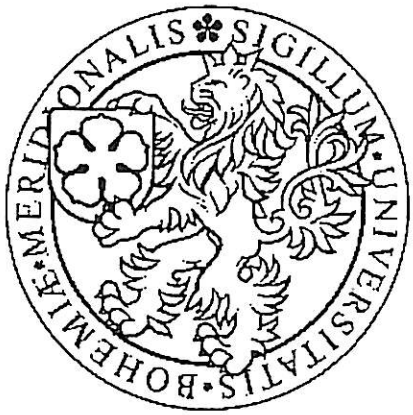
Petr Šmilauer

Modern approaches to analysis of ecological data



Faculty of Biological Sciences
University of South Bohemia
Ceske Budejovice
June 1996

prof. Kepš

Contents

Majce.

Prohlašuji, že jsem tuto práci vypracoval samostatně, pouze s použitím uvedené literatury.
České Budějovice, 22. června 1996

Petr Šmilauer

Introduction

There is a multitude of research problems that an ecologist has to deal with. There is also a multitude of approaches that can be used to evaluate the collected data. The traditional statistical methods never addressed those problems in fully adequate way. That is why the methods like the ordination methods flourished in the ecology much earlier than accepted by the 'official' statistical science. The time is changing and there are many new insights into appropriate ways of visualizing and analyzing the ecological data and the methods allowing us to do that are coming from various directions. In this thesis, I present a collection of papers covering some of my research aimed at improving the use of these methods in the field of ecology or at their application to the particular ecological research projects.

An important advance in the methods allowing us to get a global overview of the relationships in our data, to summarize their properties and to create new hypotheses to be tested, was the arrival of methods of the **direct gradient analysis** (ter Braak et Prentice 1988). These methods combine the ordination analysis traditionally used in the ecological research with the regression analysis approach and allow us to suggest more precise models of the relationship between the organisms and their environment. *Chapter 1* of my thesis makes an example of using the methods of constrained gradient analysis as a framework for an exploration of research data. The paper (which appeared in the Journal of Paleolimnology in 1994) also attempts to stress the need to exhibit the high level of self-criticism when applying models standing behind the ordination methods and check whether their particular use was an appropriate one.

The species response models models hypothesized using the results of the ordination methods can be explored in more detail by the regression methods. Even in this area the recent developments led to availability of new methods and many of the modern regression methods are much more appropriate for the application to the ecological data than the traditional linear methods of regression analysis. Among those, the **generalized additive models** (Hastie et Tibshirani, 1990) already achived attention among ecologist and have been found useful for modelling species responses to the gradients of their environment. The manuscript of paper presented in *chapter 2* discusses the advantages and problems of using generalized additive models for fitting the species - environment response models and also indicates possible implications for enhancements of methods of constrained ordination methods.

The paper in *chapter 3* (also in manuscript) provides a real-life example of using both types of statistical methods (direct gradient analysis and regression methods) to help to arrive at new findings in the ecological research.

Another example is provided in *chapter 4* which represents a single sub-chapter of the book that was accepted for publishing and has to appear at SPB Publishers before end of year 1996 (Prach et al., 1996). This chapter written with two co-authors (K. Prach, O. Rauch) deals with seasonal and inter-seasonal dynamics of Luznice river discharge and of the underground water table. During the study of these processes several types of modern regression methods were applied to arrive at realistic models.

Another chapter from the same book, presented in *chapter 5* of my thesis, uses rather different approach to modelling ecological processes. Here, a simulation model for the seasonal dynamics of nutrients and energy flows and storage in a river-floodplain ecosystem is developed. During the development of that model, a rather novel approach was used, with the architecture of the simulation model structured around the **rules-driven expert system**. The selection of expert system toolkit which employs the theory of fuzzy sets allowed me to integrate heuristic

knowledge about the ecosystem processes without the (unrealistic) need to estimate a huge array of system parameters that would need to be determined if more classical approaches were used.

While the relationship between the methods of artificial intelligence (as presented in the expert systems methodology) and the methods of constrained gradient analysis and regression analysis might not be clearly seen by the reader, the *chapter 6* of my thesis brings as an example the manuscript of paper describing an expert system being developed and helping the ecologists to manage the complexity of decisions needed when applying modern statistical methods to their data sets.

Acknowledgments

I want to thank to my wife Marie for all the help and support, enabling me to devote more time to the research than to the real life. Many thanks to Dr Jan Š. Lepš for being a fair supervisor. Last, but not least, I want to thank to Prof. John Birks, who provided me with much inspiration and support in my research.

References

R. Hastie, R. Tibshirani (1990): Generalized Additive Models. - Chapman and Hall, London.

K. Prach, J. Jenik, A. Large [eds] (1996?): Floodplain ecology and management. The Luznice River, Trebon Biosphere Reserve, Central Europe. - SPB Academic Publishers, Amsterdam, to appear.

C. J. F. ter Braak, I. C. Prentice (1988): A theory of gradient analysis. - Advances in Ecological Research, 18: 271 - 317.

Chapter 1

Exploratory analysis of paleoecological data using the program CanoDraw

Journal of Paleolimnology 12: 163–169, 1994.

Exploratory analysis of paleoecological data using the program CanoDraw

Petr Šmilauer
Faculty of Biological Sciences, University of South Bohemia, 31Branišovska, Ceske Budejovice, CZ 370 05, Czech Republic

Received 21 March 1994; accepted 3 August 1994

Key words: CANOCO, CanoDraw, exploratory data analysis, diatoms, canonical correspondence analysis, loess smoothing, generalized linear models

Abstract

This paper attempts to persuade the reader that methods of direct gradient analysis may serve as a basis for more detailed exploration of the data and that these exploration methods can be also used for checking the appropriateness of the assumptions of the applied ordination method. Generalized linear models and generalized loess smoothing together with several ways of data presentation are used to explore a sample data set.

Introduction

The introduction of methods of direct gradient analysis (DGA) (*sensu* ter Braak & Prentice, 1988) into ecological and paleolimnological research has marked the onset of significant, new approaches for examining complex data. The methods of DGA are an intuitive extension of the traditional methods of indirect gradient analysis (such as principal components analysis (PCA) or correspondence analysis (CA)) in the direction of more quantitative models of the processes that influence the actual (semi)quantitative values we have collected. The basic properties of these models are similar to those of a linear regression model, but are extended to include a whole set of dependent variables to be explained by a set of explanatory variables. In paleolimnological applications of these methods, these explanatory variables typically describe the limnological environment. The dependent (in a statistical sense) variables are not measured entities, but are composite gradients, similar to those revealed by 'traditional' ordination methods (like PCA) and are called ordination axes. The program CANOCO, written by ter Braak (1987a), provides a powerful implementation of all these techniques (including traditional indirect gradient analyses). With CANOCO, one can also test the statistical significance of the relations revealed and represented by the DGA results.

Another important feature of the way DGA methods were introduced into ecological research is the emphasis placed on the appropriate visualization of the ordination diagrams used to summarize the results of these methods. However, at the time CANOCO was released, there was no easy way to prepare appropriately scaled ordination diagrams. To make the problem worse, one often needs to prepare not only one, but several diagrams, either because one needs to look at the particular problem through its different facets or because squeezing all the multidimensional information into a single plot is incomprehensible. Also some experimenting with the presentation of results is needed if one wishes to present them in a concise, clear, and elegant way.

These considerations were the primary reasons for writing the program CanoDraw which began to be distributed with the CANOCO 3.1 package. Through further use of the methods of DGA, I realized that these methods are not only useful in their own right, but they could (and should) provide a reasonable base for exploring data by more detailed statistical and presentation methods. This is in good accordance with the now widely accepted view of ordination methods as tools to help in the formulation of new scientific hypotheses (or in the modification of existing ones). It is then very convenient if one can explore these new hypotheses in the framework in which they arose. This

could enhance and speed up the iterative process of exploring one's research data.

This is the direction that has influenced most of the contents of the new version of the CanoDraw program (version 3.0, released in 1993). In this paper, I present a few of the methods of data exploration available in CanoDraw.

Materials and methods

To demonstrate some of the capabilities of CanoDraw, I use the data set described in Cumming *et al.* (1991) and provided by H. J. B. Birks. In their paper, the authors first investigated the relationship between the abundances of scales of different chrysophyte species collected from surface lake-sediments and the measured values of several limnological variables. The data originated from a study of 25 Norwegian lakes and were summarized using 'canonical correspondence analysis' (CCA) (ter Braak, 1986). The authors used the 'forward selection' option in the program CANOCO (ter Braak, 1990) to select a subset of the measured limnological variables that seems to influence significantly the species composition of the samples. They then concentrated on the apparently most important factor, namely lake-water acidity. The other variables detected as having a significant explanatory role by the forward selection test procedure were conductivity, concentration of chloride anions (both these variables are related to the second ordination axis and may represent some influence of proximity to the sea), concentration of aluminium cations, and water colour. Based on the strong relation of the chrysophyte taxa to water acidity, the authors proceeded to model the values of the lake-water acidity. using the composition of chrysophyte assemblages recorded by their scales in lake sediments.

CanoDraw could be used to prepare the basic ordination plots, but it also provides methods for a more detailed exploration of the data. CanoDraw implements 'generalized linear models' (McCullagh & Nelder 1989) for modelling the 'responses' (or 'behaviour') of species, environmental variables, or other characteristics (such as 'fit of the species in the ordination space' or 'sites diversity') along a particular ordination axis or in the plane spanned by two ordination axes. For a less parametric approach, CanoDraw also implements a modified version of the 'loess smoother' (Cleveland, 1979). The CanoDraw' implementation uses an extension employing locally weighted generalized linear models, instead of the original locally weighted least squares regression. This allows sensible smoothing to be applied to counts or to probabilities. Another extension provided by CanoDraw is the choice of the polynomial order of the fitted model. Beside a strictly linear model, models with second or third order polynomial can be fitted, possibly with interaction terms (of the X and Y co-ordinates) for the response in the ordination plane. The choice of generalized linear models or generalized loess method is complemented by a third choice, which is an implementation of the 'universal kriging' model (Isaaks & Srivastava, 1989).

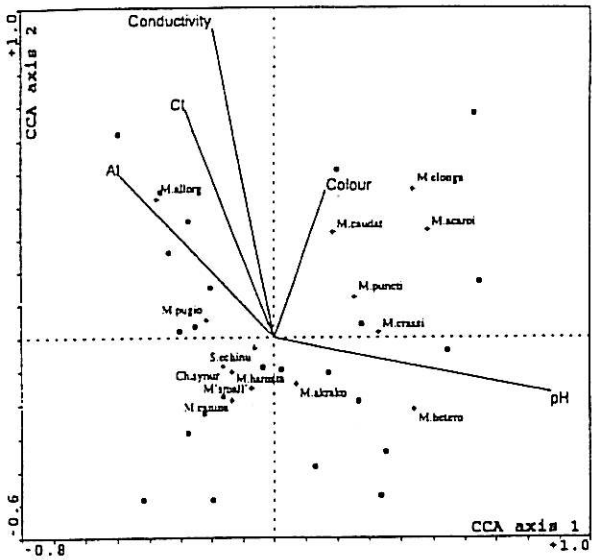


Fig. 1. Species – environmental variables biplot based on CCA with the 'significant' environmental variables chosen by CANOCO's forward-selection procedure. Environmental variables are plotted as arrows, the species positions are marked with crosses. The position of samples in the ordination plane is also plotted (filled circles), but they are not labelled.

When fitting the second-order polynomial form of a generalized linear model for species abundances/presences, CanoDraw provides estimates for the optimum (i.e. the mode of the litted unimodal curve), confidence intervals (if possible) and an estimate of the species 'tolerance' (i.e. measure of the width of the fitted response unimodal curve). These estimates are provided for fitting species response curves along an ordination axis, as well as along a gradient of a particular environmental variable.

Motivation

The application of the weighted averaging calibration procedure (and of the CCA method, too) centres around

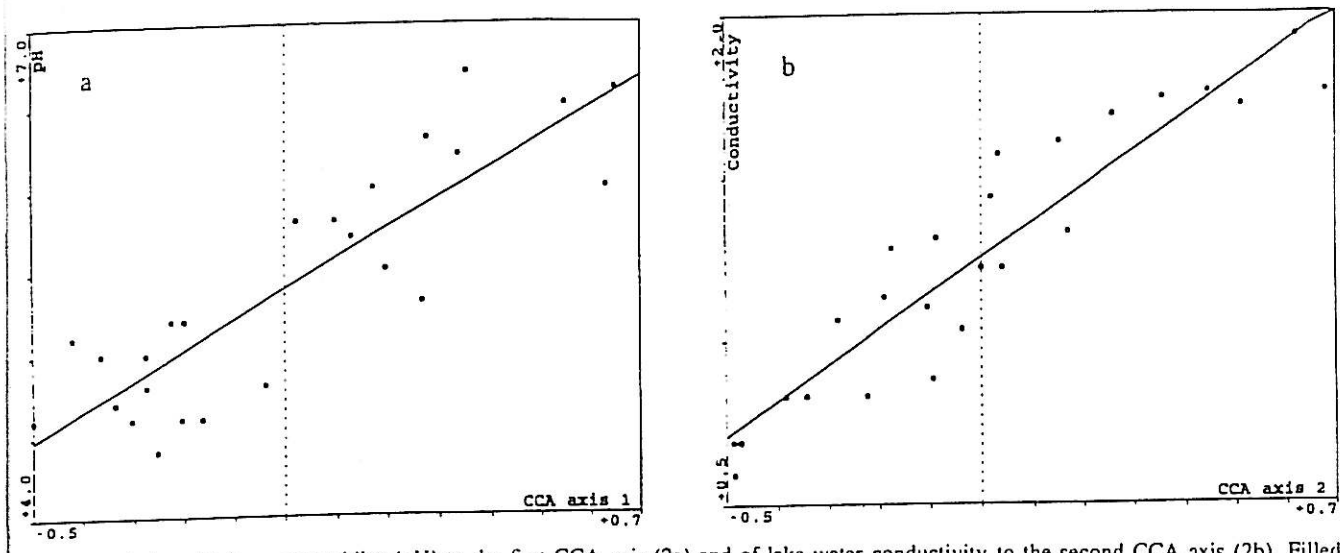


Fig. 2. Relation of lake-water acidity (pH) to the first CCA axis (2a) and of lake-water conductivity to the second CCA axis (2b). Filled circles represent individual sample values, the curve corresponds to the fitted first-order linear model. The original conductivity values were log transformed prior the analysis.

the assumption of the validity of the unimodal response model (ter Braak, 1987b). It may be useful to remember that this assumption does not imply that all the species in the analysed data are assumed to have an exact unimodal response. The linear response model where the expected value monotonically increases or decreases along the gradient of an explanatory factor (ordination axis in case of CCA, pH values in the case of the calibration problem) could be subsumed into the more general unimodal response model. Also, weighted averaging methods are known to be extremely robust to the violation of the *a priori* assumptions of the data properties. In any case, it might be worth looking at the ordination results in the same way that the methods of regression diagnostics (Cook & Weisberg, 1982) look at the regression modelling results.

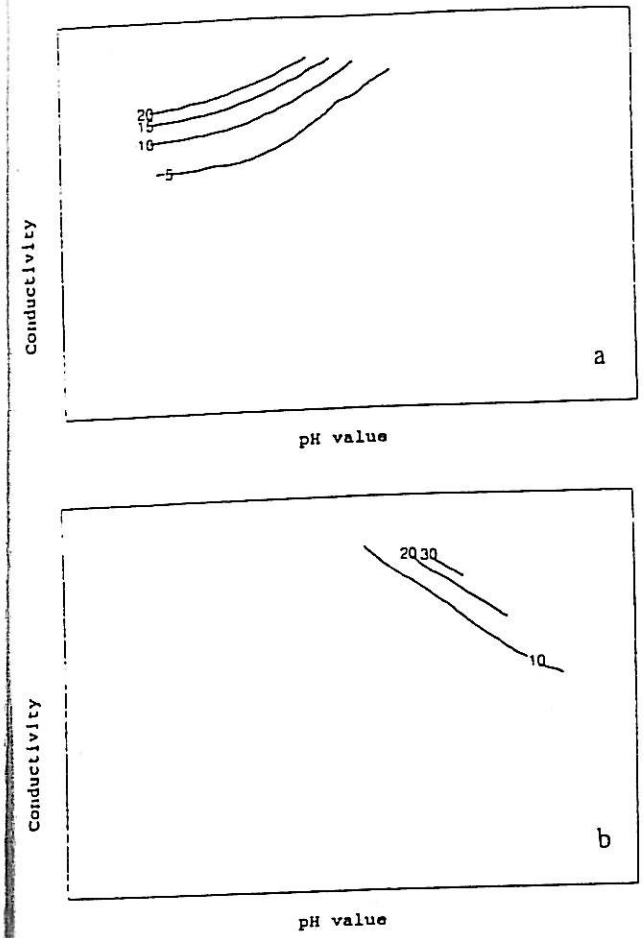
Another component of the CCA method that is worth investigating is the assumption that the ordination axes represent a sort of composite gradient along which the expected values of the explanatory variables change linearly (which is an assumption globally enforced by the definition of the ordination axes in CCA, being a 'linear combination of the explanatory variables'). Still, we should check the extent to which this assumption is fulfilled.

Results and discussion

Before starting a more detailed investigation, it is useful to recall the basic properties of the data. They can best be seen from the species – environmental biplot based on the CCA results (Fig. 1). Note that only the subset of explanatory variables selected in the forward selection procedure was used in the analysis. The contents of the figure and the method of its interpretation is described in Cumming *et al.* (1991). More general comments on the interpretation of ordination diagrams can be found in Jongman *et al.* (1987).

From this diagram, we can clearly see that the main gradient (the first ordination axis) corresponds clearly with water acidity. CanoDraw allows us to check this trend by plotting the pH values of individual samples against their position on the ordination axis. Figure 2a reveals the trend very clearly and the interpretability of the second ordination axis in terms of the conductivity values is confirmed by Fig. 2b. Note, however, that the conductivity is a compound variable, expressing synthetically the total concentration of ions in the water. The curves displayed in Fig. 2 correspond to a fitted linear model of the first order. Note, however, that the trend in Fig. 2b may be better described by a higher-order model. But the addition of a second-order term did not reduce the residual sum of squares significantly (CanoDraw provides the basic types of tests for hierarchical model selection).

The recognition of the two main environmental gradient influencing the species composition in our data allows us to provide an alternative view to the ordination diagram. We can create a 'reference plane' spanned by the values of the two environmental variables. In case of our set of environmental (explanato-



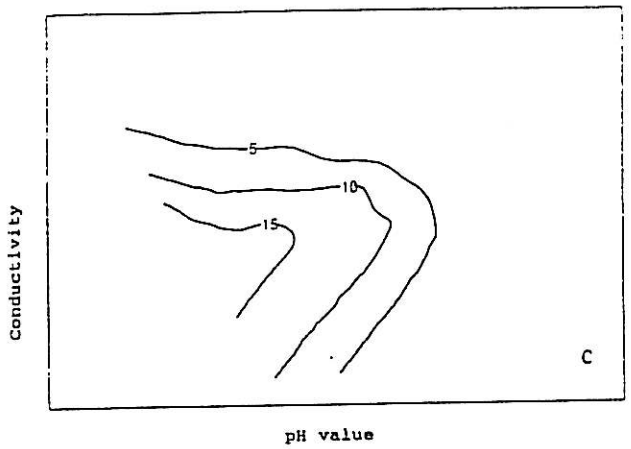


Fig. 3. Response surface of the relative abundances of *Mallomonas allorgei/lychenensis* (3a), *Mallomonas acaroides* (3b), and *Mallomonas canina/hindoni* (3c) in the plane spanned by gradients of pH and water conductivity. The surface is modelled by generalized loess with a first-order model. The contours represent the relative abundances of the species.

ry) variables including the most influential factors, we could consider such space to be a representation of the niche space of the species. The three species response-surfaces presented in Fig. 3 represent various preference types in this 'niche space'. The response surface is modelled using the generalized loess method, using the locally weighted first-order generalized linear model assuming a Poisson distribution of the abundances. *Mallomonas allorgeil/lychenensis* (Fig. 3a) shows preference for waters with high acidity and high conductivity. *M. acaroides* (Fig. 3b) prefers less acidic waters, but again with a high conductivity. *M. hindonii/canina* represents a taxon occurring in acid waters with a low conductivity (Fig. 3c).

Another way we can represent the relationship of interesting species to the most influential environmental variables is to project the abundances of the particular species onto the ordination plane, together with the directions of the most rapid changes in the values of the environmental variables. This can be illustrated for *Mallomonas crassisquama* in Fig. 4. Generalized loess smoothing was used again. Although the species response surface is much less formalized here than in the ordination diagram, the reader could still

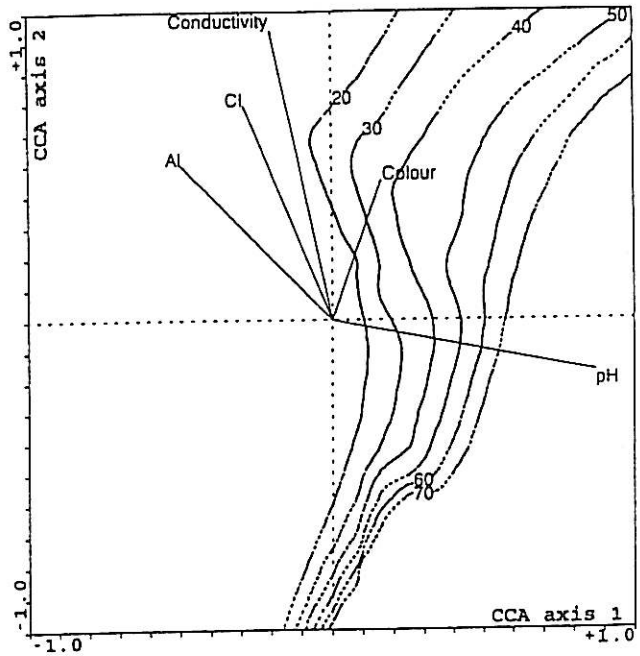


Fig. 4. Response surface of the relative abundances of *Mallomonas crassisquama* in the ordination plane spanned by the first two CCA axes. The surface is modeled with generalized loess procedure using a second-order model (including an interaction term). The contours represent the relative abundances of the species. The dotted segments of the contour lines represent extrapolated parts that are much less reliable. The arrows represent directions of the greatest increase in values of the corresponding environmental variable.

wonder whether the real pattern has been removed by the data-smoothing method. But by looking at the so-called symbol plot (Fig. 5) for the same species, we can clearly see that the trend visible in Fig. 4 has support in the actual data. In Fig. 5 the real abundances are

Table 1. Weighted averaging and maximum-likelihood estimates of optimum and tolerance of studied chrysophyte species along the gradient of water acidity

Species Name	pH - WA optimum	pH - WA tolerance	pH - ML optimum	pH - ML tolerance
Mallomonas acaroides	5.97	0.50	LINEAR	—
M. akrokomos	5.77	0.98	N.S.	—
M. allorgei/lychenensis	4.73	0.23	4.59	0.35
M. canina/hindoni	5.16	0.44	N.S.	—
M. caudata	5.52	0.61	N.S.	—
M. crassisquama	6.00	0.57	LINEAR	—
M. elongata	5.91	0.47	N.S.	—
M. hamata	5.17	0.51	5.20	0.61
M. heterospina	6.55	0.21	LINEAR	—
M. pugio	4.56	0.19	LINEAR	—
M. punctifera	5.92	0.46	N.S.	—
M. 'small'	5.41	0.58	LINEAR	—
Synura echinulata	5.32	0.77	N.S.	—
S. lapponica	5.83	0.61	LINEAR	—
S. sphagnicola	4.87	0.45	1.07	1.65
Chrysodidymus synuroideus	5.07	0.62	LINEAR	—

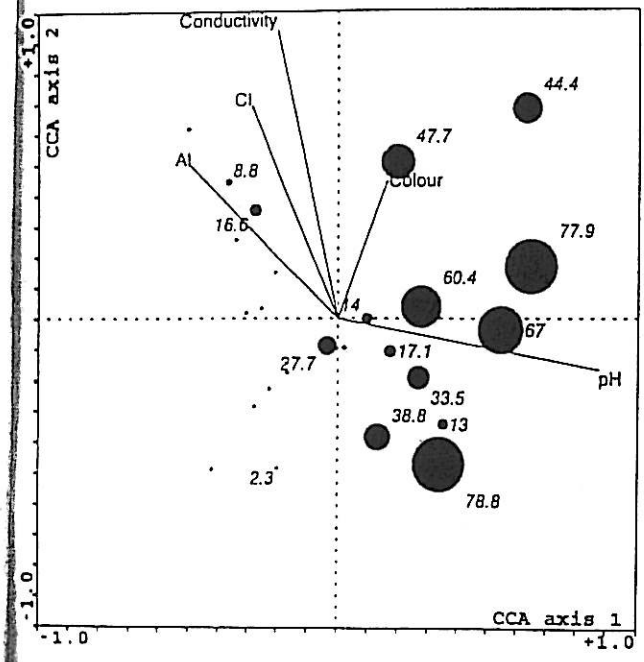


Fig. 5 Symbol plot of the relative abundances of *Mallomonas crassisquama* in the ordination plane spanned by the first two CCA axes. The size of the circle gives a visual idea of the relative abundance of the species in a particular sample, while the labels specify the exact value. The arrows represent directions of the greatest increase in the values of the corresponding variable.

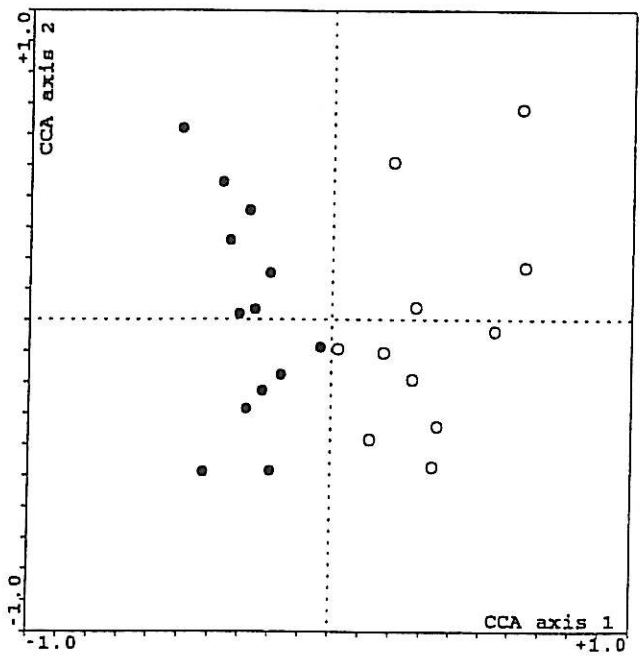
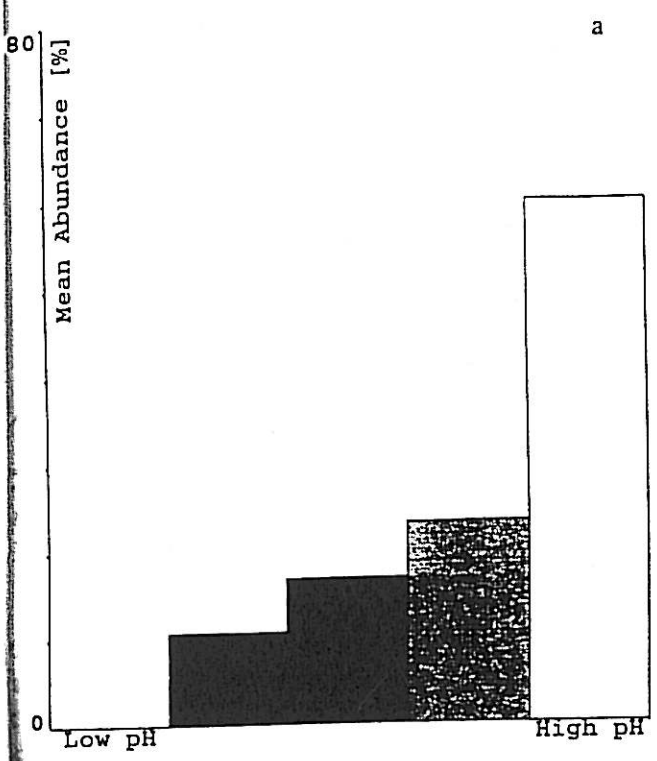
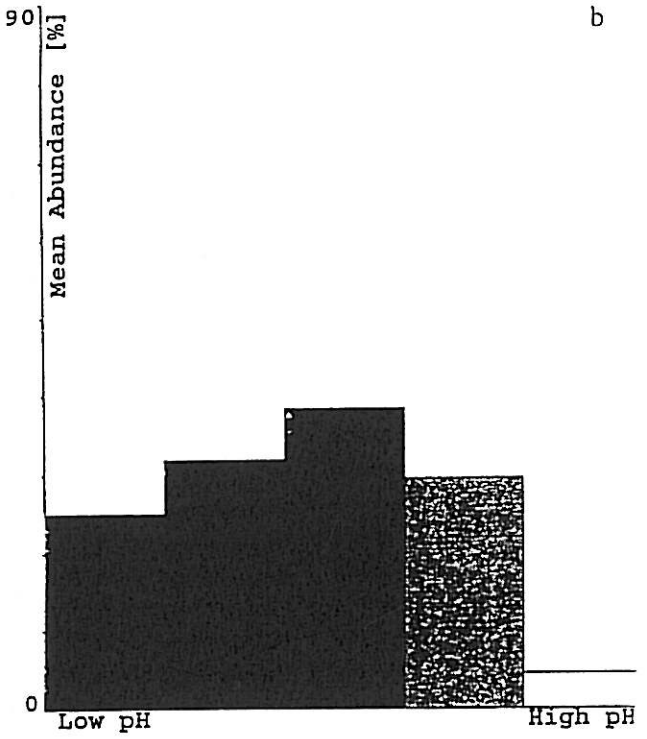


Fig. 6. Position of sites in the ordination plane spanned by the first two CCA axes. The samples are classified according their pH value into those having a value below the median pH (filled circles) and those having their value above the median (empty circles).

displayed with the size of the filled circle proportional to the abundance value. Unlabelled marks represent absence of the species.

Another method of representing the relation of a species to a particular gradient is to classify the samples





according to their values for the studied variable (pH value in our case). CanoDraw allows one to split the samples into several distinct classes using the values of either particular environmental variable or particular species. It is possible to choose one of ten possible strategies for using the values to split the set of samples. In this particular case I could have decided to split the data into two (nearly) equal groups, i.e. those below and those above the median pH value. In Fig. 6, the resulting classification of samples is plotted into the ordination plane spanned by the first two ordination axes. The filled circles represent the 'more acid half' of the data set. However, to achieve more visually informative plots, I split the samples according to pH values into five groups of the same size (i.e. 5 samples in each class). Then I could display the mean abundance for a particular species in these five classes, linearly arranged. From the five methods available for doing this in CanoDraw, histograms seemed to be most appropriate. Figure 7 displays three species representing three basic types of species relationship to the acidity gradient.

This could also be summarized in one plot, displaying the fitted response curves (of the second order) for the three species along the pH gradient (Fig. 8).

Beside visual checks of the reliability of the patterns indicated by the ordination results, we could check our assumption in a more formal way. One way is based

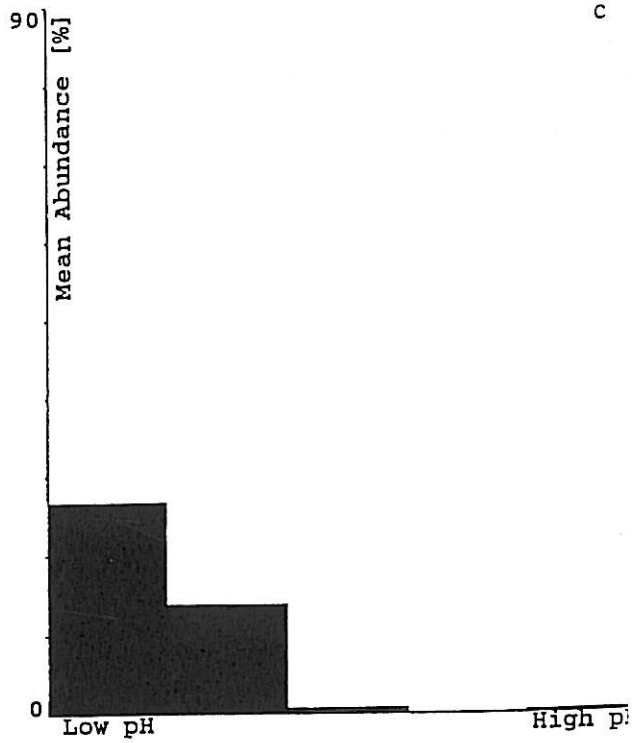
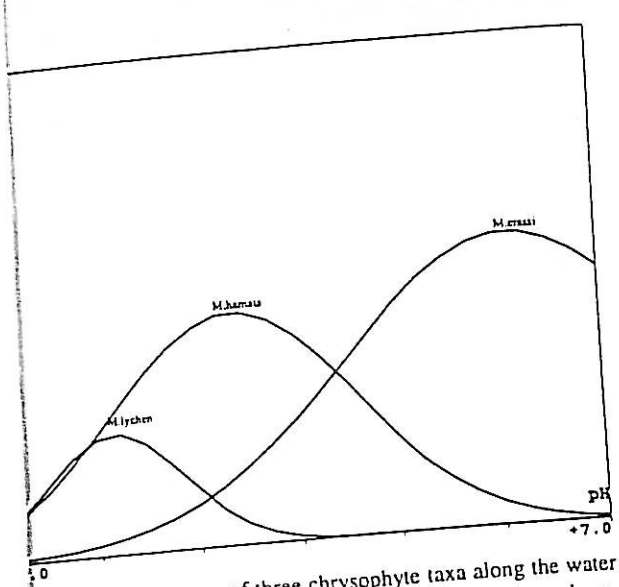


Fig. 7. Mean abundances of *Mallomonas crassisquama* (7a), *M lomonas hamata* (7b), and *Mallomonas allorgei/lychenensis* (7c) five sample classes. The classes are defined by the pH value of t samples (lakes) and they are ordered from the most acid to the le acid.



8. Response curves of three chrysophyte taxa along the water ity gradient. The curves represent a fitted second-order polyno- generalized linear model.

he idea that ordination axes resulting from weight- averaging methods represent composite gradients best fit the 'species packing model' where the cies optima are supposed to be distributed uniform- cross the gradient having unimodal response curves pproximately the same width (see ter Braak, 1986). their paper, Cumming *et al.* (1991) calculated pH ima for the most important species, using weight- averaging regression, and they also calculated their erance using the same method. In Table 1, these val- are in the second and third columns, respectively. ing CanoDraw, I calculated the optima based on the ression coefficients of a generalized linear model he second order, using the formulae presented by Braak & Looman (1986). These values are given he fourth column, but only if the regression model s statistically significant. If such a model was not nificant, the column contains either 'LINEAR' (for ignificant 'linear' relation between pH and species undance) or 'N.S.' (if there is no significant rela- n between species abundances and pH values). The reshold level for 'significance' was selected to be = 0.075. Where optima could be calculated. the tol- nces were also calculated based on the regression odel fitted.

In conclusion, I have tried to demonstrate that while e information provided by the method of direct gra- ent analysis is very valuable for certain types of pplications, it is sometimes useful to check on the viation of the real data from the underlying assump- ons of these methods. This closely parallels the role regression diagnostics in statistical modelling using nventional regression methods.

Software availability

Program CanoDraw, version 3.0 is available from Microcomputer Power, 111 Clover Lane, Ithaca, NY 14850 – 4930, U.S.A.; tlf. +1-607-272-2188, fax. +1-607-272-0782. Educational/site licenses are also available. CanoDraw can be run on IBM-PC compatible computer with I80286 microprocessor or higher. A VGA graphics card and mouse are needed.

Acknowledgments

I would like to thank to Prof. John Birks not only for providing the analysed data but mainly for his overall support in preparing this paper. Other persons who participated in collecting the data are also acknowledged. Many thanks are due to two reviewers of the paper for their valuable comments.

References

Cleveland, W. S.. 1979. Robust locally-weighted regression and smoothing scatterplots. J. Am. Statist. Assoc. 74: 829–836.

Cook, R. D. & S. Weisberg, 1982. Residuals and influence in regression. Chapman & Hall, New York. 230 pp.

Cumming, B. F., J. P. Smol & H. J. B. Birks, 1991. The relationship between sedimentary chrysophyte scales (Chrysophyceae and Synurophyceae) and limnological characteristics in 25 Norwegian lakes. Nord. J. Bot. 11: 231–242.

Isaaks, E. H. & R. M. Srivastava, 1989. An introduction to applied geostatistics. Oxford University Press, New York, 561 pp.

Jongman, R. H. G., C. J. F. ter Braak & O. F. R. van Tongeren, 1987. Data analysis in community and landscape ecology. Wageningen, Pudoc, 299 pp.

McCullagh, P. & J. A. Nelder, 1989. Generalized Linear Models. Second edition. Chapman & Hall, London, 511 pp.

ter Braak, C. J. F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67: 1167–1179.

ter Braak, C. J. F., 1987a. CANOCO – a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis, and redundancy analysis (version 2.1). Agricultural Mathematics Group. Wageningen.

ter Braak, C. J. F., 1987b. Unimodal models to relate species to environment. Doctoral thesis. University of Wageningen, 152 pp.

ter Braak, C. J. F., 1990. Update Notes: CANOCO version 3.10. Agricultural Mathematics Group, Wageningen, 35 pp.

ter Braak, C. J. F. & C. W. N. Looman, 1986. Weighted averaging, logistic regression and the Gaussian response model. Vegetatio 65: 3–11.

ter Braak, C. J. F. & I. C. Prentice, 1988. A theory of gradient analysis. Adv. Ecol. Res. 18: 271–317.

Chapter 2

Modelling species - environment response curves: can we do better?

Petr Šmilauer: Modelling species - environment response curves: can we do better?

Introduction

Statistical models for responses of plant or animal species to the important environmental gradients are an important tool for expresing and summarizing our knowledge on the behaviour of these species in their environment (Whittaker 1967, Austin et Gaywood 1994). While these regression models are used for a rather long time, only recently a more systematic approach to the modelling was suggested in Huisman, Olff, Fresco (1993). Given the fact that the traditional way of describing the relationship between species performance and the properties of the environment (namely by a linear regression model) is often too crude an oversimplification to be useful, new methods are sought that might be better suitable for that purpose.

The first logical step is to get rid of the unrealistic assumptions about the properties of the response variable (predicted performance of the population or predicted probability of occurrence of the species given particular environmental conditions): the counts of individuals, cover of the aboveground biomass, expected competitiveness indices or probability of occurrence have properties far from those assumed by the methods of statistical inference applied to the fitted linear models (Jongman, ter Braak, van Tongeren 1987). While ecologists analyzing such kind of data always had some sort of transformations available, their use is not always as intuitive and easy to do as one might like it to be. The generalized linear models (GLM, McCullagh et Nelder 1989) relieve us of this burden while holding some of the important, simplifying assumptions of the classical linear model (additivity of the influence of the explanatory variables, linearity of the regression model in its parameters, simple relationship between the predicted values and the expected variability of the predictions). The GLM already found their place in statistical analyses of ecological data (e.g. ter Braak et Looman 1986, Austin et al. 1994, Ferrer-Castan et al. 1995).

Further step towards the more faithful (and also more complex) models is represented by the generalized additive models (GAM, Hastie et Tibshirani 1990). The use of GAM in plant ecology for modelling species responses to their environment was already suggested by Yee et Mitchell (1991) and by Leathwick (1995). These models have an appeal to everyone who tried to use the more simplistic statistical models to describe the relationships emerging from his/her data and failed: in many cases the GAM do not fail on the same data. I have participated in several research projects where both GLM and GAM were successfully applied and provided new insights into the data. In this paper, I would like to: (a) discuss some motivations for applying GAM instead of the (generalized) linear regression models, (b) provide warning about the new traps that emerge hand in hand with the new freedom provided by these models, (c) suggest some improvements to the application of these models to the ecological data, and (d) show some interesting relations between semi-parametric modelling approach (as exercised by the use of GAM) and multivariate statistical methods (ordination methods).

Short exposure to GAM

Generalized additive models (GAM) can be viewed as a natural, less-parametric extension of the generalized linear models (GLM). But, let us start from the classical linear model. For simplicity, I will suppose we have one response ("dependent") variable **Y**, representing in some way the performance of the species, and we have two explanatory variables **X1** and **X2** influencing the species performance. The generalization for more explanatory variables is then easy to derive. The **classical linear regression** model describes the values of the response variable **Y** as realizations of random variate with a Normal distribution with the same (constant) variance for all the realizations (observations) and with the **expected** (mean) **value** for a particular combination of the values of the explanatory variables expressed as follows:

$$EY = \beta_0 + \beta_1 * X1 + \beta_2 * X2 \qquad (1)$$

Generalized linear models generalize the classical linear model in two respects:

(a) the type of the statistical distribution the response variable is supposed to come from any of the so-called exponential family of distributions. The most typical members of this family are the Poisson distribution for counts, gamma distribution for biomass or the binomial distribution relating to the probability of occurrence. The Gaussian distribution is one possible case, too. The important consequence of this generalization is that the variability of the response variable values is no longer constant: it is supposed to change in a systematic manner with the expected value of the postulated distribution. For example with Poisson distribution the variance is supposed to be equal to the expected value ("mean") of the response variable.

(b) while the GLM still maintain the linearity and additivity of the influence of individual predictors (explanatory variables), these hold only on the scale of the **linear predictor**. For our example, the linear predictor might be defined as:

$$\eta = \beta_0 + \beta_1 * X1 + \beta_2 * X2 \qquad (2)$$

This specification is in GLM complemented by the choice of **link function**, which relates the expected value of the response variable to the linear predictor:

$$\eta = g(EY) \qquad (3)$$

The link function *g(.)* is a simple monotonic function that transforms the values onto the scale of linear predictor (generally any real value) from the restricted scale of the response variable (values in the range 0 to 1 for probability of occurrence, positive values for biomass, non-negative values for counts etc.). These link functions will look very familiar to the practitioners who used the more traditional transformation methods, e.g. logarithmic link function for counts. It must be stressed, however, that in the GLM the transformations like logarithmic one do refer to **expected** values, not to the observed data.

Approaching finally the **generalized additive models**, I must start from the fact that these are a logical extension of GLM, differing only in the more flexible specification of the systematic part

of the model, namely linear predictor of GLM. This one is called **additive predictor** here and its specification is written in the form

$$\eta = \beta_0 + f_1(X1) + f_2(X2) \qquad (4)$$

where $f_i(.)$ are some smooth, semi-parametric curves, of different shape for different explanatory variables. Note that a GLM for our example is just a special case with $f_1(x) = \beta_1 * X_1$ and $f_2(x) = \beta_2 * X_2$. The property of additivity in the GLM and in the classical linear models is retained (on the scale of the additive predictor), so we can visualize a fitted model by displaying the individual fitted functions $f_i(.)$ of each of the predictors. One important consequence emerging from the extra freedom in modelling the contribution of individual explanatory variables is that we can select for a particular predictor from various functions differing in their complexity. Clearly, we cannot select the one fitting "best" our data sample: we would invariably end up with functions interpolating our data points but doing little for the generalization of their patterns or for prediction of new values. The complexity of the curve can be expressed in the terms of the **degrees of freedom** taken from our data by the model term corresponding to a particular explanatory variable (Hastie et Tibshirani, 1990). There are various **model selection strategies** known for the classical linear regression models. All of these are in principle available in GAM, as well. A complication is that when working with GAM, we should find out not only **which predictors** enter the final model but also **what amount of information** these should convey in the model. Beside the complexity of the function f_i used for a particular predictor, a further problem we need to address is what type of smoother should be considered. Nevertheless, this problem does not seem to be as essential as the previous ones: because the data 'drive' the selection of the appropriate shape of f_i, the resulting curves are rather similar for the types of smoothers commonly used with the GAM (most often smoothing cubic spline - Eubank, 1988 or the loess smoother - Cleveland, 1979).

Advantage of GAM

1. The most important feature that makes GAM different from the linear models is their localized behaviour: Fig. 1 shows an example of fitting ecological response model describing the dependence of the relative frequency of grasses (pollen type Gramineae) in a pollen sample from an area with particular amount of precipitation (non-published data, from H.J.B. Birks). Two models were fitted to the data, each of them with the same amount of complexity (two degrees of freedom) in the model term corresponding to the rainfall amount: the thin line displays a GLM with second-order polynomial dependency of Gramineae occurrence on the rainfall. The thick line is a GAM, with cubic smoothing spline with two degrees of freedom. We can see that the parametric model follows a predetermined bell shape, while the GAM describes more faithfully the data, with an increase of occurrence for precipitation up to 1500 mm and then with the fraction levelling of (or slightly decreasing).

******** Figure 1 about here *********

Similarly, GAM with the equivalent complexity are much better at describing unimodal responses in the case where the response shapes are not symmetric around the optima. In summary, the models like the GAM (but not only these) do not expose us to the undesirable consequences of linearity in the traditional regression models. Linearity in X further implies that the extent of response of a species performance to an unit change in the environmental factor is still the same (or that the response changes in a monotone way for the models involving transformations) whatever position on the environmental gradient we consider.

2. Another neat consequence of having a semi-parametric model plugged into GAM is the more reasonable way we can extend them to model interactions among the factors. Ecological factors often do not exhibit interaction over the whole range of their gradient (availability of various nutrients or elements or food resources) or differ in the extent (or form) of their interaction for different parts of the gradients (interactions among nutrients availability and water availability). This is too complicated for an appropriate description using the (generalized) linear models. Of course, modelling interactions among factors directly is always quite difficult task and one of the main strengths of the GAMs is that the complexity of the non-parametric description of patterns is limited by their assumption of additivity (on the additive predictor scale) of influence of individual factors (explanatory variables).

3. The last advantage of GAM I need to mention lays in the similarity and continuity of their philosophy with the more traditional approaches, namely GLM. As the GLM are just a special case of GAM, we can include the GLM as the ultimately simple solution on our quest of faithful but parsimonious model. The Occam's Razor principle asks us here to pick up the simplest model suitable for our purpose. For model selection based on statistical inference, methods similar to those used to compare competing linear models (usually hierarchically arranged) are available for GAMs as well (selection based on **analysis of deviance** - McCullagh et Nelder, 1989, p. 35).

Problems with GAM

The most difficult problem I have faced when working with GAM relates to model selection: more freedom means more danger of getting off the right track. There have been several methods suggested for dealing with the problem of model selection - those similar to the stepwise selection approaches of classical regression analysis are still available (Hastie et Tibshirani, 1990, p. 260). The search space where the (semi-)automated selection procedure has to look for the model is much more complex, however. We have to consider not only the presence of the explanatory variable in the model, but also the complexity of description brought into the model by the particular variate. For example, the model terms considered for a particular explanatory

variable might include its absence, linear form (as part of GLM), smooth term with 2 degrees of freedom and smooth term with 4 degrees of freedom. But in reality, we should not be restricted to integral number of degrees of freedom: a smooth curve with 2 degrees of freedom might be **oversmoothed** (not following sufficiently the trends in the data), while the curve with 3 degrees of freedom is already **undersmoothed** (too "wiggled").

Beside the problem with the complex search space for the model, the stepwise selection procedures often tend to select too complicated models. For classical regression models, model selection measures were devised that weight the extent the model fits the data against its complexity. The best known is the Mallow's criterion (Cp - Mallow, 1973). This criterion has a generalization for the case of GLM (Akaike information criterion, AIC) and this one might be applied with approximate validity during the selection of GAM as well (Hastie et Tibshirani, 1990, p. 158).

An alternative (and more appropriate) way of selecting model is to minimize the prediction error, i.e. to optimize the model performance in respect to the new, not yet collected observations. This is very appealing and intuitive approach for modelling species responses along environmental factors. The so called **cross-validation** approach attempts to estimate the true prediction error for a particular model by repeatedly dividing the data set into one part used for fitting the model and the other part used for evaluating the model performance. The commonly applied type of cross-validation is the **one-leave-out** type, where in each iteration single observation is omitted from the data set, the remaining observations are used to fit the model and the performance of the model (using for example the squared difference between the true and predicted value of the response variable) is evaluated using that single observation (Efron et Tibshirani, 1993). It turns out that under certain assumptions the model selection statistics like Mallow's Cp or AIC approximate the results obtained by the one-leave-out cross-validation method (Stone, 1977).

When working with various types of data sets, fitting response curves and surfaces for different types of organisms and different types of environmental factors, I have found that the model selection using the AIC - based procedure often leads to 'over-complicated' model specifications, where the visualized responses of the species performance to the environmental gradients are too difficult to generalize and interpret. The literature about this topic is scarce, but published research papers (Shao 1993, Shao et Wu 1989) indicate that the problem lays in the inherent similarity of the Cp or AIC statistics to the prediction error estimate based on the one-leave-out cross-validation. These papers suggest that a raise of the fraction left out from the 'training sample' makes the estimated prediction error less biased towards the over-complicated model specifications (Shao 1993).

I have been experimenting with the cross-validation procedures and found that the models based on the **k-leave out** cross-validated residual deviance are more parsimonious and almost always interpretable if compared with the models selected by other methods. I have been applying the procedure where the data set is randomly splitted into two halves, one being used for

model fit and the other for the model evaluation. Models selected in that way often suggest only simple linear relationship, while the models selected using stepwise selection procedure based on a deviance analysis indicate more complicated response curve.

Suggestion for use of GAM

For modelling response curves (i.e. in the situation where a single environmental factor is considered) the following schema is suggested:

1. The average predictive squared error (PSE, Hastie et Tibshirani, 1990, p. 42) is approximated using the cross-validated deviance (which is a generalization of the usual cross-validated sum of squares for the more general notion of **model deviance** - see Nelder, McCullagh, 1989) and evaluated for a series of model differing in the type and complexity of the explanatory variable entering the model.

2. I have found that GAM based on smooth curves with more than 5 degrees of freedom are often too complicated for ecological interpretation of their shape or (alternatively) differ from those with lower complexity in too subtle (not interpretable) ways. Most often, I have been evaluating the PSE of the model, considering the candidate specifications for the explanatory variable given in Table 1.

The inclusion of the $X+X^2$ term (describing parametric, second-order polynomial dependence) is sometimes questionable and largely depends on the context and purpose of the study. In many situations, the smooth term with the same amount of complexity (i.e. with two degrees of freedom) provides a better fit to the data.

3. Each of the candidate models is evaluated several times by creating independent random splits, because the *k*-leave-out cross-validation tends to produce estimates of PSE of increasing variance as the *k* increases up to *n/2*, which is the value suggested here and elsewhere (Shao, 1993) The variance of the estimates increases with the term complexity, too. The model with the minimum average estimate of PSE is then selected and fitted to the whole data set.

********* Table 1 about here **********

If we want to define a **response surface** (where more environmental factors are involved) we should not base our model selection on marginal performance of the individual model terms. The environmental factors often exhibit high degree of relatedness - a phenomenon which in the context of GAM was named **concurvity**, a parallel to the traditional term of **collinearity** (Hastie et Tibshirani, 1990, p. 123). As the reader might expect, the model terms selected jointly into the "best" model are usually less complicated if compared with their specification when the factors are considered individually, because they share some part of their information. With more environmental factors considered jointly in the fitting of species response surface, another

problem emerges and that is whether to explicitly model any interaction terms and how to appropriately express such interaction. This area is rather unexplored even in the statistical literature (Hastie et Tibshirani, 1990, p. 264), the more then in its application in ecology. In fact, ecologists have a different understanding to the term 'interaction', namely interaction among them in their effects on a response variable. Using this view, the interaction between various ecological factors entering our model of the response surface might emerge from a comparison of the marginal response curves (where each of the environmental factors is used separately) with their shape in the joint GAM.

Response models and multivariate methods

Modelling the response of species to environmental gradients using GAM leads to an important, conceptual problem that the ecologist using these models has to tackle: *Shall we adopt such scaling of the environmental factor that makes the data fit our simple conceptual model (like symmetric unimodal response curve) in the best way or is it better to stay with the original scale of measurement (pH values, nutrient concentrations in mg.g^{-3} etc.)?* This problem was already discussed for example by Økland (1986) for modelling community composition response to the environmental gradients. Økland suggested to accept the scale of the ordination axes of DCA (Detrended Correspondence Analysis) method because this scale fits best the response of individual species with the symmetric unimodal response curve, underlying the DCA method (ter Braak, 1985).

The same question has to be answered for the univariate response models where a single environmental factor takes the role of the composite gradients represented by the ordination axes. Both problems mix together in the case of the **direct gradient analysis** (sensu ter Braak et Prentice, 1988). For example, in the CCA (Canonical Correspondence Analysis, ter Braak et Prentice, 1988) the unimodal responses of species along the gradients represented by the ordination axes are assumed, but the ordination axes are further restricted to be a linear combination of the explanatory variables (environmental factors, most often) - see Carlton (1990) for an example of using CCA analysis solution as a basis for modelling species responses.

Consequently, we would profit enormously from the situation where scale of the individual environmental factors would lead to constrained ordination axes that fulfil best the assumption of unimodal responses of species along these gradients. In other words, restricting the CCA axes to be a **linear combination** of submitted environmental variables expressed on *ad hoc* scales often leads to sub-optimal ordination results. Technically, an improvement might be achieved by an iterative procedure, where:

1. The environmental variables are used on their original scale in a CCA.

2. A semi-parametric (GAM-like) transformation of these environmental variables at their current scale is done, to linearize their relation with the current first (few) ordination axis (axes). A

suitable method for doing that is provided by AVAS (Additivity and VAriance Stabilization, Tibshirani, 1988).

3. The transformed environmental factors are used for a new CCA and steps **2** and **3** are repeated until convergence is achieved (signified by the semi-parametric transformation suggested for all the variables being near to linear). The experiments done so far showed that the convergence happens almost instantly (in 2 - 3 iterations) and the suggested transformations are most often easy to interpret (e.g. transformation similar to log transformation for the age of stands when modelling changes in plant communities during succession).

This is an area of further research, but results achieved so far are promising.

References

M. P. Austin et M. J. Gaywood (1994): Current problems of environmental gradients and species response curves in relation to continuum theory. - J. Veg. Sci., 5: 473 - 482

M. P. Austin et al. (1994): Determining species response functions to an environmental gradient by means of a β-function. - J. Veg. Sci., 5: 215 - 228

T. J. Carleton (1990): Variation in terricolous bryophytes and macrolichen vegetation along primary gradients in Canadian boreal forests. - J. Veg. Sci., 1: 585 - 594

W. S. Cleveland (1979): Robust locally weighted regression and smoothing scatterplots. - J. Am. Stat. Assoc., 74: 829 - 836

B. Efron, R. J. Tibshirani (1993): An Introduction to the Bootstrap. - Chapman and Hall, N. York

D. Ferrer-Castán et al. (1995): On the use of three performance measures for fitting species response curves. - J. Veg. Sci., 6: 57 - 62

T. J. Hastie, R. J. Tibshirani (1990): Generalized Additive Models. - Chapman and Hall, London

J. Huisman, H. Olff, L. F. M. Fresco (1993): A hierarchical set of models for species response analysis. - J. Veg. Sci., 4: 37 - 46

J. R. Leathwick (1995): Climatic relationships of some New Zealand forest tree species. - J. Veg. Sci., 6: 237 - 248

P. McCullagh, J. A. Nelder (1989): Generalized Linear Models. Second Edition. - Chapman and Hall, London

R. H. Okland (1986): Rescaling of ecological gradients. II. The effect of scale on symmetry of species response curves. - Nord. J. Bot., 6: 661- 669

J. Shao (1993): Linear model selection by cross-validation. - J. Am. Stat. Assoc., 88: 486 - 494

J. Shao, C. F. J. Wu (1989): A general theory for jackknife variance estimation. - The Annals of Statistics, 17: 1176 - 1197

C. J. F. ter Braak (1985): Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. - Biometrics, 41: 859 - 873

C. J. F. ter Braak, C. W. N. Looman (1986): Weighted averaging, logistic regression and the Gaussian response model. - Vegetation, 65: 3 - 11

C. J. F. ter Braak, I. C. Prentice (1988): A theory of gradient analysis. - Advances in Ecological Research, 18: 271 - 317

R. J. Tibshirani (1988): Estimating transformation for regression via Additivity and Variance Stabilization. - J. Am. Stat. Assoc., 83: 394 - 405

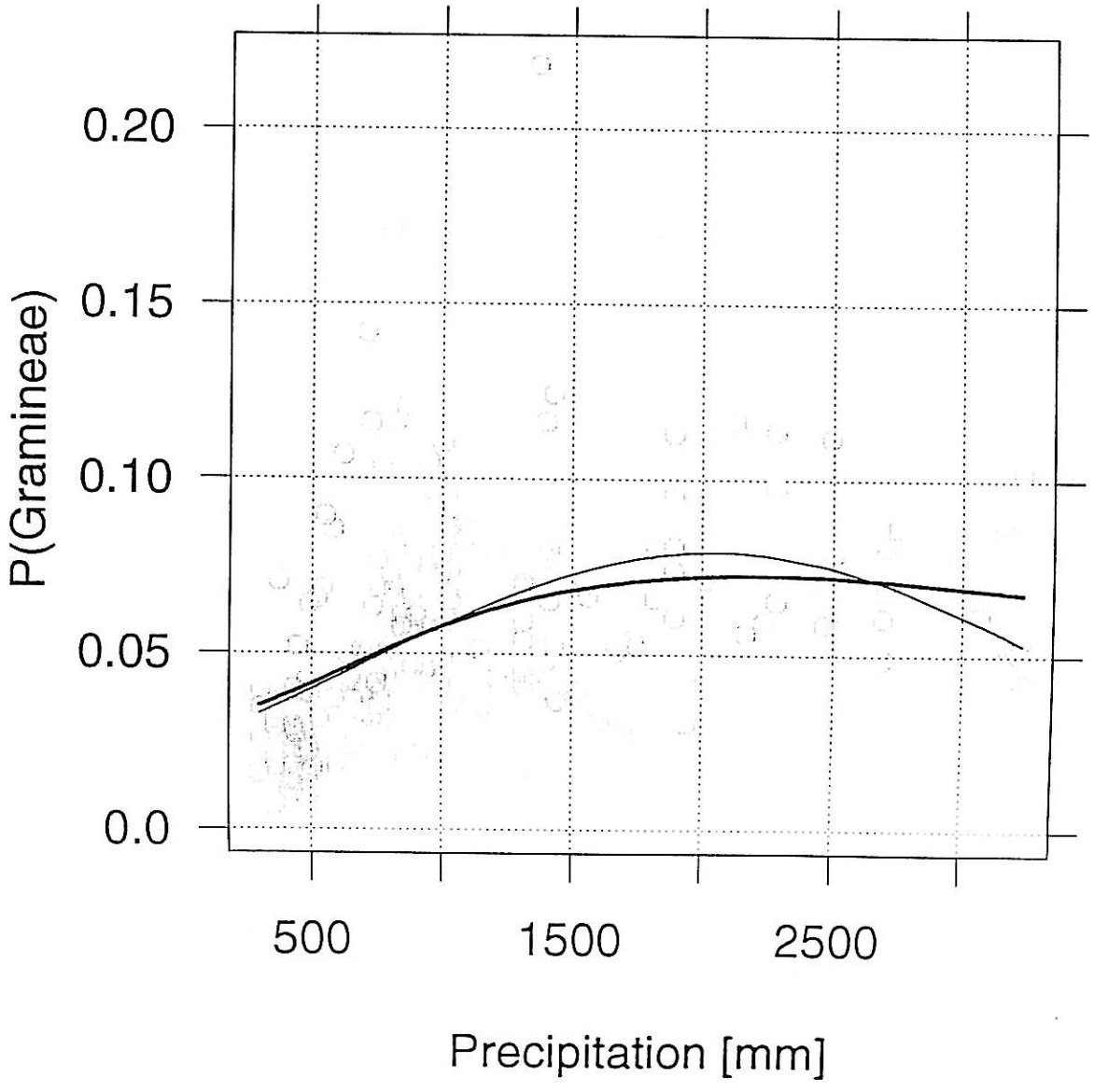
T. W. Yee, N. D. Mitchell (1991): Generalized additive models in plant ecology. - J. Veg. Sci., 2: 587 - 602

0	null model, the response variable is not expected to change with the values of the predictor)
X	response curve is a GLM
$X+X^2$	response as a second-order polynomial in a GLM
s(X,df=2)	GAM with smooth term with 2 degrees of freedom (d.f.)
s(X,df=3)	GAM with smooth term with 3 d.f.
s(X,df=4)	GAM with smooth term with 4 d.f.
s(X,df=5)	GAM with smooth term with 5 d.f.

Table 1 The considered terms for a generalized additive model when selecting model using the k-leave-out cross-validation. Fit more predictors in the model, all combinations of their corresponding terms of various complexity should be tried when looking for the "best" model specification.

Captions:

Figure 1 Dependency of relative frequency of pollen type *Gramineae* in a pollen sample on the amount of rainfall (based on unpublished data of John Birks). The fitted second-order polynomial generalized linear model is displayed with the thin solid line. The fitted generalized additive models of the same complexity (with two degrees of freedom) is displayed by the thick solid line.


0.20
0.15
0.10
0.05
0.0
P(Gramineae)
500
1500
2500
Precipitation [mm]

Chapter 3

Small-scale heterogeneity in plant cover: can be explained?

Petr Šmilauer: Small-scale heterogeneity in plant cover: can be explained?

Abstract

The spatial heterogeneity of the soil nutrients available for plants was measured in a grid of soil samples, each taken from the area of 16 cm^2. The plants rooting in individual samples were recorded on semi-quantitative scale. A significant relation between availability of water-soluble phosphorus and plant cover composition was found and explored by statistical methods. While the causality of the dependencies found in this observational study cannot be determined, the clear relation of patches with relatively higher nutrients content to the dominance of *Poa angustifolia*[1] is apparent.

Introduction

As part of my research on the role of root systems in the maintainance of spatial heterogeneity and biodiversity in grassland communities, I have started a field experiment exploring the influence of endomycorrhizal symbiosis on the plant community spatial structure and its interaction with the nutrients availability. A supplementary observational study was performed to assess the natural levels of nutrients availability at the study site, their spatial distribution and its possible relations with the small-scale heterogeneity in plant cover. This supplementary study revealed some interesting results that are presented in this paper.

The study was expected to provide some support data for the a priory hypothesis that part of the high spatial heterogeneity in the small-scale distribution of plant populations in the studied grassland can be explained by the corresponding heterogeneity (patchiness) in the soil nutrient resources. Different species have different ability to explore the soil volume for the nutrients and in many types of plant communities (including temperate grasslands) this differential ability might be substantially influenced by the extent of endomycorrhizal symbiosis the individual species exhibit (Koide 1991). Nevertheless, the spatial heterogeneity in the soil resource might enable plants with different strategies for soil-volume exploration and nutrients acquistion to coexist (e.g. Fitter 1982).

It is very difficult, if not impossible, to say to which extent the position of individual plants is influenced by the soil nutrients heterogeneity and in what extent the causality goes in the opposite direction (i.e. the dynamics of aboveground and belowground biomass creating the patchiness in soil nutrient resource). Despite of that, there was some interest in this problem (e.g. Jackson et Caldwell 1993, Hook et al 1991). From the point of view of my study, the studies reported by those authors tackled the problem on too gross spatial scale (e.g. the scale of 12.5 cm in the study of Jackson et Caldwell) and in the type of vegetation having more simplistic spatial structure than is that of the type of grassland studied here.

Study site and methods

The study site is located near the village Zvikov, approx. 10 km E from the Ceske Budejovice, Czech Republic. It is an oligotrophic meadow, being managed in a traditional way at least since 1827 (Stabilni katastr 1827). Even in the 1827, the grassland is reported as nutrient poor, stony meadow, being cut at most once a year. The meadow is positioned at the slope of a small brook,

[1]Plant species nomenclature according Rothmaler 1976

formed by its sandy banks. At present, the upper margin of the meadow is somewhat influenced by the wash-out of the nutrients from the adjacent arable field (with larger dominance of *Alopecurus pratensis, Deschampsia cespitosa, Dactylis glomerata*), but the central part (where this study was done) is still species rich with the most important species being *Poa angustifolia, Festuca rubra, Avenochloa pubescens, Nardus stricta, Luzula campestris, Plantago lanceolata, Achillea millefolium, Galium boreale* and many others. The low nutrients availability on this site is very important for studies with the focus on the influence of increased nutrients inputs on the vegetation.

To assess the small-scale heterogeneity in the soil nutrients availability a block of the soil was divided into 9 x 5 adjacent square subsamples, each with side of 4 cm. The soil was collected into depth of 6 cm, corresponding to the zone where at least 90% of the root biomass is situated. It was felt that some species might occupy a rather separate niche characterized by their roots penetrating deeper into soil horizon (see also Fitter 1982), but separate analysis of lower soil layer was impossible due to funding constraints. The plants rooting in each of the 45 soil blocks were collected and the presence and semi-quantitative abundance was recorded using a very simple scale (0=absence, 1=one or few individuals, depending on the species' constitution, 2=large abundance of that species). The soil was sifted through the 2 mm sieve and analysed within few hours for the contents of nitrogen in the NH_4^+ form, the contents of anorganic nitrogen in the NO_3^- form and of the water-soluble phosphate were determined in the next two days. The nutrient contents was expressed as μg of N (or P) per g of dry sieved soil, as well as the nutrient contents per soil volume. The water content (expressed as weight percentage) was measured, as well.

To test for relationship between plant cover composition at the spatial scale of approximately 4 cm and the soil nutrients availability, the Redundancy analysis - RDA (a constrained type of linear ordination method, closely related to Principal components analysis - PCA, see ter Braak et Prentice 1988) was done with forward selection of the explanatory variables, based on the Monte Carlo permutation test (ter Braak, 1987). The PCA was also applied to the species data to summarize the trends of the co-occurrence of the species in the studied segment of vegetation. The analyses were done with the program CANOCO (ter Braak 1987) and ordination diagrams prepared with program CanoDraw (Šmilauer, 1992) and CanoPost (Šmilauer, unpublished).

The relationship between species occurrence and the one of the soil nutrients found significantly related with the species composition of the vegetation was studied by fitting a generalized linear model (McCullagh et Nelder, 1989) with proper model selection based on forward selection using the AIC statistics (Chambers et Hastie, 1992). For the simple abundance estimate, the model with expected Poisson distribution was used, but assuming (in calculating the AIC statistics) that the response variable might be in fact under-dispersed (McCullagh et Nelder, 1989).

Results

The concentrations of the two anorganic forms of nitrogen (the NH_4^+ form and the NO_3^- form) and of the water-soluble phosphate are displayed in Fig. 1, Fig. 2, and Fig. 3, respectively. The concentrations are given as μg of N or P per g of dry soil, as this form presented a slightly better relationship with the species composition data. The Fig. 4 displays the ratio between concentration of both form of anorganic nitrogen to the concentration of available phosphorus. Comparing these values to the data about concentration of these two elements in plant dry matter (generally N:P is approximately 10:1), it can be clearly seen that phosphorus is probably limiting

the biomass production more substantially. Generally, all the studied nutrients show very low levels.

Fig. 7 displays the main results of PCA, where composition of the plant cover was studied and the corresponding changes in the nutrients availability were subsequently projected into the ordination space. From that ordination diagram, it might be seen that there are two main groups of species forming the "matrix" of the community: one represents small-scale patches of *Poa angustifolia*, the other one small tussocks of *Luzula campestris* being accompanied by *Festuca rubra*. The third important grouping, with species like *Carex hirta, Plantago lanceolata, Anthoxanthum odoratum* is somewhat independent in its occurrence to the previous two (which are to large extent mutually exclusive) and corresponds to small-scale "gaps" between the patches dominated by the species of the one of the two above-mentioned grouppings. The reason for the independence apparent in the ordination plane of the first two axes of PCA lays probably in the fact that the mosaic of these patches with gaps is on a scale smaller than that used in this study (i.e. smaller than 4 cm). It might be seen from the projection of the variables describing the soil nutrients availability that higher nutrient contents are rather under the patches of the dominant species (*Poa angustifolia* or *Luzula campestris*). The spatial distribution (corresponding more or less to the dominance of *Poa angustifolia* vs. *Luzula campestris* and *Festuca rubra*) of the scores of the first PCA axis is displayed in Fig. 6. The black regions (corresponding to the bars going downwards from the zero plane) correspond roughly to the subplots where *Poa angustifolia* dominated. This might be seen also from the Fig. 5, where gray shading correspond to the abundance value of 2 for *Poa angustifolia*, the black color to the value 1 and white areas mark the absence of that species.

The forward selection of explanatory variables using the RDA method revealed that there is a significant relation between availability of water-soluble phosphorus and plant cover composition ($P = 0.050$, $N=1999$). The variable with the second best explanatory power was concetration of nitrate form of nitrogen, but it was not significant ($P = 0.087$, $N=1999$). This might be caused by the lack of relation as well as by the low power of the test (due to low number of replicates), of course. Fig. 8 presents the ordination space of RDA with the PO_4^- concentrations as the only explanatory variable. Consequently, only the projections of the species' arrows on the first (horizontal) ordination axis reveal something about the relations of the occurrence of those species with the phosphorus availability. Only species where significant relation was found are displayed. In the results of fitting generalized linear model to the dependency of species abundance on the phosphorus availability, only *Poa angustifolia* had significant increase in its abundance with increased availability of phosphorus in the soil, the other species having significant change in their "abundance" along the phosphorus availability gradient (*Achillea millefolium, Anthoxanthum odoratum, Carex hirta, Festuca rubra, Veronica chamaedrys*) show varying extent of decrease in their abundance.

Discussion

This study is presented in hope to provide interesting insight into this rather unexplored field of correlation between low-scale nutrient patchiness and the patchiness in the plant cover composition. I am well aware of the many shortcoming, that such study undoubtly has, the most significant of them being:

a) the distribution of individual plants on such a small scale is inevitably influenced by their life history related of the history of that particular place, where the sample was taken. The replicate blocks for the study would be extremely needed.

b) the studied nutrient resources have distinct patterns of their intra-seasonal variability which this study fully ignored. The study was done in time where large nutrient input (namely of

phosphorus) is needed, as most of the plants are in the phenophase of flowering or shortly before that phase. Another important part of the season is the period of rapid accumulation of aboveground biomass, starting shortly after snow thawing, and the story that would be told by such study done in that part of season can be much different.
c) no explanation of the causality of the observed patterns or relations could be attempted, a manipulative study would be clearly needed
d) even the expected relationship between the point of rooting of a plant and the nutrient availability of the surrounding block of soil is limited, as many of the species are clonal plants, with expected transfer of nutrients between the ramets via the rhizomes.

Acknowledgments

Many thanks to my wife Marie for help with the data collection, to Ota Rauch for the soil analyses and to the Grant Agency of Czech Republic providing the financial support (GACR 204/96/0522).

References

J. M. Chambers, T. J. Hastie [eds.] (1992): Statistical Models in S. - Wadsworth & Brooks, Pacific Grove, California. 608 pp.

A. H. Fitter (1982): Influence of soil heterogeneity on the coexistence of grassland species. - J. Ecology, 70: 139 - 148.

P. B. Hook, I. C. Burke, W. K. Lauenroth (1991): Heterogeneity of soil and plant N and C associated with individual plants and openings in North American short-grass steppe. - Plant and Soil, 138: 247 - 256.

R. B. Jackson, M. M. Caldwell (1993): Geostatistical patterns of soil heterogeneity around individual perennial plants. - J. Ecology, 81: 683 - 692

R. T. Koide (1991): Nutrient supply, nutrient demand and plant response to mycorrhizal infection. Tansley Review 29. - New Phytologist, 117: 365 - 386

P. McCullagh, J. A. Nelder (1989): Generalized Linear Models. - 2nd Edition. Chapman and Hall, London, 511 pp.

W. Rothmaler (1976): Exkursionsflora für die Gebiete der DDR und der BRD. Kritischer Band 4. - Volk und Wissen Volkseigener Verlag. Berlin. 811 pp.

Stabilni Katastr (1827): Kreis Budweis, Zvikov. Nummer 428.

P. Šmilauer (1992): CanoDraw 3.0 User's Guide. Microcomputer Power, Ithaca, USA, 118 pp.

C. J. F. ter Braak (1987): CANOCO - a FORTRAN program for Canonical Community Ordination by [Partial] [Detrended] [Canonical] Correspondence Analysis, Principal Components Analysis and Redundancy Analysis (Version 2.1). - Agricultural Mathematics Group, Wageningen.

C. J. F. ter Braak, I. C. Prentice (1988): A theory of gradient analysis. - Advances in Ecological Research, 18: 271 - 317

Figure captions

Tab. 1 : Significant relations between abundances of the species and the water-soluble phosphorus availability in the soil. All the models (fitted with Generalized linear model) were linear and only those significantly improving the null model are presented in the table.

Fig. 1 : Spatial arrangements of the anorganic nitrogen availability, in the NH_4^+ form. The concentration is given as µg of N per gram of dry soil weight. Each block is a square with a side of 4 cm.

Fig. 2 : Spatial arrangements of the anorganic nitrogen availability, in the NO_3^- form. The concentration is given as µg of N per gram of dry soil weight. Each block is a square with a side of 4 cm.

Fig. 3 : Spatial arrangements of the anorganic phosphorus availability, in the water-soluble PO_4^{3-} form. The concentration is given as µg of P per gram of dry soil weight. Each block is a square with a side of 4 cm.

Fig. 4 : Spatial arrangements of the ratio of available anorganic nitrogen (both form) to the available water-soluble phosphorus. The ratio uses the concentration given as µg of N or P per gram of dry soil weight. Each block is a square with a side of 4 cm.

Fig. 5 : Spatial distribution of the *Poa angustifolia* in the analysed segment of the grassland vegetation. Each block is a square with a side of 4 cm. The gray area corresponds to large abundance of *P. a.*, the black area to an intermediate abundance, and the white area to the absence of *P. a.*

Fig. 6 : Spatial arrangements of the scores of subsamples on the first ordination axis of PCA. The scores are based on the species composition at each of the subsamples. Each block is a square with a side of 4 cm. For the interpretation of the scores, see the ordination diagram in Fig. 7.

Fig. 7 : The ordination diagram displaying the first two ordination axes of PCA. These two axes account together for 35% of the variability in the vegetation composition data. The expected abundance (for the species) or expected concentration of the nutrients (for the explanatory variables) increases linearly through the ordination plane in the direction indicated by the corresponding arrow.

Fig. 8 : The ordination diagram displaying the first two ordination axes of RDA. The first ordination axis (horizontal) is constrained, corresponding to the increase in water-soluble phosphorus availability (from left to right). It accounts for 4% of the variability in the vegetation composition data. The second axis is un-constrained and account for 18% of the variability in the composition data. The expected abundance (for the species) increases linearly through the ordination plane in the direction indicated by the corresponding arrow.

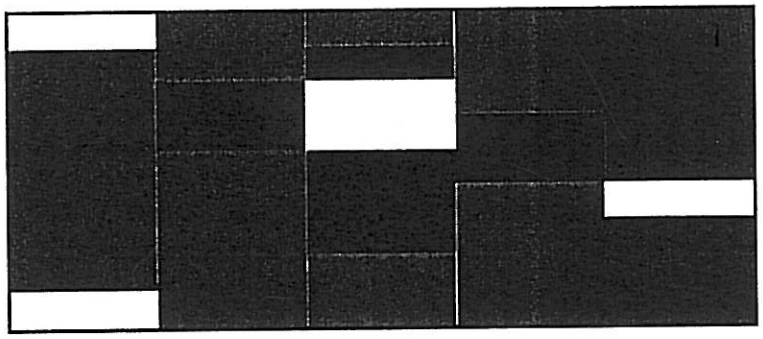


Figure 5

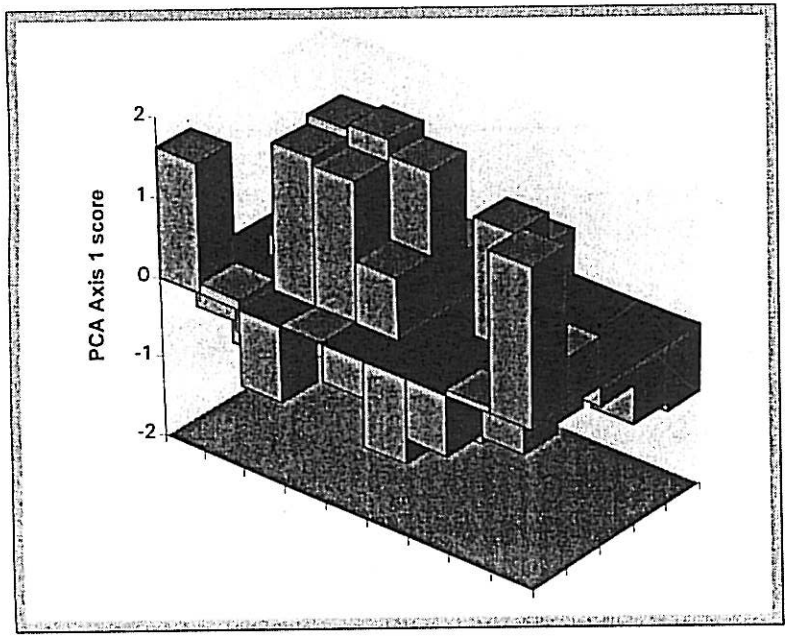
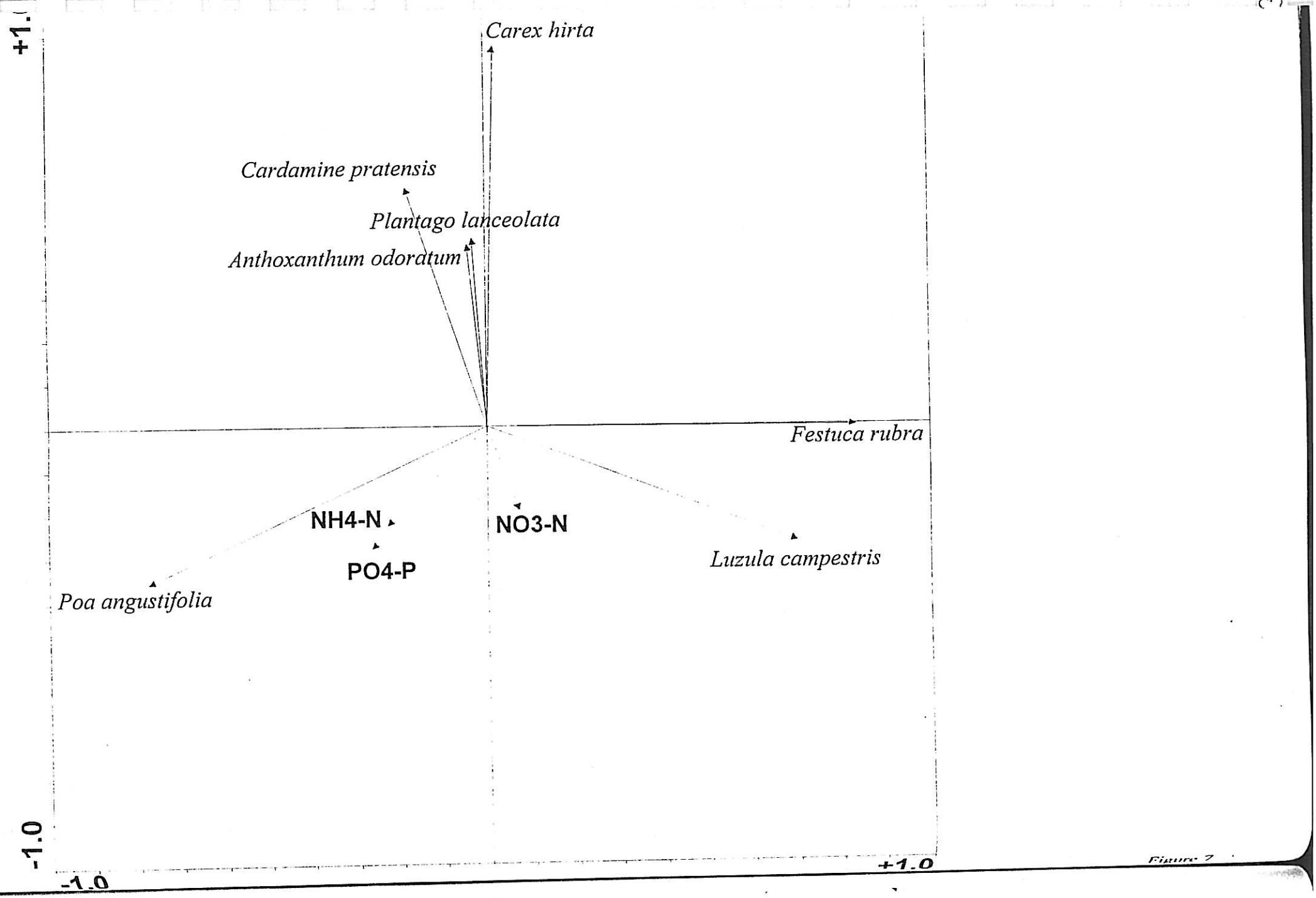


Figure 6


+1.
Carex hirta
Cardamine pratensis
Plantago lanceolata
Anthoxanthum odoratum
Festuca rubra
NH4-N
NO3-N
PO4-P
Luzula campestris
Poa angustifolia
-1.0
-1.0
+1.0

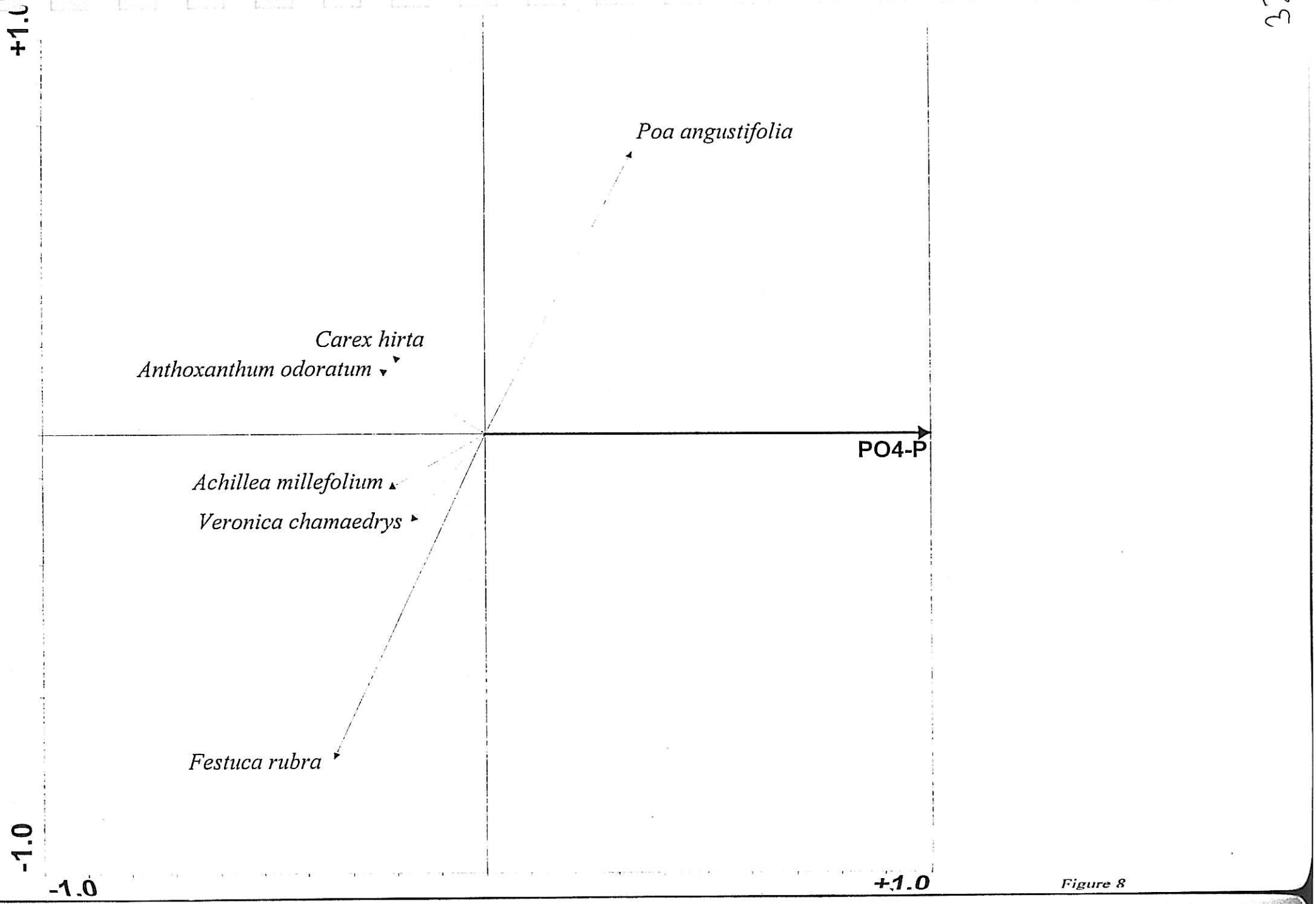

+1.0
Poa angustifolia
Carex hirta
Anthoxanthum odoratum
PO4-P
Achillea millefolium
Veronica chamaedrys
Festuca rubra
-1.0
-1.0
+1.0

Figure 8

Chapter 4

Hydrology and water table dynamics

Co-authored with K. Prach and O. Rauch

In: K. Prach, J. Jenik, A. Large [eds]: Floodplain ecology and management. The Luznice River, Trebon Biosphere Reserve, Central Europe. - SPB Academic Publishers, Amsterdam

3.3. HYDROLOGY AND WATER TABLE DYNAMICS

Petr Šmilauer, Karel Prach and Ota Rauch

The Lužnice River is characteristic of those rivers where the average annual discharge has its peak in spring because of melting snow in the headwater region. In summer time a high, but short-lasting, discharge often occurs after heavy rains (usually thunderstorms). The lowest average discharges typically occur during the autumn and early winter between the months of September and January. Fluctuations of water discharge are ameliorated by the influence of water contained in deeper horizons of the permeable sediments in the Lužnice catchment (Chábera 1985).

In the Upper Lužnice River, there are two permanent measurement points equipped by limnigraphs operated by the Hydrometeorological Institute, located just outside the main research area. One is located 143 stream kilometre upstream from the confluence with the Vltava at a bridge crossing the Lužnice near the village of Nová Ves, the other, 117 stream kilometres from the confluence is located near Chlum u Třeboně. Using standard hydrological coefficients, it was possible to recalculate all long-term hydrological parameters measured at these two sites to provide a surrogate hydrological profile for the centre of the research area. The centre of the research sector was located at a bridge at the village of Halámky, 137 stream kilometres upstream of the confluence with the Vltava.

The following parameters characterize the river at this point:

Long-term average discharge: 4.85 $m^3.s^{-1}$
maximum 100 yr discharge 129
maximum 50 yr discharge 112
maximum 10 yr discharge 76
Average monthly discharge: Jan 3.73; Feb 5.29; Mar 8.39; Apr 7.61; May 4.95; Jun 4.27; Jul 5.56; Aug 4.66; Sep 3.30; Oct 4.12; Nov 3.44; Dec 2.86

The long-term average precipitation in the respective part of the catchment was calculated as 753 mm per annum occurring over the 645 km^2 section of the 4225 km^2 Lužnice catchment which lies upstream of Halámky (Chábera & Šabatová 1965; Krásný 1980; Homolka 1984; unpublished data of the Hydrometeorological Institute, České Budějovice). The downstream movement of the water in the river channel, together with related horizontal and vertical movements of ground water in the hyporheic zone surrounding the river, are the key factors governing the many hydrological processes occurring in the river-floodplain complex. In this section, both horizontal and vertical ground water movements are discussed, along with flow variability for the reach in question.

The horizontal movement of underground water in the floodplain sediments was measured by marking water samples by Br nuclide and detecting its spread over time. The average rate of water movement was estimated to be several cm per day at the boundary with the floodplain terrace, and approximately 10^{-1} m. day^{-1} in the middle of the floodplain sector examined, equidistant from the terrace and the river channel itself. Although not directly measured, this rate of flow would be expected, using these

Floodplain Ecology and Management, pp. 000–000
ed. by K. Prach, J. Jeník and A.R.G. Large

spb/prach/3-3/conv/040696

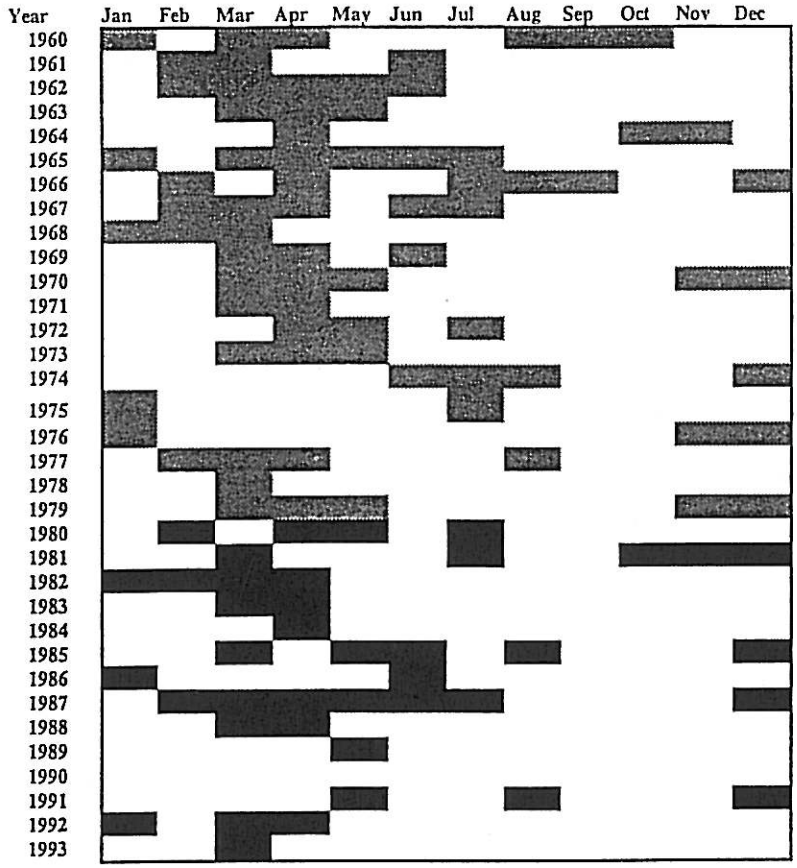


Fig. 3.16. Occurrence of flooding events in individual months during the period 1960-1993. The occurrence is derived from monthly averages of water discharge for the years 1960 to 1979 (using the threshold value of 7 $m^{-3}s^{-1}$). The information for the period 1980-1993 is based on daily average data. Here, the threshold used is 10 $m^{-3}s^{-1}$, which is much more reliable indicator of flood in the area. The measurements were conducted on the Pilař profile at 117 stream kilometre by Hydrometeorological Institute.

dynamics. Firstly, the rises and falls of discharge volume during the "spike events" proceed at different paces. Fig. 3.18 displays the large differences among the daily averages in the period of 1989-1993. From this, it can be seen that the discharge volume rises much more swiftly than it decreases afterwards. Secondly, the speed of the daily change of the discharge volume increases almost exponentially with the absolute amount of water involved (Fig. 3.19). Finally, the localized precipitation events in the river discharge area, often unpredictable from the meteorological situation at the site under consideration, could result in sharp volume spikes (even resulting in short-duration floods) in the summer months and the occurrence (in a less pronounced form) in other parts of year of sudden step-increments in the discharge volume.

An attempt was made to simulate daily river discharge volume dynamics for the sector under investigtion. This, it was hoped would feed directly into the modelling of the primary production of the plant communities in the river floodplain (see Chapter 6.2). In doing so, efforts were made to include all the peculiarities of the flow dynam-

Table 3.4. The probability of one or more floods occurring in individual months of year, based on the data from 1980-1993 (left hand column) and 1960-1993 (right hand column). March and April, the two months with a probability of flooding higher than 0.5 are highlighted with a grey background.

	1980-93	1960-93
Jan	0.2	0.2
Feb	0.2	0.3
Mar	0.6	0.6
Apr	0.5	0.6
May	0.4	0.4
Jun	0.2	0.3
Jul	0.2	0.3
Aug	0.1	0.2
Sep	0.0	0.1
Oct	0.1	0.1
Nov	0.1	0.1
Dec	0.3	0.3

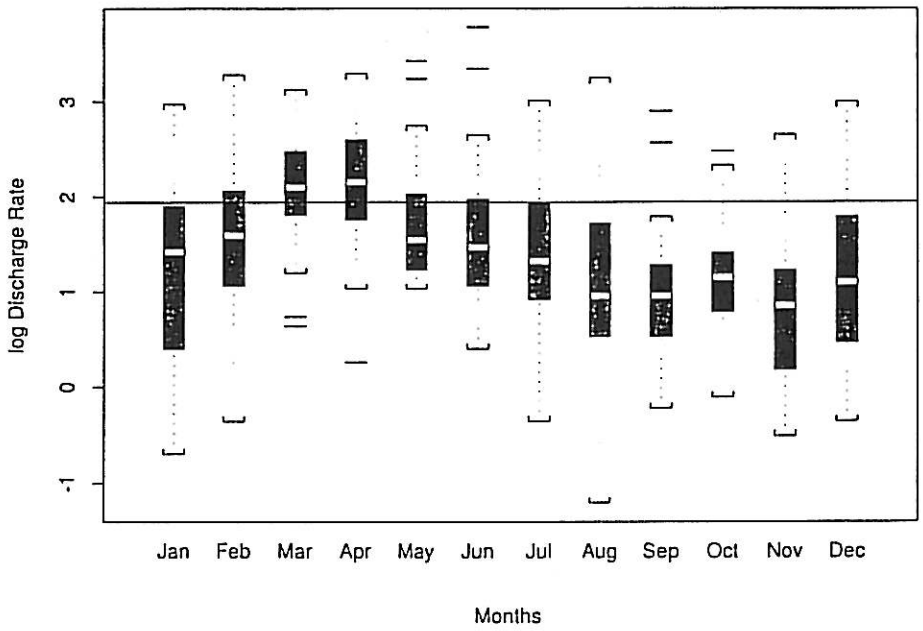


Fig. 3.17. Distribution of monthly averages of discharge rate values in individual months based on a loge distribution. The reference line corresponds to the threshold value for flooding, as used in the upper part of Fig. 3.16, namely loge 7. The distribution of values is shown by the box-and-whisker diagrams with the usual meaning (after Tukey 1977).

ics outlined above. However, the primary limitation of this approach – not including the meteorological events occurring in the whole catchment area discharged by the river – makes the simulation results somewhat less reliable. In the simulation displayed in Fig. 3.20, the occurrence of local "spikes" or step-increments, as well as the pace of the discharge volume decrease, are based solely on local meteorological conditions. This gives to the simulated series an inevitably different pattern from what would be produced by integrating precipitation events and temperature over the much larger area of the river catchment.

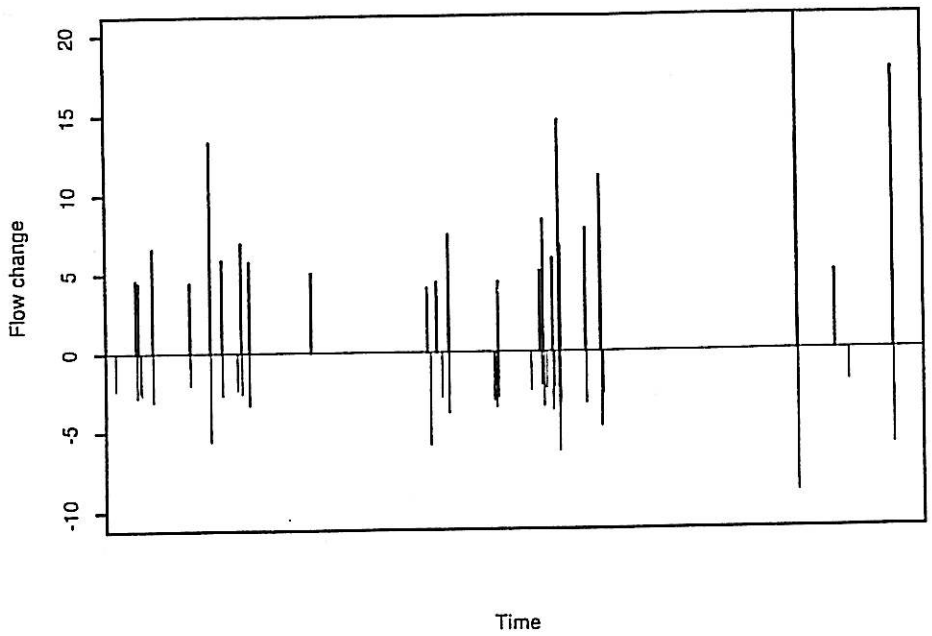


Fig. 3.18. First-order differences between daily means of river discharge volumes during the period 1989-1993. Only positive volume increase of 4 m^3s^{-1} or more and negative ones volume decreases of -2 m^3s^{-1} or less are displayed. The different thresholds were used so that the decrease events displayed approximately cover the same set of "spike" events represented by the increase events.

The simulation of the discharge volume is based on the simulation of the dynamics of the first-order differences between daily values. The increase of the volume comes from three principal sources:

– for the early spring, one flood event with a fixed probability of recurrence is scheduled for each year. If the event is scheduled for a given year, the actual start (depending intrinsically on the start of significant snowmelt in the headwater region) will be triggered by the threshold value of average daily air temperature. The value used (mean air temperature at 2 m averaged over last 5 days, exceeding 0 °C) was derived from the observed meteorological data.

– for all periods of the year, daily precipitation above a certain level triggers an increase in the river discharge volume. The simulation of discharge volume data divides the year into several distinct periods (corresponding roughly to the four seasons of the year), and the duration and extent of the precipitation-induced increases differ among these periods.

– at times of the year where discharge volumes are generally low and slowly decreasing (a feature typically seen in autumn months, but which often occurs also in early winter), random step-increments are scheduled with a fixed frequency, which patterns the influence of local precipitation events in the river catchment area.

In Fig. 3.20, the daily river discharge volume data for two years (1980-1981) is compared with the data for the first two years of one simulation run of the model. There is significant variability among individual years in the real data, so the similarity of the simulated data with the real data should not be compared in terms of the exact distribution of the peaks throughout the years or the absolute height of the peaks (which happen to be lower in the simulated run used, as compared with the "real-

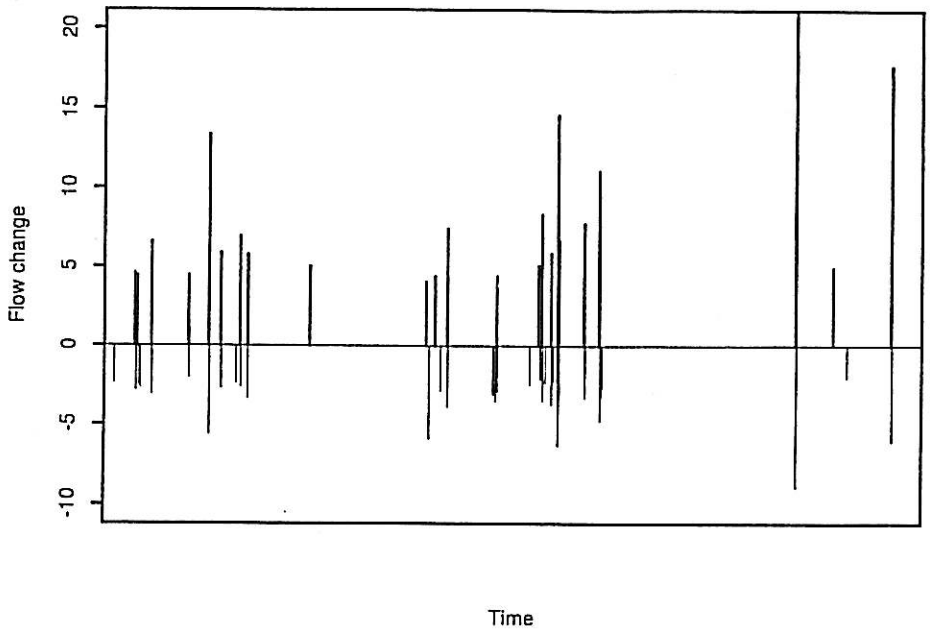


Fig. 3.18. First-order differences between daily means of river discharge volumes during the period 1989-1993. Only positive volume increase of 4 $m^{-3}s^{-1}$ or more and negative ones volume decreases of -2 $m^{-3}s^{-1}$ or less are displayed. The different thresholds were used so that the decrease events displayed approximately cover the same set of "spike" events represented by the increase events.

The simulation of the discharge volume is based on the simulation of the dynamics of the first-order differences between daily values. The increase of the volume comes from three principal sources:

– for the early spring, one flood event with a fixed probability of recurrence is scheduled for each year. If the event is scheduled for a given year, the actual start (depending intrinsically on the start of significant snowmelt in the headwater region) will be triggered by the threshold value of average daily air temperature. The value used (mean air temperature at 2 m averaged over last 5 days, exceeding 0 °C) was derived from the observed meteorological data.

– for all periods of the year, daily precipitation above a certain level triggers an increase in the river discharge volume. The simulation of discharge volume data divides the year into several distinct periods (corresponding roughly to the four seasons of the year), and the duration and extent of the precipitation-induced increases differ among these periods.

– at times of the year where discharge volumes are generally low and slowly decreasing (a feature typically seen in autumn months, but which often occurs also in early winter), random step-increments are scheduled with a fixed frequency, which patterns the influence of local precipitation events in the river catchment area.

In Fig. 3.20, the daily river discharge volume data for two years (1980-1981) is compared with the data for the first two years of one simulation run of the model. There is significant variability among individual years in the real data, so the similarity of the simulated data with the real data should not be compared in terms of the exact distribution of the peaks throughout the years or the absolute height of the peaks (which happen to be lower in the simulated run used, as compared with the "real-

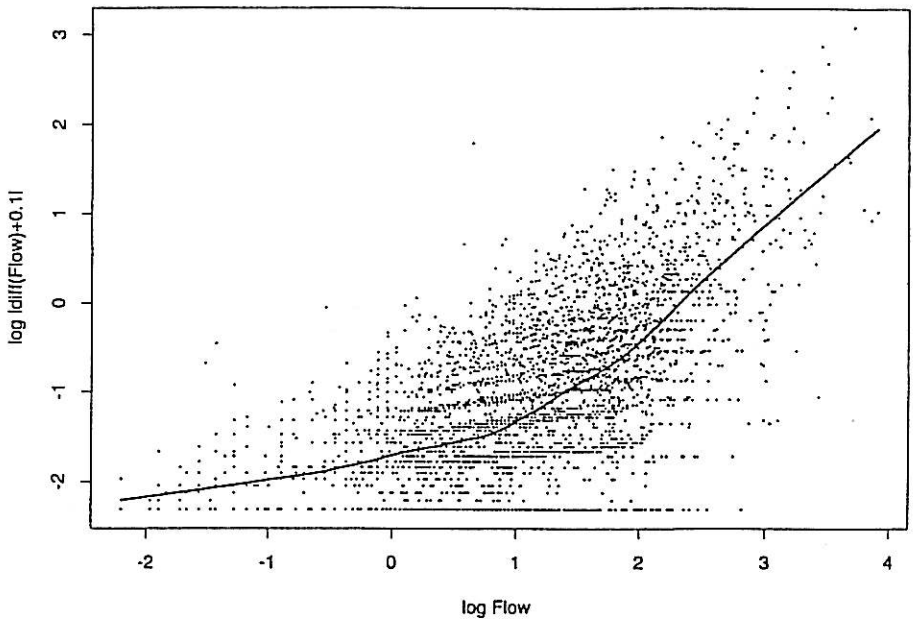


Fig. 3.19. The loge – transformed absolute values of differences between consecutive daily discharge average values (Y axis) plotted against the loge – transformed discharge volume values (X axis). The curve attempts to visualize the trend and is based on a loess-model with parameters (a=0.3, l=1) (after Cleveland 1993).

world" example). Rather, they should be compared in terms of the shape and general distributional properties of the "ups" and "downs" of the series. When this is done, the similarities of the simulation to the real data series becomes apparent.

Underground water table simulation

At the study site at Halámky, the dynamics of the underground water table was observed over a number of years at more or less regular intervals. The water table was observed over a cross-section through the river floodplain, starting near the river bed at one end of the cross-section and ending at the first river terrace (see Chapter 4.1.3). This cross-section spanned a distance of approximately 150 metres. An attempt was made to simulate daily positions of underground water table, based on data relating to river discharge volume and precipitation events – two factors presumed to influence the underground water table dynamics in general.

An important feature of the underground water table in the studied cross-section is its curved shape with a small depression in the middle of the cross-section and large rise from the bottom of the first river terrace. This feature, together with the non-linear dynamics of the water-table position during the season, led to the choice of a generalized additive models (GAM) approach (Hastie & Tibshirani 1990) to simulate the water table position throughout the year. To enhance suitability of graphical presentation without loss of accuracy or predictive power, a Gaussian distribution/identity link family of GAMs was selected. GAMs are less parametric that the more traditional linear models and are more suitable for graphical presentation.

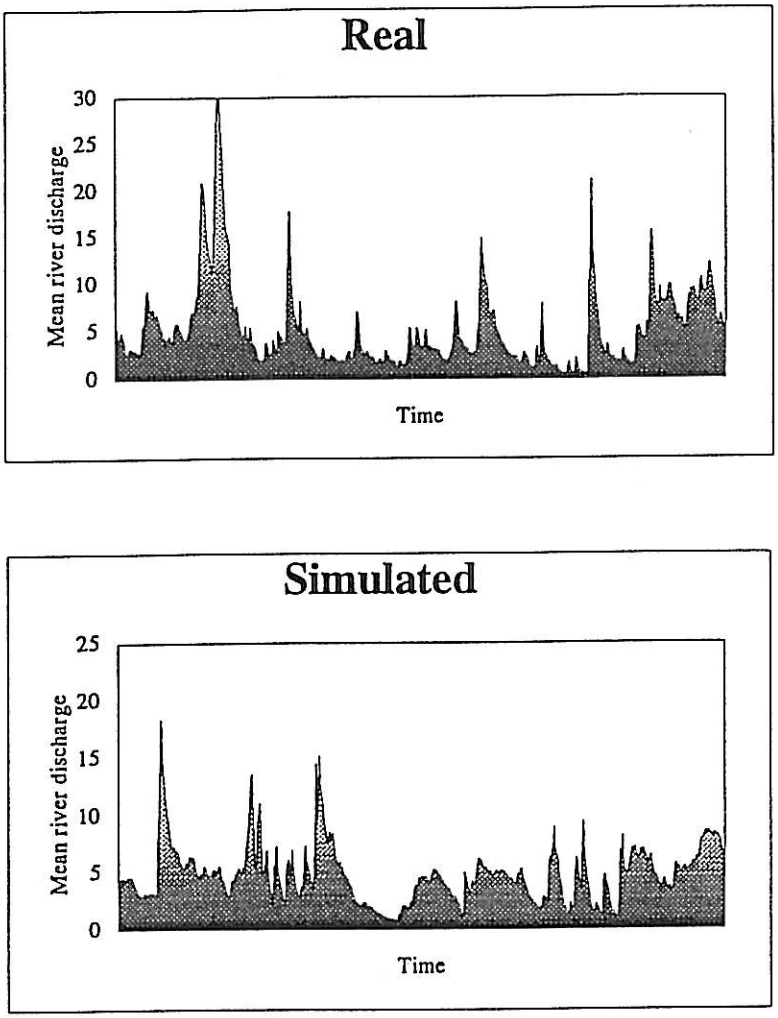


Fig. 3.20. The comparison of the real data (upper curve) – from years 1980-1981 with the simulated data (lower curve) describing the daily averages of river discharge volume (Y axes, $m^{-3}s^{-1}$). The horizontal scale represents days 1-731 in the time series.

Four predictor variables were considered during model selection, namely: actual daily average river discharge volume (Flow), the average discharge volume for the last ten days (Flow10), the amount of precipitation in the last five days, and location on the cross-section across the floodplain, expressed as distance from the river (Pos). The most parsimonious model was selected by a series of stepwise model changes, using the AIC criterion (Akaike 1973) to find the "best" model.

In the model, the amount of precipitation, in addition to the river discharge amount descriptors, turned out to have little predictive power despite the fact that there was a clear evidence of a greater role being played by the precipitation events on the part of the cross-section near the river terrace. The final fitted model explains about 86% of the variability in the values for the underground water table position. As expected, the major explanatory power was exercised by the variables Pos and Flow, while the

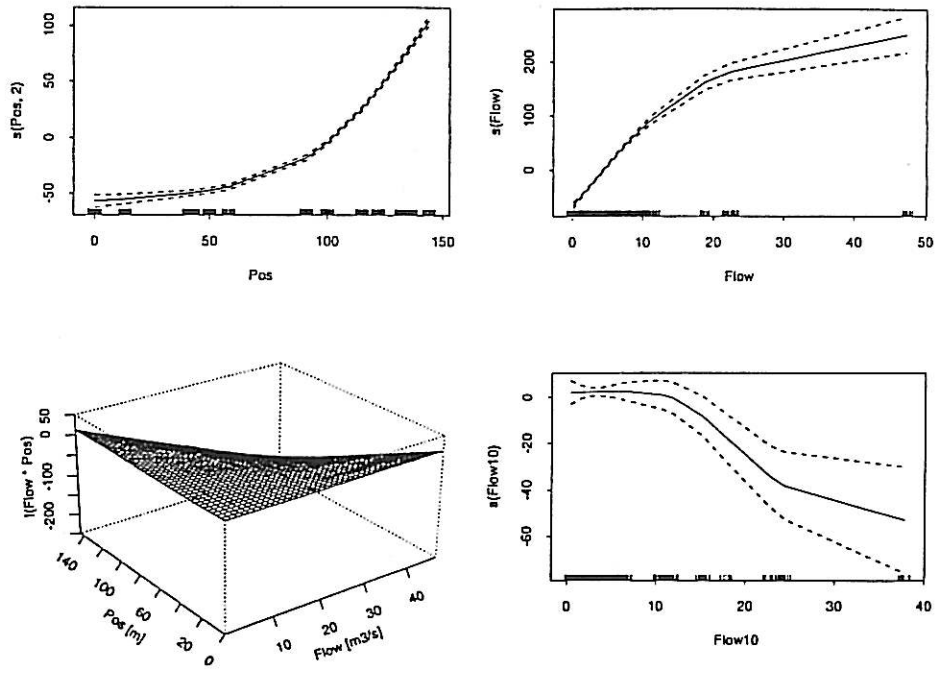


Fig. 3.21. Graphical presentation of the fitted generalized additive model to the underground water table data. The solid-line curves in the individual diagrams (except the bottom left one) display the fitted smooth terms together with their point-wise confidence regions (see Hastie & Tibshirani 1990 for more details). The diagram at the bottom left displays the two variables' interaction term, fitted by a means of a linear term (plane). The bars on the horizontal axis represent jittered positions of individual observations in the space of the predictor variables.

Flow10 and linear-form interaction between Flow and Pos (expressing different response of water table to the river flow at different points across the floodplain) also proved influential. Fig. 3.21 presents in graphical form the terms of the fitted model. The values of the predictors are displayed on the horizontal scale of the corresponding graph. Transformed values can then found on the vertical axis using the displayed curve. The contributions of individual terms are then added and corrected by a constant (99.4 in this case) to yield a predicted water table level relative to the arbitrary reference point of 455.0 m a.s.l. Resultant water table height values are expressed as relative altitude in centimetres.

Fig. 3.22 displays a simplified response surface fitted by the GAM model (with variable Flow10 kept at its average value). It is apparent that the slope of the water table has a much steeper gradient when the underground water is lower down in the soil layer. The decrease at higher stream discharge rates in the slope of the underground water table across the river floodplain provides evidence that re-charge of the underground water from the river itself takes place. Fig. 3.23 shows the position of the underground water table for a period of three years, based on the observed river flow data and fitted generalised additive model. A regular sequence of rises and falls in the height of the water table can clearly be seen.

As the relative altitude of the soil surface on the studied cross-section is rather well determined, we can combine this knowledge with the predicted underground water ta-

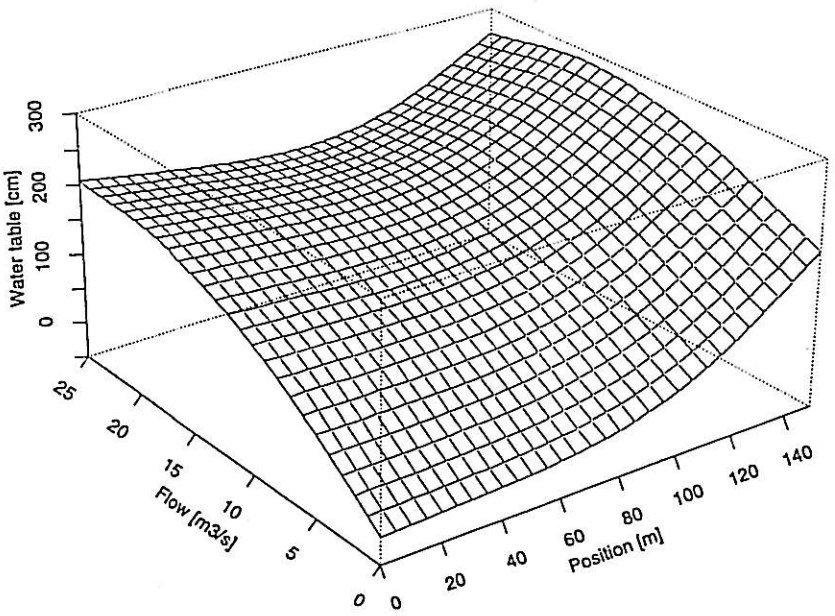


Fig. 3.22. Three-dimensional projection of the response surface fitted by means of the generalized additive model, presented in Fig. 3.21. Here, the values of the third explanatory variable (Flow10) are kept at its average. The calculation of the predicted water table level (on the vertical axis) also takes into account the contribution of the (Flow.Position) interaction term.

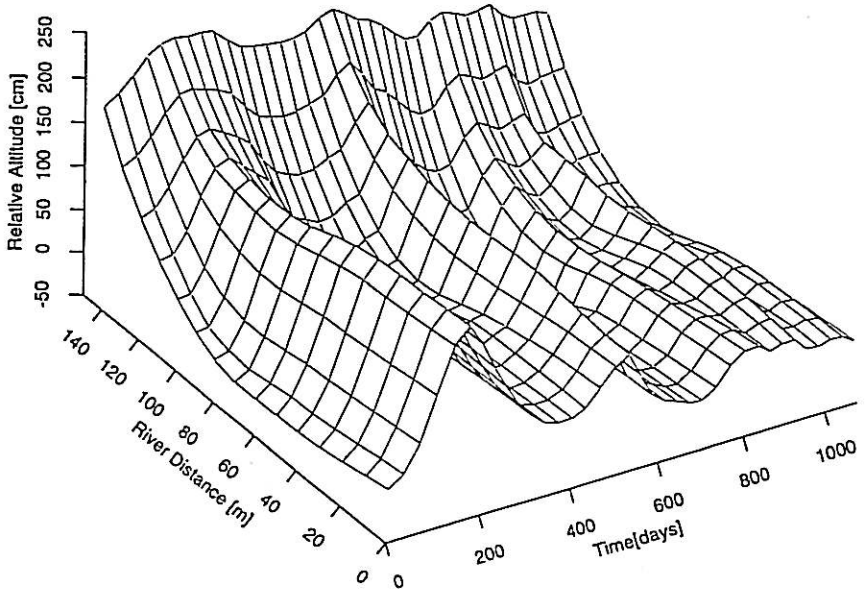


Fig. 3.23. Underground water table dynamics as predicted by the fitted generalized additive model (see Fig. 3.21 and Fig. 3.22), using the real river flow data for three consecutive years (autumn 1986-autumn 1989). The relative altitude on the vertical axis refers to the reference point at 455 m a.s.l., which is the approximate altitude of the river bed at the study site.

ble position to simulate the occurrence and duration of flooding events based on these data. Caution is needed however, when extrapolating these results into the broader realm of the floodplain. The sector of the floodplain subjected to detailed investigation is covered with well-drained, sandy river sediments. A significant part of the floodplain area (usually associated with remnant palaeochannel features) has different hydrological properties however, with less well drained soil horizons buffering the influence of river flow dynamics and increasing the relative importance role of precipitation in the maintenance of water table position. In these parts of the floodplain, much slower lowering of the water table in late summer / early autumn should be expected, along with significantly slower retreat of flood waters, once they rise above the level of the soil surface.

References

Akaike, H. 1973. Information Theory and an Extension of the Maximum Likelihood Principle. In: Petrov, B.N. & Csaki, F. (eds.), Second International Symposium on Information Theory . Akademia Kiado, Budapest. pp. 267-281.

Chábera, S. (ed.) 1985. Geography of south Bohemia (in Czech). Jihočeské nakladatelství, České Budějovice, pp. 270.

Chábera, S. & Šabatová, E. 1965. Survey of hydrography of Southern Bohemia (in Czech). Pedagogical Institute, České Budějovice, pp. 71. .

Cleveland, W.S. 1993. Visualizing data. Hobart Press, Summit, New Jersey.

Hastie, R. & Tibshirani, R. 1990. Generalized additive models. Chapman and Hall, London.

Homolka, M. 1984. Hydrogeological investigations in the southern part of the Třeboň Basin (in Czech). Report of the Institute of Geology, Prague.

Krásný, J. 1980. Hydrogeology of south Bohemian basins (in Czech). Sbornik geol. věd, Ser. Hig., Prague, 14: 7-81.

Tukey, J.W. 1977. Exploratory data analysis. Addison-Wesley, Reading, Massachusetts.

Chapter 5

Modelling primary production and nutrient dynamics

In: K. Prach, J. Jenik, A. Large [eds]: Floodplain ecology and management. The Luznice River, Trebon Biosphere Reserve, Central Europe. - SPB Academic Publishers, Amsterdam

6.2 MODELLING PRIMARY PRODUCTION AND NUTRIENT DYNAMICS

Petr Šmilauer

There have been numerous attempts to model primary production in various terrestrial ecosystems, with a range of different approaches being taken by different authors. An excellent example is provided by the extensive ELM model created as part of the International Biological Program (IBP) of UNESCO in the 1970s, by the U.S. Grassland Biome research team (Cole 1976; Innis 1978). The ELM model divides the prairie ecosystem into several independent blocks (for example primary producers, abiotic sections, mammalian consumers, decomposers etc.) and expresses the state of individual blocks by an array of state variables. Any changes to that state are described by mechanistically conceived processes, which are implemented chiefly as difference equations, with the time steps on the scale of days. The ELM model was successful in collating the efforts of many ecologists and helped to identify areas where further detailed investigation was much needed.

Many other published models concentrate on finer-scale ecophysiological approaches, particularly simulating the processes of photosynthesis, respiration and nutrients assimilation, translocation and conversion (e.g. Botkin 1969; Ondok & Gloser 1983). Several contributions have been published concerning detailed simulation of light and heat conditions in the vegetation canopy, the diurnal course of evapo-transpiration, etc. These are, however, concerned with a scale of resolution too fine for the purposes of our project.

Several studies have been done focusing on the productivity of grassland ecosystems in the Czech Republic during last 30 years, namely by the team of Prof. M. Rychovská (Rychnovská 1985, 1993). The results of the study, carried as a part of the Czechoslovak contribution to the IBP programs on the grassland ecosystems of the river floodplains in South Moravia, are of particular relevance to the Lužnice River project, especially in relation to modelling (Rychnovská 1972).

General aims of the model development

The approach taken by the authors of the ELM model seemed to be at a scale suitable for our purposes, but the required amount and quality of input data was apparently beyond the budget of the Lužnice River project. Yet, we have considered the development of a simulation model to be still desirable for a number of reasons:

(a) The simulation model brings together formulations for a number of research hypotheses, many of them supported by the collected data. The phenomena covered by the individual hypotheses are often inter-related and the model helps also to express and clarify their relationships to each other.

(b) The model could potentially serve as a research tool, allowing us to see the consequences of changing our views of the processes running in the ecosystem. A well-designed simulation model allows for easy integration of new knowledge about underlying ecosystem processes, so that the knowledge can be immediately compared with already existing hypotheses.

Floodplain Ecology and Management, pp. 000–000
ed. by K. Prach, J. Jeník and A.R.G. Large

(c) During the research project management, the model-development requirements can provide supporting information for prioritizing individual research tasks (this however does not always work perfectly, mainly due to budget limitations).

(d) The model might point to the most important bottle-necks in our ability to describe the processes running in the ecosystem. While it is not always managable to act on these during the project, this certainly helps in designing future projects.

Requirements for the modelling approach

Beside the inability to investigate and quantify so many various facets of ecosystem processes influencing solar energy fixation, uptake of nutrients from the soil system, the fate of nutrients during litter decomposition etc., we wished to incorporate several novel aspects into our modelling approach to make it more useful and appealing to the ecologist:

(a) The model has to be easily maintainable. The general-purpose programming languages are too much technical and the code they are written in often unreadable for non-specialists, while the programming systems specialized for simulation of ecosystems are usually restricted in their abilities by the modelling philosophy of their authors and their availability is generally limited.

(b) The model should be, at least as far as the parameters are concerned, easily modifiable for individual model run by individual users.

(c) A major problem with traditional simulation models is their strictly quantitative nature. At its extreme, the prevailing simulation approaches make it possible to incorporate scientific knowledge into the model only as long as it can be re-expressed by means of differential equations, with all the relevant parameters precisely quantified. While it is still possible to put heuristic, procedural-knowledge such as "If the rainfall amount in the late autumn is very high, then a large part of standing dead biomass moves to the litter" into a surrogate form of differential equations, using a combination of indicator variables and threshold values, that way of implementation is totally counter-intuitive for the ecologist and only increases his/her scepticism towards the simulation model. So, an easy incorporation of heuristic rules into the model was much needed.

(d) Even if we can use the IF...THEN... rules to express some of the important knowledge ecologists have about the subject, there still remains the difficulty when dealing with sometimes vague or uncertain terms contained in the hypotheses. In the example sentence from the previous paragraph, the terms like "...rainfall amount ... is very high..." or "...a large part of ... biomass..." are difficult to change into rules with an exact threshold value for the amount of precipitation needed to move, say 0.73% of the standing dead biomass to the litter compartment. While it is ultimately necessary to quantify that process if it has to enter the simulation model, it is probably better to keep the knowledge in its original, semi-quantitative form as long as possible. This is where the approach of linguistic variables, as used in the theory of fuzzy sets (Zadeh 1965, 1983), can be applied.

(e) Another problem situation, which occurs when we ask ecologists to explicitly formulate their knowledge, is that when the ecologist is not fully convinced about the applicability of the knowledge. This might be a result of several alternative expectations which differ subtly in their assumptions, but in the given context is not possible to discriminate among them. A similar problem situation might result from the conflicting opinion of different experts. This is a problem widely recognized in the field of

knowledge acquisition for building expert systems and the traditional way of approaching it is to use some form of the truth maintenance system.

(f) An important specific requirement for our simulation model is that it should be, in its global architecture, applicable to the all three dominant vegetation types found on the floodplain of the Lužnice river where biomass dynamics were studied. While the model is currently developed sufficiently only for the *Alopecurus* type, its architecture is easily applicable to the *Urtica*- and *Phalaris*-dominated stands (see Chapter 6.1).

(g) An important role in the modelled ecosystems is played by sudden events – mowing and flooding – which might or might not be scheduled for a particular time or might (in case of the flood) happen several times in a year. These features should be incorporated into the model.

Bearing in mind all the considerations outlined above, a prototype form of simulation model was developed for the Lužnice floodplain. Its structure is somewhat heterogeneous, as the desire to use the appropriate method for every sub-problem dominated over the criterion of an unified solution. Part of the system was developed in procedural language (programming language C), the major part was developed using an expert system development language called CLIPS (Giarratano & Riley 1994; CLIPS 1993). To enable the use of fuzzy logic based reasoning as well as the use of truth maintenance via certainty factors (Giarratano & Riley 1994), the modification of the base CLIPS system was used, called Fuzzy CLIPS [version 6.02] (Orchard 1994). The CLIPS language is a language based on the mix of three programming principles; (i) rule-based programming used in the traditional expert systems, (ii) object-oriented programming and (iii) procedural programming. It is portable among various platforms and, in fact, freely distributable.

The simulation system developed does not use the object-oriented extensions of CLIPS and it is mainly oriented towards the procedural programming with much stronger control of sequential operation than it is usual for the traditional expert systems. Despite the rather strong sequential control over the system execution (particularly for the sequential switching between individual modules – see below for their list) exercised via several mechanisms provided by CLIPS (system of salience levels, manipulation of the focus stack and use of control facts – see Giarratano & Riley 1994 for the details), the most of the code is written using inferential rules. The CLIPS language syntax is similar to the LISP programming language (or, more closely, to the OPS5 ES toolkit) syntax and a rather simplified example of the rules used in our simulation model looks like the following:

```
(01)  (defrule BIOM::ALOP_EAMRgr
(02)  "RGR-s for EarlyPostMow period"
(03)  (GrowthPhase (phase EarlyPostMow))
(04)  (not (rgr-set))
(05)  (Biomass (SystemName Alopecurus))      ;//specific for Alopecurus stnd.
(06)  ?rD <-(rgr-Dom ?)
(07)  (SinceChange ?sc)
(08)  =>
(09)  (retract ?rD)
(10)  (assert (rgr-Dom (randVal (- 0.0500 (* ?sc 0.000600))  0.05)))
(11)  )
```

The numbers in the parentheses on the beginning of each line are not actually part of the code. They were added only to facilitate discussion of the demonstrated code.

The code in lines (01) and (02) is part of the CLIPS-specific syntax for defining rules, giving the name to the rule and an optional comment in quotes (" "). The following lines until line (08), represent so called antecedent of the rule, which presents fact patterns (either specific facts which must exist before the rule is activated, or conditions the existing fact must fulfil). Line (03) specifies that the current phase of the biomass growth must be the one named EarlyPostMow (see section about biomass simulation below). The control-fact pattern specified in line (04) prevents this rule from activation ("firing" in the expert-system terminology) if the values that this rule has to produce were already set (in that case, the fact rgr-set would already exist). The fact pattern in line (05) assures that this rule applies only to one type of the vegetation, namely the mown meadows with dominating *Alopecurus pratensis* (as the comment at the end of the line states). This is needed for this rule, as the coefficients defined by this rule are specific for particular types of vegetation. The original rule defined totally six coefficients of the relative growth rate (RGRs), but all except that for the live biomass of the dominant species were omitted for simplicity. The pattern on the line (06) binds the existing value for that coefficient (rgr-Dom) so that it can be later removed from the knowledge base (in line (09)). At line (07), the number of days passed since the current growth phase began is acquired into the variable ?sc, so that it can be used to calculate the new value of the RGR coefficient (which is done in line (10)). The lines (09) and (10), separated from the previous one by the => symbol, form the consequent of the rule, whose actions are performed (removing the old fact and creating a new one with the calculated value, in our case) only if all the conditions given in the antecedent of the rule are satisfied.

The presented example could seem to be somewhat confusing for the first time reader, but it is generally accepted, that the syntax used by the CLIPS system is easy enough to learn for most people. The interested reader is again referred to Giarratano & Riley (1994).

Data used for simulation model development

The model is primarily concerned with the simulation of the main features of the seasonal dynamics of the three modelled vegetation types (see Chapter 6.1) with consideration given to the role of the floods (Chapter 3.3) and of the mowing regime (in the case of the *Alopecurus*-dominated stands). For the abiotic parts of the model, the source data consisted of the daily meteorological observations from the nearby meteorological station (Chapter 3.2), the data about the underground water levels across the river floodplain, and the daily average river discharge volume data (see Chapters 3.2, 3.3 and 4.1.3 for details). To obtain a detailed information about the pattern of changes in the biomass production, a detailed sampling campaign was run during two seasons – 1991 and 1992 (Chapter 6.1). The following biomass categories were considered in the model:

1. The living aboveground biomass of the dominant species;
2. The standing dead aboveground biomass of the dominant species ("standing dead" being defined as all the dead biomass not lying on the top of the soil);
3. The living aboveground biomass of the other species;
4. The standing dead biomass of the other species;
5. The litter biomass, representing all the decomposing biomass lying on the top of

the soil and not attached to any living parts of plants;

6. The living root biomass;
7. Contents of C and N in the above mentioned categories, changing during a year;
8. The rate of litter decomposition.

The exact methods of collecting biomass samples, and all respective figures are described and presented in Chapter 6.1.

Meteorological variables subsystem

The simulation of the seasonal patterns of biomass production needs certain input from the weather simulator. The module WEATHER written in the CLIPS language reads the daily predictions of the meteorological variables values from an external text file. Currently, only the precipitation amount and average air temperature at 2 m are directly utilised in the system, beside the average precipitation and air temperature for the last five days, provided from another input file. These input values are then used at the beginning of each simulated day to update values of two fuzzy ("linguistic") variables called FuzzyTemp and FuzzyPrec which are defined with several fuzzy quantifiers like low, high, low-for-spring etc. This allows to define rules requiring conditions such as "IF the precipitation is not too high...", which is translated into CLIPS syntax in the form:

```
...
(FuzzyPrec NOT [ very high ])
...
```

The input file with the weather simulation data is read not only by the main simulation model, but by the river flow/underground water table dynamics simulator as well. The temperature and precipitation data are used to simulate the daily average river discharge values. The meteorological simulation is done in a way similar to that used by Randell & Gyllenberg (1972), as it structures the generation of the values of meteorological characterics according to the so called meteorological situations (or air-mass types). In the Czech Republic, a classification of the current meteorological situation into one of the 28 distinct types has been done on a daily basis since 1946. The values published in Starý (1989) together with tables listing the sequences of meteorological situations for the years 1979-1989 were used to create a generator of meteorological situations. The generator uses different parameters for the three parts of the year, distinguished by the meteorological sub-model. The division of the year was based on the results of a multivariate analysis (principal components analysis) of the monthly averages of the meteorological variables. For each part of the year, the sequence of the meteorological situations is generated using transition matrices simulating first-order Markov process. For each new meteorological situation, its length (in days) is generated based on the known frequency distribution of the lengths of the given situation type over the past 30 years. The values of the meteorological variables are generated from the table of distributional parameters (mean values and standard deviation) of the variable for a particular meteorological situation in the particular part of the year.

River flow and underground water table simulation

The module RIVER uses two input files to update the river discharge rate as well as the underground water table profile values (the relative altitude of the water table at

distance 10, 20, ..., 150 meters from the river) at the beginning of each day. The water table altitude values are then compared with the input parameter (specifying the distance of the simulated site from the river and the relative altitude of the soil surface above the reference point) and the flood event is generated for the given day, if appropriate (i.e. if the predicted water table is above the surface). For information about the method for generating the river discharge values and modelling the underground water table profiles, the reader is referred to Chapter 3.3.

Season phase subsystem

The whole year is divided into several phases. Their sequence depends on the presence of mowing in the particular year, which might be set as one of the input parameters. The names, meaning and typical starting dates for the individual phases are summarized in Table 6.8. The specification of rules and parameters was done for the *Alopecurus* type, tuning the parameters to both the mowing/no-mowing situations. The maintenance of the current date and of the phases of season is done in the DATE module.

Biomass growth subsystem

This is the core subsystem (module BIOM), where the dry weight of biomass is predicted, in steps of one day, for all the six simulated compartments (dominant species aboveground biomass live or standing dead, other species aboveground biomass live or standing dead, litter compartment and root system compartment). The changes in the compartments are modelled using the RGR (relative growth rate) coefficients expressing the change in the biomass weight as fraction of its weight in the previous day. No attempts were done to separate counteracting processes (e.g. the development of new live biomass versus its movement to the standing dead biomass). Only the changes in the litter compartment are not simulated in this way, but their change is modelled as a fractional change related to the sum of the weights of both aboveground standing dead compartments on the previous day. There are other rules which modify the RGR coefficients before they are applied, based on the fuzzy meteorological variables (Fuzzy Temp and FuzzyPrec). Other rules, acting directly on the biomass compartments (not on the RGR coefficients), are defined for the mowing day and for the processes that occur during the flood (removal of the part of the dead biomass compartments, no growth for the live biomass compartments) and in the period shortly after the flooding ceased (increased growth of live biomass, increased rate of litter decomposition).

Additional rules can be be easily added or existing ones modified as the main simulation system (written in CLIPS language) is a interpreted one and the system reads the rules and facts on the beginning of each simulation run from human-modifiable source files.

Fig. 6.3 shows the dynamics of the root compartment in the results of the one year simulation run for the *Alopecurus pratensis* stand, where the meadow was mown at the beginning of June. There is visible influence of the start of the season of vegetative growth, the change at the beginning of the LatePreMow season, where the amount of the root biomass starts to recover, the rapid depletion following the hay-cut and finally the "recovery" before the early autumn. The dynamics of the above-ground live biomass of Alopecurus pratensis, from the year of the simulation run is shown in Fig. 6.4. In this run, the peak biomass at the end of summer was lower than that achieved before the cutting, but this is not necessarily so for every year.

Table 6.8. Season phases of biomass development in the *Alopecurus pratensis* stand. Mow = mowing.

PHASE	MEANING	TIME PERIOD START
Quiescence	the period of vegetative quiescence, with no living biomass	last week of November / first week of December
EarlyPreMow	beginning of growin season, with fast increase of biomass amounts and little production of standing dead biomass. The production of above-ground biomass is expected to proceed partly at the expense of the root system	depends on the vegetation type. Usually last week of March or first week of April for the *Alopecurus* type
LatePreMow	the period of further accumulation of nutrients. The production of new above-ground biomass is based on the photosynthetic assimilation	beginning of May. The phase ends depending on the date of mowing
EarlyPostMow	short time period after above-ground biomass clipping, with intensive re-growth and depletion of the below-ground biomass resources	at the mowing date; end 11 days after end of LatePreMow phase
LatePostMow	the long period of above-ground and below-ground biomass accumulation	when the EarlyPostMow ends
EarlyNonMow	the meaning of this phase depends on the stand type, in *Alopecurus* the live biomass accumulation slows down and percentage of standing dead increases	mid June
LateNonMow	the above-ground biomass is approximately at its peak and significant amount of "ageing" takes place	mid August
Slowdown	gradual decrease of live above-ground biomass and of its quality; accumulation of standing dead biomass and litter	mid September

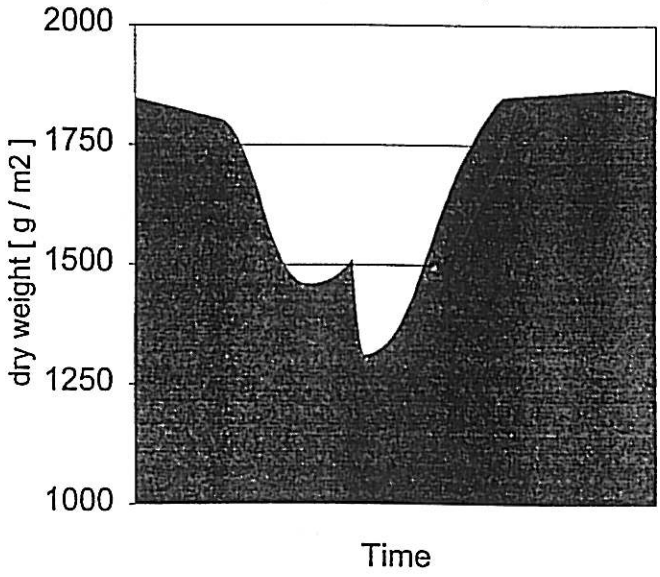


Fig. 6.3. The simulated dynamics of the root system compartment biomass in the *Alopecurus pratensis* stand during one year. The vertical scale shows the dry weight of the functional roots in g . m^{-2}, the horizontal scale starts at 1st January and ends at 31st December.

Table 6.8. Season phases of biomass development in the *Alopecurus pratensis* stand. Mow = mowing.

PHASE	MEANING	TIME PERIOD START
Quiescence	the period of vegetative quiescence, with no living biomass	last week of November / first week of December
EarlyPreMow	beginning of growin season, with fast increase of biomass amounts and little production of standing dead biomass. The production of above-ground biomass is expected to proceed partly at the expense of the root system	depends on the vegetation type. Usually last week of March or first week of April for the *Alopecurus* type
LatePreMow	the period of further accumulation of nutrients. The production of new above-ground biomass is based on the photosynthetic assimilation	beginning of May. The phase ends depending on the date of mowing
EarlyPostMow	short time period after above-ground biomass clipping, with intensive re-growth and depletion of the below-ground biomass resources	at the mowing date; end 11 days after end of LatePreMow phase
LatePostMow	the long period of above-ground and below-ground biomass accumulation	when the EarlyPostMow ends
EarlyNonMow	the meaning of this phase depends on the stand type, in *Alopecurus* the live biomass accumulation slows down and percentage of standing dead increases	mid June
LateNonMow	the above-ground biomass is approximately at its peak and significant amount of "ageing" takes place	mid August
Slowdown	gradual decrease of live above-ground biomass and of its quality; accumulation of standing dead biomass and litter	mid September

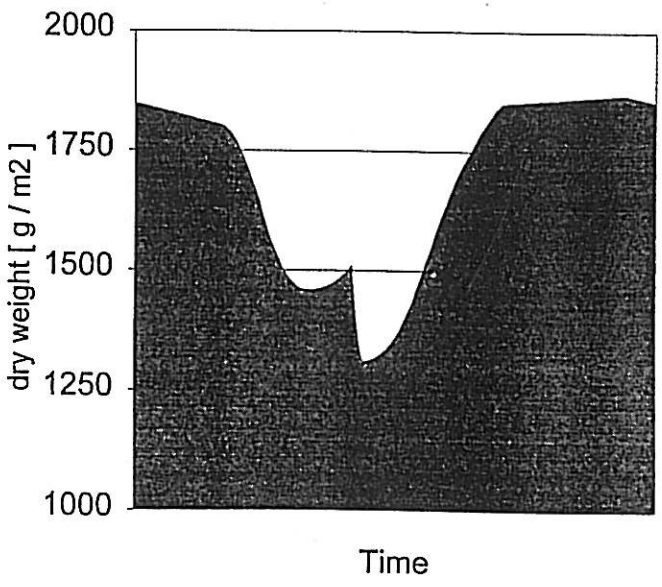


Fig. 6.3. The simulated dynamics of the root system compartment biomass in the *Alopecurus pratensis* stand during one year. The vertical scale shows the dry weight of the functional roots in g . m^{-2}, the horizontal scale starts at 1st January and ends at 31st December.

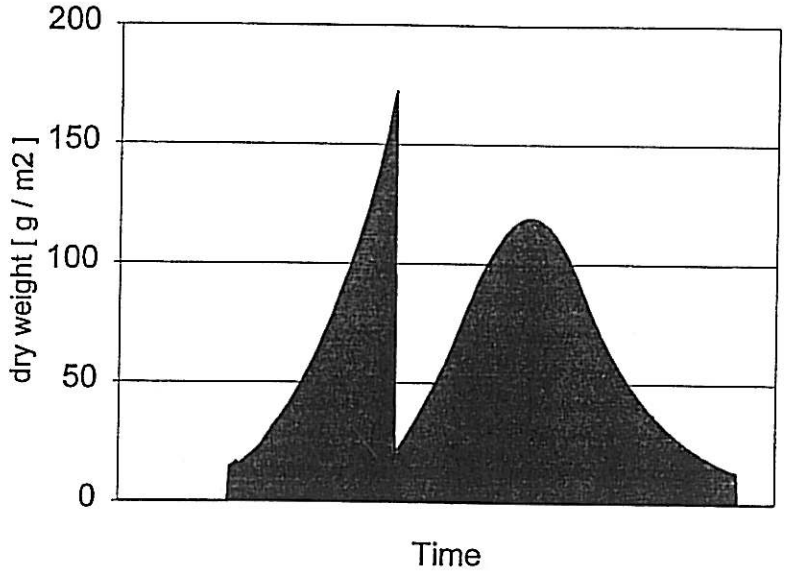


Fig. 6.4. The simulated dynamics of the live aboveground biomass of the dominant species in the *Alopecurus pratensis* stand during one calendar year. The vertical scale shows the dry weight in g . m^{-2}, and the horizontal scale starts at 1st January and ends at 31st December.

Nutrient contents subsystem

The information about the nutrient contents dynamics for the individual biomass compartments does not enable us to model gradual changes in these. This is particularly because a more detailed study of response to particular types of ecological stress would be needed. Also, the structure of the model is too rough to support the simulation of nutrient content changes effectively, namely in the litter compartment, where mixing of different cohorts of the litter – with different C/N ratio – takes place. Therefore, the percentage of N or C is modelled as a fixed value for the given compartment and season phase. These percentage values are then multiplied by the modelled biomass amount to yield the amount of nitrogen or carbon fixed in the biomass compartment. An example of dynamics of the amount of nitrogen in the live aboveground biomass of *Alopecurus pratensis* is shown in Fig. 6.5. The irregularities in the curve are caused by the sudden changes of the percentage of N at the time of change from one seasonal phase to another.

Technical details

The input parameters for the model run are provided in the configuration file. A simple program was writen providing a user-friendly interface for setting-up the parameters (see Fig. 6.6). The input parameters include the position of the simulated point in the river floodplain (to enable appropriate generation of the flooding event), the range of the simulated years, the names of the three input files and the one output (log) file, as well as the dates of the mowing in the individual years. The output log file (by default

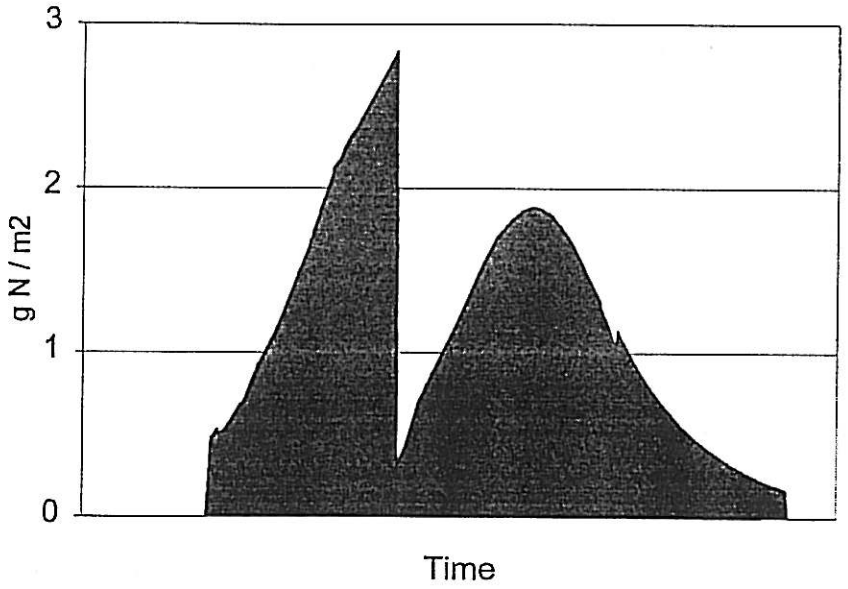


Fig. 6.5. The simulated amount of nitrogen stored in the live aboveground biomass compartment of the dominant species in the *Alopecurus pratensis* stand during one year. The vertical scale shows g N . m^{-2}, and the horizontal scale starts at 1st January and ends at 31st December.

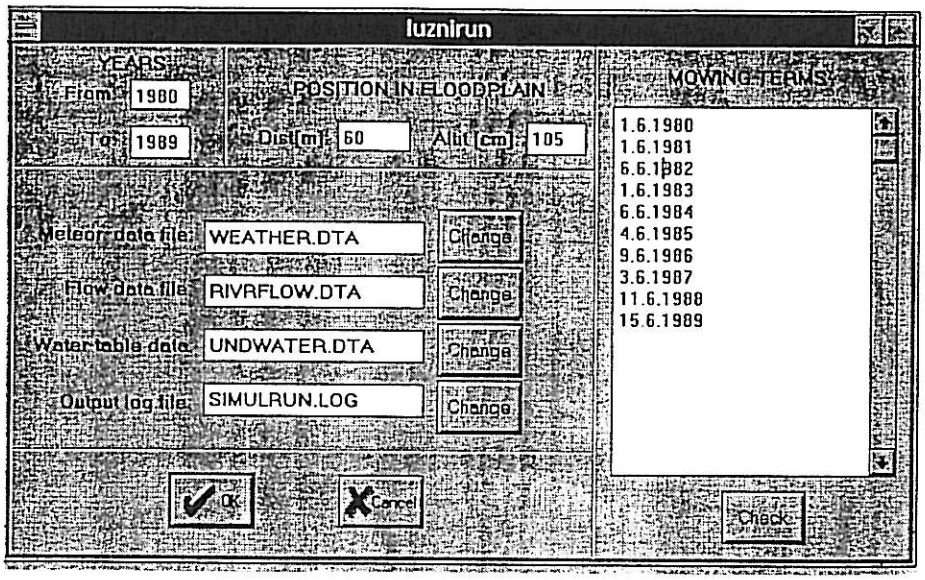


Fig. 6.6. An example of the user-friendly interface for setting up parameters in the model.

called SIMULRUN.LOG) contains one row of information for each simulated day, starting with information about the flood (if any) and amount of dominant/other species biomass removed by the cutting (only at the day the meadow was mown), fol-

lowed (for each day) by the date, name of the season' phase and the current values for the individual biomass compartments, as well as the amount of N and C held in the selected compartments. To enable easy visualization of the simulation results, a program called STRIPLOG.EXE was written which collects the non-regular events data (mowing, flooding) and creates a file which is easily importable into all spreadsheets (tab-separated ASCII format), with one column for each characteristic and the data about floods and mowing events during the whole simulation run summarized in a table at the end of the file.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N. & Csaki, F. (eds.), Second International Symposium on Information Theory, pp. 267- 281. Akademia Kiado, Budapest.

Botkin, B. 1969. Prediction of net photosynthesis of trees from light intensity and temperature. Ecology 50: 854-859.

Cleveland, W.S. 1993. Visualizing data. Hobart Press, Summit, New Jersey.

CLIPS 1993. CLIPS 6.0 Reference manual. Volume I and II. Artificial Intelligence Section, Lyndon B. Johnson Space Center, NASA.

Cole, W.D. 1976. ELM version 2.0. Range Science Dept., Science Series 20. Colorado State University, Fort Collins.

Giarratano, J.C. & Riley, G. 1994. Expert systems – Principles and programming, 2nd ed. PWS Publ. Comp., Boston, Massachusetts.

Hastie, T.R. & Tibshirani, R. 1990. Generalized additive models. Chapman and Hall, London.

Innis, G.S. (ed.) 1978. Grassland simulation model. Ecological Studies, Springer, New York.

Ondok, J.P. & Gloser, J. 1983. Leaf photosynthesis and dark respiration in sedge-grass marsh. I. Model for mid-summer conditions. Photosynthetica 77: 77-86.

Orchard, R. 1994. FuzzyCLIPS version 6.02 user's guide. Knowledge Systems Laboratory, Inst. for Information Technology, National Research Council, Canada.

Randell, L. & Gyllenberg, G. 1972. Weather simulation models. In: Matador Project (IBP), 5th Annual Report, pp. 111-140. Saskatoon, Saskatchewan.

Rychnovská, M. (ed.) 1972. Ecosystem study on grassland biome in Czechoslovakia. PT-PP/IBP Report No. 2., Czechosl. Academy of Sci., Brno.

Rychnovská, M. (ed.) 1985. Ecology of grasslands (in Czech). Academia, Praha.

Rychnovská, M. (ed.) 1993. Structure and functioning of seminatural meadows. Academia, Praha.

Starý, K. 1989. Analysis of the occurrence of synoptic situations and weather related to them (in Czech with English summary). Transactions of the Czech Hydrometeorological Institute 35: 1-163.

Tukey, J.W. 1977. Exploratory data analysis. Addison-Wesley, Reading, Massachusetts.

Zadeh, L.A. 1965. Fuzzy sets. Information and Control: 338-353.

Zadeh, L.A. 1983. The role of fuzzy logic in the management of uncertainty in expert systems. Fuzzy Sets and Systems 11: 199-227.

Chapter 6

Multivariate gradient analysis in ecology: helping ecologist to do it better

Petr Šmilauer: Multivariate gradient analysis in ecology: helping ecologist to do it better

Summary

The applications of multivariate gradient analysis (MGA) in ecological research are many and cover various directions of ecological research. The use of these methods become so widespread and fashionable that inevitably an increased rate of inappropriate applications or inadequate interpretation of their results followed. In this paper, I argue that a software-based guidance is needed that would help the ecologist with the more difficult decisions to be made when using the MGA methods efficiently. An expert system being developed with that purpose is described and its further development discussed.

Using MGA in ecological research

The methods of MGA are used in the ecological research for a rather long time. Nevertheless a "renaissance" of their use happened with their extension and assortment done by Cajo ter Braak and summarized for example in ter Braak et Prentice (1988). The program CANOCO (ter Braak, 1987) implements the majority of the "model-based" methods of multivariate gradient analysis (i.e. leaving out only the methods of (non-metric) multidimensional scaling type). The "taxonomy" of the MGA methods features two important criteria. First, the methods are distinguished based upon the underlying model of response of the variables in the primary data (most often populations of plant or animal species) to the (hypothetical) gradients of their "environment" into *linear* and *unimodal response methods*. Next, two approaches to analysing the change in the primary data (typically the composition of assemblages of species populations) are distinguished: either these compositional gradients are hypothethical and maximize the fit of the primary data to the underlying response model (here the MGA method with linear response model is Principal Components Analysis - PCA and the method with the unimodal response model is (Detrended) Correspondence Analysis - (D)CA) or they are constructed under the constraint of being linear combinations of measured explanatory variables (the resulting MGA method for the linear response model is Redundancy Analysis - RDA and for the unimodal response model the Canonical Correspondence Analysis - CCA). While there is certainly a danger of indiscrimate use of these methods in ecology (which is the danger of any of the statistical methods), the methods like the CCA are generally accepted and used for analysis of ecological phenomena on various spatial and temporal scales (Palmer 1993).

The use of the methods of MGA in ecology flourished through the last ten years and the sophistication of their application increased as well. This might be seen from the bibliography published by Birks et al. (1994), where only a subset of the MGA methods is covered (the constrained MGA, including the CCA, RDA, and Canonical Variate Analysis - CVA). Yet almost 380 studies employing one or other form of constrained MGA were collected from the time period 1986 to 1993. Two most important fields of their application are terrestrial plant ecology and limnology, with a prominent place of paleolimnology. Nevertheless, there are important applications to the ecology of habitats of various taxonomical groups of animals (insects, birds, spiders).

While the MGA methods represent potentially very useful tools for exploring ecological data, their potential cannot materialize for free. A substantial background knowledge on these methods and the methods of visualizing their results is needed. And because the theory related to the application of MGA methods is rather extensive, the software implementing these methods

gets overly complicated, too. Based on these facts, courses covering the theory and practice for these methods are becoming an important part of the university curricula of plant and animal ecologists.

The widespread use and popularity of the MGA methods brought the problem of researchers applying these methods without necessary knowledge of their theoretical background or even the practical rules for their use and interpretation of their results. Based on my experience with teaching the MSc-level students and with the researchers applying these methods in their studies, I consider the following list to represent the most frequent problems that the users of the MGA methods face:

- users fail to distinguish different applicability of the methods based on the linear versus unimodal response model and its dependence on the properties of their data. Most often, the CCA or DCA (Detrended Correspondence Analysis) is used even when the corresponding methods based on the linear response model (RDA or PCA, respectively) are more appropriate
- users tend to use constrained MGA methods (such as CCA or RDA) at any occassion where explanatory variables ("environmental variables" in the terminology of the program CANOCO) are available. Even if the factors (presumably influencing the primary data) are measured, it is useful to distinguish the different purposes of focusing the analysis on the variability that might be explained by those explanatory variables (the constrained methods) and the assessment of importance of the measured variables in explaining the main patterns in primary data (as done by using non-constrained method with passively interpreted explanatory variables)
- when using the constrained MGA, the user indiscriminately includes all the explanatory variables in the analysis, even when there is strong collinearity among some of the explanatory variables
- similarly, the user applies and interprets the results of constrained MGA even if the results of these methods do not indicate any relation between the primary data and the explanatory variables
- users often do not realize that even if the MGA method used is the appropriate one and it summarizes the data with reasonable efficiency, the reliability of the positions of species optima along an environmental gradient is not the same for all the species and that for many species no reliable conclusion can be made
- users do not use the results of the MGA methods efficiently and only the very basic features among those that might be seen in ordination diagrams are interpreted

The expert system

The problems described in the previous section represent a potential niche for expert system software, that would be able to partly mimic the role of an expert user of the MGA methods and advise less-experienced users of these methods on their appropriate use.

An expert system is a kind of software that attempts to emulate the decision-making ability of human expert in a restricted and precisely delimited field of expertise (Giarratano et Riley 1994). The expert systems are employed in various areas of human activities including medical diagnosis, mineral exploration, financial analysis, military applications. The reasons the expert systems are developed and used range from the need to retain knowledge that a particularly skilful expert might possess up to the ability to apply the human expert's knowledge with substantially higher speed than any human being can (a good example is the application of expert system during the pre-flight controls on the space-shuttles). Nevertheless, the most frequent reason of applying the expert systems in practice is to bring an expert knowledge to

many users where availability of human expert is not always practical. In this way, we might look at the creation of expert system as a particular form of transferring knowledge from one (or several) person(s) to the users by the means somewhat complementary to the more classical papers, textbooks or lectures. The expert systems advising about the statistical analysis are also available, e.g. the Statistical Navigator system (Anonymous 1993).

A common mis-conception about expert system is that it is a kind of general reasoning software being able to productively 'think' about the problems it is going to solve. On contrary, the success of expert systems is based on their 'shallow knowledge'. Expert systems match the knowledge they gain (by asking the user, looking into the available data etc.) with their knowledge base and based on that stored knowledge assert new facts, ultimately leading to suggestions or diagnostic that the system is supposed to produce. The knowledge base has usually the form of so-called **production rules**, with the general form:

IF conditions that hold **THEN** conclusions that can be made
or
IF goal to be satisfied **THEN** actions that should be taken

The difference between the two above forms is usually not so strong and human expertise in a particular area is usually composed from both types (beside other knowledge including the semantic knowledge about relation between various terms and meta-knowledge, concerning the applicability of the "knowledge proper" in particular context or in particular order). Yet, this distinction might lead to different approaches to building expert systems: an expert system has two substantial parts - the knowledge base and the inference engine, which is a set of algorithms for applying the knowledge base to a particular problem. The inference engine could take one of two approaches for reasoning. The **forward chaining** algorithm applies the rules to the available facts, makes some intermediate conclusions (by asserting new facts), which are matched to other rules etc., until the final conclusion is reached. This strategy is most suitable for expert activities like planning experiments (or data analysis), monitoring or controlling processes, and generally to any problems with a multitude of possible outcomes. On the other hand, the inference engine based on **backward chaining**, starts from the desired goal that a particular run of the expert system has to solve. The production rules in the knowledge base decompose the given goal(s) into several subgoals, each of them dependent on other subgoals or on certain facts being true. This kind of problem-solving strategy is most suitable for the diagnosis of problems (e.g. medical diagnosis) or generally for problems where the desired conclusions do not differ much in their structure. While the method of reasoning selected for implementation of particular expert system influences the structure of the knowledge-base to a significant extent, there is no clear and hard line between both approaches, as any kind of knowledge can be implemented with greater or lesser difficulty with any one of them.

CANOEXP system

The CanoExp program is a first version of an expert system which collects practical knowledge related to the appropriate use of MGA methods in ecological research. The Fig. 1 displays the general structure of the current version of the CanoExp program.

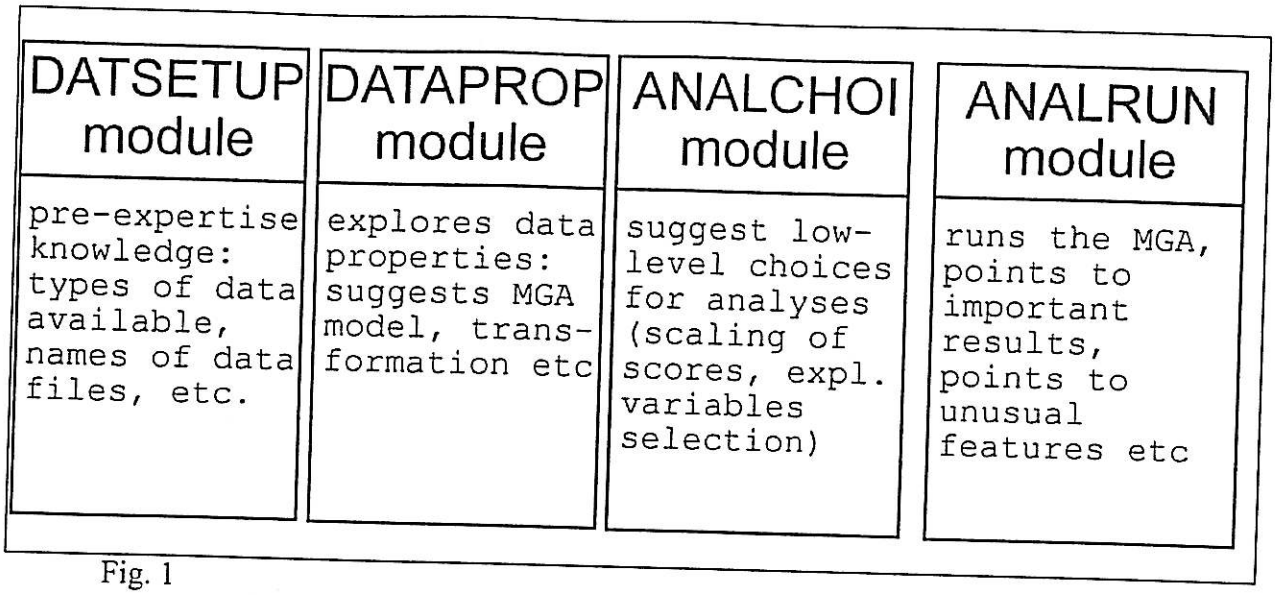


Fig. 1

Currently, the knowledge base is divided into four functional modules. There is fifth module (called MAIN), but it contains rules not related to the expert domain (rules controlling the sequential flow of the analysis, logging the knowledge, methods for asking user about particular facts, etc.).

The work of the expert system proceeds through the modules in Fig. 1 sequentially, from left to right. During this process, new knowledge about the particular data-analytical problem is derived, based on various information sources. Inside each of the modules, the rules belong to one of four groups, differing in priority:

1) highest priority is given to an initialization rule which asserts new facts based on the knowledge gained so far (e.g. in the DATAPROP rule, a fact collecting information about properties of explanatory variables is asserted only if the user indicates that the explanatory variables are available). The **slots** for such fact are initially set to 'unknown' and subsequently filled-in by the rules of lower priority.

2) the second highest priority is reserved for rules that check for **constraints**. These constraints guard the solution achieved so far against being inconsistent with the semantic model of the domain. For example, if a rule changes the type of response model for a MGA from unimodal to linear, a constraint rule might be activated if there is any setting for detrending (because the detrending is a technique available only with the MGA methods similar to DCA).

3) most of the rules have intermediate priority and represent the main knowledge about the particular phase of data analysis selection.

4) the rules with lowest priority are those which initialize particular slots in the facts by asking the user. The CanoExp program asks the user about statistically-oriented decisions only if there is no alternative way of reaching the answers.

In following paragraphs, the purpose of the individual modules is described, together with the changes intended for the next release.

Module **DATSETUP** asks the user about the general description of the research problem, in terms of what is the primary goal of the analysis and what data sets are available for the analysis. Currently, this phase expects the user has at least intermediate knowledge of the theory behind the MGA methods or of the relevant terms in regression analysis etc. The next version should make the inquiry more oriented towards the ecologist's way of thinking. The most suitable method here seems to be to present the user with a collection of examples of research problem and asking the user to indentify himself with one of those. The user has to provide the system

58

with the location of the data files to be used in the analysis, as the system is build with the assumption that it will be consulted only after the source data were made available.

Module **DATAPROP** builds on the knowledge acquired in the previous module, opens the data files relevant for the selected type of analysis, checks the properties of the variables (are there only positive values in the data, are there only binary variables etc.) and asks user about information that would be too difficult to guess (are the variables in the primary data all measured on the same scale - e.g. percentage cover). Based on that, the system runs an exploratory run of a MGA to check the beta-diversity in the primary data set and based on the results, it decides about the species response model which determines the type of MGA (linear vs. unimodal).

Module **ANALCHOI** finalizes the choices needed to run the MGA with the program CANOCO (namely, to generate a response file of the CON type - see ter Braak, 1987). The choices include use of detrending, the method of scaling ordination scores, downweighting the occurrences of rare species etc.

Module **ANALRUN** currently only generates the response file for CANOCO program and runs the program. In future release, it will more fully integrate with the CANOCO program for interactive procedures, such as forward selection of explanatory variables (ter Braak, 1987). Also, the heuristic knowledge about the procedures useful in **ordination diagnostics** (Šmilauer, 1994) will be incorporated. This will enable iterative problem solving, with possible returns to the earlier phases of data analysis selection and with employing parallel, complementary strategies of looking at the particular research problem.

Integration with other tools

The described expert system should make user's life easier, not more difficult with having to run another piece of software. Therefore, CanoExp should integrate transparently with the future Windows version of CANOCO. This version of CANOCO will allow to select the options for a MGA analysis in a similar, but more user-friendly way than is used by CANOCO version 3.1. The CanoExp integration would allow selection of default options based on the properties of the data to be analysed and on the answers of the user to few non-statistical questions.

Further integration aims at employing the knowledge about the research problem, gained during the CanoExp session, in advising user on the appropriate ways of visualizing the ordination results. This field is also very appropriate for casting into expert systems, as there are many pieces of shallow knowledge ("rules of thumb") relating to the quality of ordination diagrams, to the selection of variables to be displayed in the ordination diagram, to the ways of illustrating particular aspects of the data or addressing particular research problem, etc. Consequently, the CanoExp shall integrate smoothly with the future version of the program CanoDraw (Šmilauer, 1992).

Technical aspects and future developments

The CanoExp program is developed as a combination of a procedural system for reading and analysing source data, parsing the output of program CANOCO and of the knowledge based implemented using the expert system shell CLIPS, version 6.04 (CLIPS 1993). The CLIPS shell is a high-quality, open software which is distributed with source code, so that integration of special procedures (like the code for analyzing CANOCO input files and running CANOCO program) and their calls inside the rules is possible. Also, the whole expert system might be embedded in other programs.

The expert system CanoExp is currently implemented so that all the domain-specific knowledge is represented in authoritative form. Nevertheless, the implementation would be more realistic with the following two types of extensions:

- the expertise about the use of MGA methods operates with somewhat inexact terms. A typical example is the decision between linear response- and unimodal response-type methods. This might be based on the range of the first ordination axis for a trial run of DCA (ter Braak et Prentice, 1988): there is not clear-cut threshold of the length of the gradient represented by the first ordination axis beyond which the linear model is entirely inappropriate. The rule would be formulated in terms like
  IF the beta-diversity of axis 1 is not too-high THEN use linear model
  Such in-precise, but more real-life formulations are supported by the theory of linguistic variables (Zadeh 1979) and implemented e.g. in the FuzzyCLIPS extension of the CLIPS 6.0 system (Orchard 1994).
- beside the inexactness of many terms in the domain of MGA, there is also some uncertainty about applicability of individual rules and about certain input information. For example, it might be too restrictive to ask the user whether his / her attention is focused on the configuration of variables (taxa) or of the sites. (S)he might prefer to select only a certain extent of preference to one of the options. This, as well as the certainty bound with the conclusions of individual rules is also implementable with FuzzyCLIPS (Orchard 1994). A more ambitious goal of a self-improving knowledge base which updates the rules' certainty factors given the user's assessment of the appropriatness of the conclusions made in previous sessions is also under consideration.

References

Anonymous (1993): Statistical Navigator Professional. - Idea Works Inc., Columbia, Missouri.

H. J. B. Birks, S. M. Peglar, H. A. Austin (1994): An annotated bibliography of Canonical correspondence analysis and related constrained ordination methods 1986 - 1993. - Botanical Institute, University of Bergen, Norway. 58 pp.

CLIPS Reference Manual (1993): Volume I. Basic Programming Guide. - Software Technology Branch - NASA, Johnson Space Center, USA.

J. Giarratano, G. Riley (1994): Expert Systems. Principles and Programming. Second Edition. - PWS Publishing Company, Boston. 644 pp.

R. Orchard (1994): FuzzyCLIPS version 6.02 User's Guide. - Knowledge Systems Laboratory, Inst. for Information Technology, National Research Council, Canada.

W. M. Palmer (1993): Putting things in even better order: the advantages of Canonical Correspondence Analysis. - Ecology, 74: 2215 - 2230.

P. Šmilauer (1992): CanoDraw 3.0 User's Guide. - Microcomputer Power, Ithaca, USA. 118 pp.

P. Šmilauer (1994): Exploratory analysis of paleoecological data using the program CanoDraw. - J. of Paleolimnology, 12: 163 - 169.

C. J. F. ter Braak, I. C. Prentice (1988): A theory of gradient analysis. - Advances in Ecological Research, 18: 271 - 317.

C. J. F. ter Braak (1987): CANOCO - a FORTRAN Program for Canonical Community Ordination by [Partial] [Detrended] [Canonical] Correspondence Analysis, Principal Components Analysis and Redundancy Analysis (Version 2.1). - Agriculture Mathematics Group, Wageningen.

L. A. Zadeh (1979): A Theory of Approximate Reasoning. - Machine Intelligence, 9: 149 - 194; edited by J. E. Hayes, D. Michie, L. I. Mikulich.