,248 July and August observations from 2,289 lakes in the Finnish lake monitoring network. The lakes in the network have been classified into nine lake types for the determination of ecological status.

The hierarchical approach has advantages both over lake-specific linear models (which are unbiased but with potentially high model uncertainty) and over lake-type-specific models (which have smaller model uncertainty but are potentially biased). The hierarchical model is favoured by model evaluation criteria such as R^2, deviance information criterion (DIC) and deviance statistics. The hierarchical model pools information from different levels of the lake hierarchy (lake, lake type, and all lake types), which reduces both uncertainty and potential bias in lake-specific predictions.

The Bayesian hierarchical model was used to predict lake chlorophyll a concentrations and to simulate a response surface for probability of chlorophyll a standard exceedance. This probability response surface is ideally suited for setting nutrient criteria, which can be directly used within a risk assessment framework. Posterior predictive simulations of a Bayesian approach are more informative than the traditional regression modelling approach, because of the use of Bayesian credibility interval directly conveys the probabilistic risk of a management decision.

Using the hierarchical modelling approach allows the model to borrow strength from lakes within a lake type, which is important for the lakes with few observations and limited observational range. When lake-specific data are unavailable, a lake-type-specific model can be used to generate a prior for an initial management strategy.

Using lake-specific model predictive distributions, we can provide information on monitoring network optimization to minimize uncertainty for all lakes. The hierarchical model presented here is proposed for lake eutrophication management in Finland to comply with the European Union Water Framework Directive. The Finnish lake monitoring network has produced a large cross-sectional water quality data set from lakes that have been classified by type. In the future, monitoring efforts will be concentrated on the lakes that are likely to violate current eutrophication-related water quality standards and lakes with high predictive uncertainty. The lake-specific target nutrient loads will be estimated with a hierarchical nutrient retention model together with the hierarchical chlorophyll a model. Fish and zooplankton effects will be added to the model after the collection of necessary data, to predict the effect of biomanipulation.

Target nutrient load estimated using our hierarchical model will benefit lake management work for lakes with few observations or lakes newly added to the monitoring network. Target load estimates for all lakes will be adjusted as additional monitoring data are collected annually, such that an adaptive management scheme can be developed to implement the European Water Framework Directive (Saloranta 2003).

3.2 Bayesian networks: application to classification of lake status

Sakari Kuikka

3.2.1 Why a Bayesian network approach was selected

The European Water Framework Directive (WFD 2000/60/EC) aims at enhanced protection and improvement of the aquatic environment... In order to design necessary and WFD-compliant mitigation actions, 'programme of measures', an appropriate ecological classification system with consistent boundaries separating classes from each other is needed. This has been the task for the Intercalibration Exercise aimed to be finished by the end of 2007. As a next step, each waterbody has to be classified by using biological observations and appropriate classification techniques.

The information that we usually obtain by sampling our water body includes several uncertainties, caused by sampling and measurement errors. Thus, when classifying a waterbody, we always have a risk of misclassification. The management principles stated in the WFD includes the precautionary principle, which refers to the concept of uncertainty and calls for a proper uncertainty management within the implementation of the WFD. A precautionary approach can be understood to be risk-averse decision making, i.e. to take more precautionary actions if the risks are high. This uncertainty should be taken into account in many steps of the implementation process, including the design of classification systems and classification of single water bodies.

Annex V of the WFD (section 1.4.2) states that the ecological status classification of a water body shall be represented by the lower of the values for the relevant biological and physico-chemical ele-

ments. The principle called the 'one-out all-out' principle was introduced in the Classification Guidance (European Commission 2003). This principle means that the quality class of a waterbody is defined according to the biological quality element (phytoplankton, macrophytes, benthic invertebrates or fish) that gives the worst classification result, regardless of the classification results based on the other biological quality elements. This procedure is theoretically in accordance with the precautionary principle by setting a high emphasis on the most sensitive element responding to a human pressure or to a combination of several pressures. However, it should also be considered that misclassification occasionally may result in a worst classification that is lower than it should. Such misclassifications may be due to high uncertainties.

As pointed out in the Classification Guidance, the risk of misclassification may be amplified by the number of quality elements that are taken into account in the one-out all-out system. The higher the number of quality elements (or the number of parameters indicating the condition of these elements), the more likely it is that one of them happens to be assessed as bad even if the water body would actually be in good state. The uncertainties within biological assessments are generally high, and often these uncertainties may differ for the different biological quality elements. Thus, the risk for misclassification is likely higher when applying the one-out-all-out-principle including all four biological elements compared to using a different classification approach based on some kind of averaging or integration of the results from the different biological elements. The WFD Annex V also points out that elements with very high uncertainty should not be used in assessment.

Bayesian statistics is an effective way to estimate uncertainties in statistical inference. In addition to providing probability distributions for the variables of interest, it also offers a possibility for answering complicated questions by probability models. For example, we may be interested in estimating what is the probability of a lake belonging to classes ‘High’ or ‘Good’.

We used a Bayesian approach for two purposes:

1) To evaluate the “one-out all-out" principle under uncertainties in sampling and in assessments based on classification models
2) To study, as an alternative, the integration of information contents of the four elements with a Bayesian model that takes specifically into account the uncertainties of each element

We focused on the risk of making wrong conclusions (misclassification) by combining different elements and their likelihoods. In particular, we focused on the boundary between the quality classes “Good" and “Moderate", because this boundary is the criterion for the decision of starting appropriate management actions to improve the ecological status of a lake. Even though the WFD includes several stressing factors, we focus on eutrophication and on the role of phosphorus in this process, since this the most important pressure to lakes we have included in our study.

3.2.2 Material and Methods

Taking a very deterministic approach, it can be stated that the (ecological) state of a lake is defined by the physical, chemical and biological properties and processes (including human impact) within a lake and its drainage basin. Whilst the Intercalibration Exercise will set the boundaries between different ecological quality classes, we still face another problem: the sampling and measurement errors as well as the lack of good and precise indicators, partly due to our incomplete understanding of the ecosystem functioning, frequently lead us to a position where biological information gives us controversial information of the state of the lake. For example, phytoplankton data may suggest that the lake is in 'good' status, while the benthos data indicate 'moderate' status.

We decided to apply so called Naïve Bayesian nets (Figure 17). This means a Bayesian net where there is one parent node (independent variable) and several child nodes (dependent variables). The parent node has two states (good or moderate), which is represented by TP in two discrete intervals.

The child nodes represent the measured variables of the elements, like biomass, number of species, etc., in discrete intervals. Each element will contribute to the classification through their specific likelihood probabilities (the probability of value of the element given the true state of the class). In a way, likelihoods describe how “trustworthy" the different measurements are, i.e. how well they describe changes in the state of the lake.

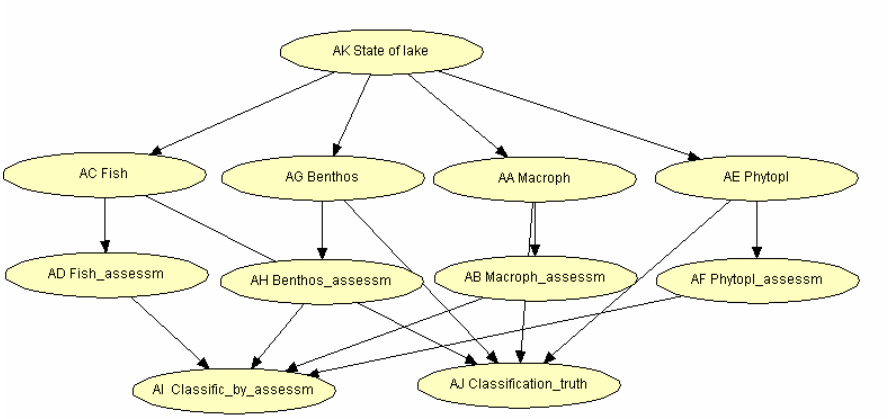


Figure 17. Naïve Bayesian net: a Bayesian net where there is one parent node (independent variable: lake status) and several child nodes (dependent variables: the measured variables). In this case, lake status has two discrete classes (Good or better + Moderate or worse), represented by TP in two discrete classes. The child nodes represent the measured variables of element, like biomass, number of species etc.

Naïve Bayesian nets have been found to be among the best classifiers in several studies (Langley et al. 1992, Walley & Dzeroski 1995). Even though there are not yet comparisons of e.g. multivariable methods and Naïve Bayesian nets, we are convinced that the effectiveness of this method is very high. These numbers used in this Naïve Bayesian model are then applied in a meta-model (Section 3.2.3), where the various elements can be combined and which mimics the principles of the WFD (including one-out all-out principle).

This meta-analysis model (metamodel in the sense that it utilizes the values of other models, i.e. element specific likelihood estimates) is described in Figure 18. It follows the notation of Bayesian nets in general. While it is a Naïve Bayesian net, there is only one parent node and several child nodes. While the model aims to be able to summarise the probabilities for studies of each element, it does not include the detailed models applied for the estimation of each element.

The probabilities in Bayesian models can be estimated by using some theoretical statistical distribution (see for example Gellman 1995) or, alternatively, by using directly the frequencies of data. Here we used the frequencies of dicretised data, which frees us from the assumption of statistical distribution. This is possible here because we have a relatively high number of observations in the discrete intervals.

The model was tested and the likelihoods estimated by the so-called “leave one out method". The resulting likelihoods can be called empirical likelihoods. The method mimics the uncertainty that we have when we sample a new lake for which we have data from elements but no classification result. In this error rate estimation it is assumed that parameter values of the assessment model have been estimated from other lakes. It leaves out the “correct value" of a lake, and estimates the model parameters by data sets of other lakes and uses the child nodes (dependent variables) to estimate the value of the parent node (independent variable, here the class of the lake). This method estimates the frequency of correct and incorrect classification, and it provides the probabilities to the likelihood function, i.e. the probabilities applied in the meta model.

3.2.3 Meta-model

The above estimated probabilities were then used as input values to a meta-model (Figure 18). They represent conditional probabilities between the 'true' state of each element and the assessment result, i.e. the probability that a certain assessment result predicts the true state of the element. This is an empirical likelihood function.

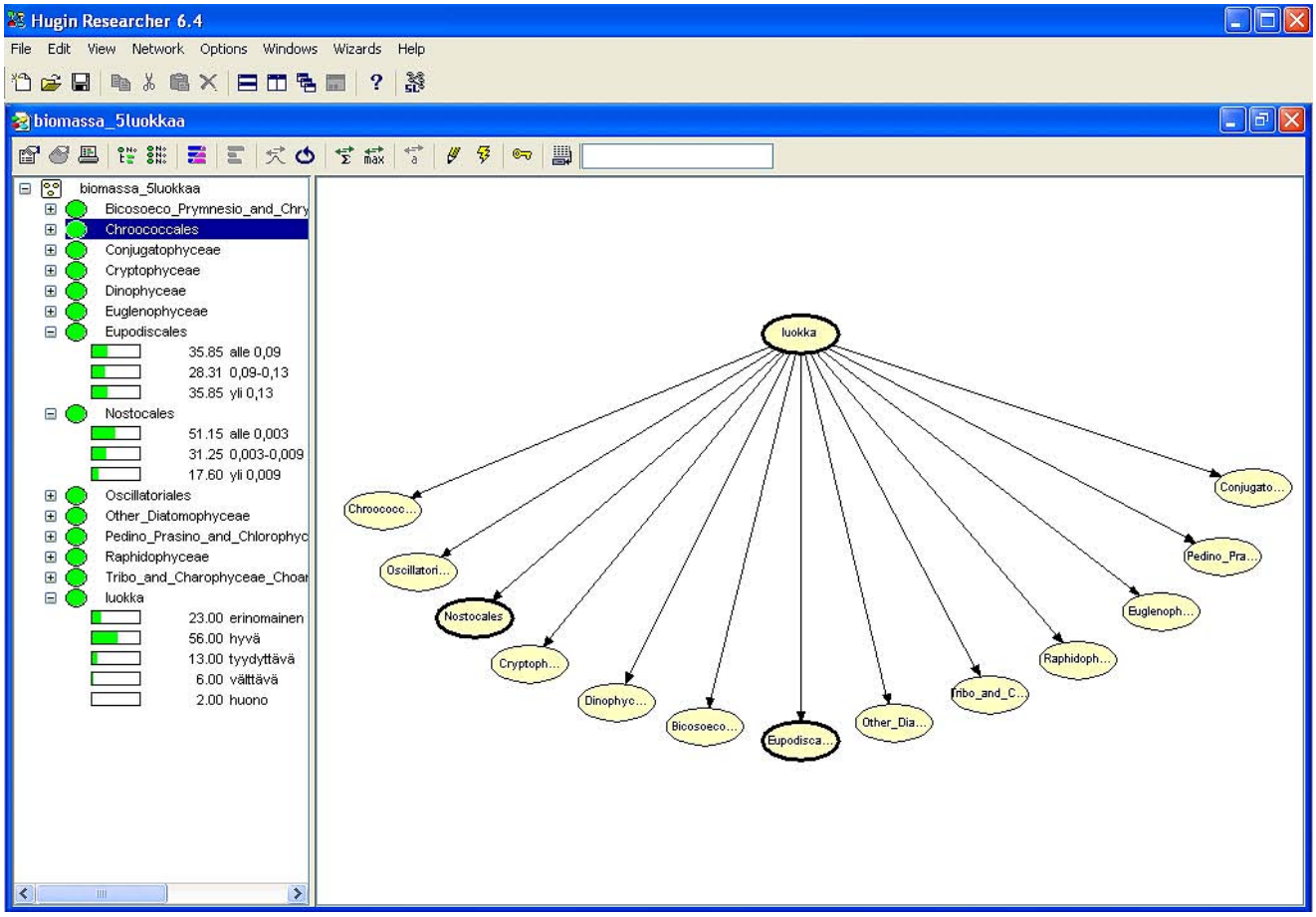


Figure 18. Bayesian meta-model. Each node (phytoplankton group) represents conditional probabilities between the 'true' state of each element and the assessment result, i.e. the probability that a certain assessment result predicts the true state of the element.

The uppermost variable of the meta-model (Figure 18) is ‘State of lake’, which has two probabilistic outcomes: "No action" (i.e. good status or better) and "Action" (i.e. moderate status or worse). These outcomes are based on the WFD-compliant ecological quality classes as in Section 3.2.1. They conditionalise the next layer of the model, i.e. the true state of the four biological quality elements. The conditional probability distribution of this layer as follows: if the state of the lake is ‘Action’, then the probability of the true state of a biological quality element is 1.0 for the outcome "moderate status or worse" (i.e. "Action"). This means that this version of the model does assume that a certain state of lake will be deterministically reflected to each element. In other words, we assume that we have complete information of natural and human factors determining the abundance and structure of a biological element. However, this is likely not true. It is more likely that a certain state of a biological element can be linked to unknown causal relationships and noise. However, we did not take this uncertainty into account in this exercise.

3.2.4 Results and discussion

According to our Naïve Bayesian Nets (results not shown here), fish and phytoplankton were most reliable biological elements, while benthos is the most unreliable indicator for lake status. These differences can be partly explained by differences in data quality: our benthos data represented only 22 lakes, whilst we had phytoplankton observation from almost 400 lakes. The other reason for higher uncertainty connected to benthos is that it is more indirectly impacted by eutrophication, and are thus less sensitive than the directly impacted phytoplankton. These differences in Naïve model results will be reflected in the behaviour of the meta-model.

Our results (not shown here) indicate clearly that classification, which includes the 'one out – all out' logic, has a low probability (30% in our case) for 'No action', because it calculates the probability that at least one of the biological quality element assessments shows the outcome 'Action'. This result fits well to the precautionary approach: the bias is rather to the direction of "too much management" than "too little management".

Furthermore, if we assume that the 'true' state of the lake is either 'Action' or 'No action', the model can be used to illustrate what are the likelihoods of the two optional outcomes ('Action'/'No action') of assessment results (taken into account the asymmetric likelihood probabilities for different errors in different biological elements). The results clearly indicate that it is very likely (probability 95%) that the 'one out – all out' principle gives a general assessment result of 'Action', even though the probabilities in the assessment results of single quality elements are potentially more erroneous due to the uncertainties in single assessments. It is very likely that at least one of them shows 'Action'.

However, this precautionary attitude has its price if we assume a 'truly' good lake. Because there are quite high probabilities to get wrong assessment results in the elements, the wrong conclusion ('Action') produced by 'one out – all out' principle is as high as 0.4. This means that 40 % of the lakes that would not need any abatement actions would be classified to need some, because at least one element is considered to be in moderate or worse state.

The Bayesian meta-model described above was also used as a tool to assist in classification of a lake, assuming that we have new information of one or several biological elements. In doing this we used the model in a Bayesian way, i.e. we calculated the *a posteriori* probabilities of the different variables using this new information. The results (not shown here) show that this new information on one biological element changes our understanding about the likely outcomes of the other biological elements. This can be explained by the dependencies in the model: new information on one element updates the first the 'a posteriori' probability of the 'true' state of this element, this in turn affects the probability of the general state of the lake (as described above). Finally, this updates the probabilities of the 'true' and assessed states of other biological elements. The lower the uncertainties between the true and assessed relationships (based on the Naïve model results), the larger improvement is obtained in the assessed classification result of the other biological elements. The strength of the Bayesian model becomes even clearer, where new monitored information is available from more than one biological element. The fact that two information sources, even though both uncertain, support the same conclusion, decreases the uncertainties in classification effectively.

The use of Bayesian models to combine the information contents of the uncertain assessment results is a promising way to carry out the overall classification of water bodies. It will enable such risk management practice, where the decision makers have a possibility to judge what the uncertainties matter in a specific case. This may increase the rational use of economic resources.

4. References

Cade B.S., Noon B.R. 2003 A gentle introduction to quantile regression for ecologists. Front. Ecol. Environ. 1: 412-420

Carvalho L., I. Dodkins, F. Carse, B. Dudley & S. Maberly. 2006. Phytoplankton classification tool for UK lakes. Final report, Project WFD38. SNIFFER, Edinburgh, UK.

Clark, J.S. 2005. Why environmental scientists are becoming Bayesians. Ecology Letters 8:2-14.

Clark, J.S. 2007. Models for ecological data: An introduction. Princeton University Press, Princeton, New Jersey, USA.

Clarke K.R. and Warwick, R.M. 1994. Change in marine communities: an approach to statistical analysis and interpretation. Plymouth Marine Laboratory, Plymouth, 144p.

De Boor, C. 1978. A practical guide to splines. Appl. Math. Sci. 27

Dodkins I & Rippey B. 2006. mCBASriv: A Method for Assessing the Ecological Status of Rivers Using Macrophytes. Scientific Summary v3.0, Internal Report to NSShare, March 2006.

Dodkins I, Rippey B & Hale P, 2005. An application of canonical correspondence analysis for developing ecological quality assessment metrics for river macrophytes. Freshwater Biology, 50, 891-904.

Gauch, H.G. 1982. Multivariate analysis in community ecology. Cambrigde university press, Cambridge

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. 2005. Bayesian Data Analysis. Chapman & Hall.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. John Wiley & Sons, New York

Hastie, T. and Tibshirani, R. 1986. Generalized Additive Models. Statistical Science 1(3): 297-310

Koenker, R. 2006. quantreg: Quantile Regression. R package version 4.02. http://www.r-project.org

Koenker, R. and G. Bassett. 1978. Regression quantiles. Econometrica, 46:33-50

Koenker, R., Ng, P., Portnoy, S. 1994. Quantile smoothing splines. Biometrika 81: 673-680

Kruskal, J.B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1-27

Lyche-Solheim, A. 2006. Dose-response relationships between biological and chemical elements in different lake types. REBECCA Deliverable 11. www.environment.syke.fi/rebecca.

Malve, O. 2007. Water quality prediction for river basin management. Dissertation for the degree of Doctor of Science in Technology. Helsinki University of Technology, Department of Civil and Environmental Engineering, Water Resources Laboratory, Espoo, Finland. TKK-DISS-2292. ISBN 978-951-22-8749-9. URL: http://lib.tkk.fi/Diss/2007/isbn9789512287505/index.html. 126 p. + app. 73 p.

Malve, O. and Qian, S. 2006. Estimating nutrients and chlorophyll a relationships in Finnish Lakes. Published in Environmental Science & Technology Sept. 2006.

Muggeo, V. M. R. 2004. segmented: Segmented relationships in regression models. R package version 0.1-4.

Portnoy, S., Koenker, R. 1997. The Gaussian hare and the Laplacian tortoise: Computa bility of squared-error versus absolute-error estimates. Statistical Sci. 12: 279-300

R Development Core Team. 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

R Development Core Team. 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Reynoldson, T.B., Bailey, R.C., Day K.E., and Norris, R.H. 1995. Biological guidelines for freshwater sediment based on BEnthic Assessment of SedimenT (the BEAST) using a multivariate approach for prediction biological state. Australian Journal of Ecology 20: 198-219

Saloranta, T., Kämäri, J., Rekolainen, S. and Malve, O. 2003. Benchmark criteria: A tool for select-ing appropriate models in the field of water management. Environmental Management 32(3):322—333.

Scharf FS, Juanes F, Sutherland W. 1998. Inferring ecological relationships from the edges of scatter diagrams: comparison of regression techniques. Ecology 79: 448–460

Scheffer, M., Carpenter, S. R., Foley, J. A., Folke, C., Walker, B. 2001. Catastrophic shifts in ecosystems. Nature 413: 591–596.

Scheffer, M., van Nes, E. H. 2004. Mechanisms for marine regime shifts: can we use lakes as microcosms for oceans? Prog. Oceanography 60: 303–319

Shepard, R.N. 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. Psychometrika 27: 125-140

Vollenweider, R. A. 1976. Advances in defining critical loading levels for phosphorous in lake eutrophication. Memorie dell'Istituto Italiano di Idrobiologia 33:53-83.

Wahba, G (1990) Spline Models for Observational Data CBMS-NSF Applied Mathematics No. 59 ISBN-13: 9780898712445 180 pp

Willby, N., Pitt, J., Phillips, G. 2006. Summary of approach used in LEAFPACS for defining ecological quality of rivers and lakes using macrophyte composition. Draft Report January 2006.

Wood, S.N. 2000 Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society (B) 62(2):413-428.

Wood, S.N. 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

Wright, J.F., Moss, D. Clarke,R.T. and Furse, M.T. 1997. Biological assessment of river quality using the new version of RIVPACS (RIVPACS III) Freshwater Quality: Defining the Indefinable (eds. P.J. Boon and D.L. Howell, pp 102-108. HMSO, Edinburgh.