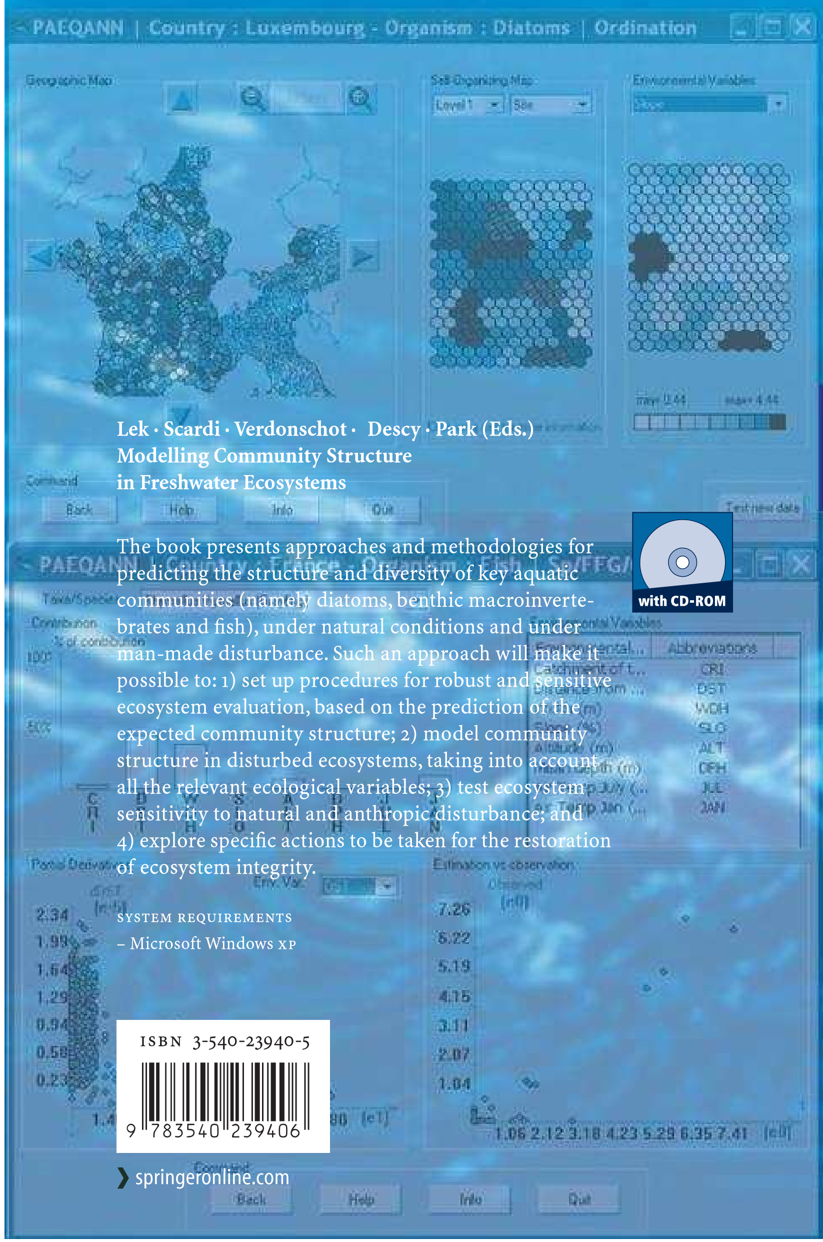
Gestalter
**ERICH KIRCHNER**
Heidelberg

Druckfarben
HKS 47 blau
HKS 82 braun

Dieser Farblaser-Ausdruck dient nur als Anhaltspunkt für die farbliche Wiedergabe und ist nur bedingt farbverbindlich.



Lek · Scardi · Verdonschot · Descy · Park (Eds.)
Modelling Community Structure
in Freshwater Ecosystems

with CD-ROM

The book presents approaches and methodologies for predicting the structure and diversity of key aquatic communities (namely diatoms, benthic macroinvertebrates and fish), under natural conditions and under man-made disturbance. Such an approach will make it possible to: 1) set up procedures for robust and sensitive ecosystem evaluation, based on the prediction of the expected community structure; 2) model community structure in disturbed ecosystems, taking into account all the relevant ecological variables; 3) test ecosystem sensitivity to natural and anthropic disturbance; and 4) explore specific actions to be taken for the restoration of ecosystem integrity.

SYSTEM REQUIREMENTS
– Microsoft Windows XP

ISBN 3-540-23940-5
9 783540 239406

❯ springeronline.com

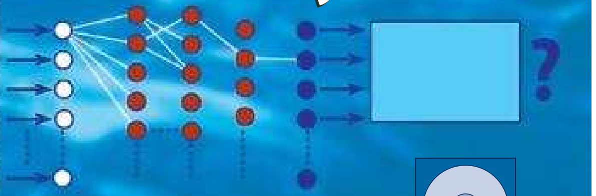
Lek · Scardi · Verdonschot · Descy · Park (Eds.)

Modelling Community Structure in Freshwater Ecosystems

with CD-ROM

Sovan Lek
Michele Scardi
Piet F.M. Verdonschot
Jean-Pierre Descy
Young-Seuk Park
Editors

# Modelling Community Structure in Freshwater Ecosystems



with CD-ROM

Springer

tive abundances present at a given site. At each site, major habitats were selected over a 10 to 30 m long stretch of the waterbody and were sampled by means of the same sampling effort. The sampling effort was thus standardised for each site.

At shallow sites, vegetation habitats were sampled by sweeping a pond-net (200 mm × 300 mm, mesh size 0.5 mm) several times through each vegetation type over a length of 0.5-1 m. Bottom habitats were sampled by vigorously pushing the pond net through the upper few centimetres of each bottom type over a length of 0.5 to 1m. The habitat samples were then combined for the site to give one sample with a standard area of $1.5\ m^2$ $1.2m^2$ of vegetation & $0.3\ m^2$ of bottom). At sites lacking vegetation, the standard sampling was confined to the bottom habitats. At deeper sites, five samples were taken from the bottom habitats with an Ekman-Birge sampler. These five grabs were equivalent to one 0.5 m pond net bottom sample. Vegetation habitats were sampled with a pond net as described above. Again the total sampling area was standardised as $1.5\ m^2$. Macroinvertebrate samples were taken to the laboratory, sorted without any external aid, counted and identified to species level.

The sampling dates were spread over the four seasons as well as over several years (1981 up to and including 1985). Season was taken into account by defining sampling periods as nominal "environmental" variables within the analysis.

A data sheet was used to note a number of abiotic and some biotic variables in the field. Some were measured directly (width, depth, surface area, temperature, transparency, percentage of vegetation cover, percentage of sampled habitat), others (such as regulation, substratum, bank shape) were classified. Field instruments were used to measure oxygen, electrical conductivity, stream velocity and pH. Surface water samples were taken to determine chemical variables. Other variables, such as land-use, bottom composition, and distance from source, were gathered from additional sources (data from water boards, maps). In total, 70 abiotic variables were measured at each site.

### Diatoms

Several teams of the PAEQANN network were involved in providing diatom records and the corresponding environmental data: *CEMAGREF* (France), *CRPGL* (Luxembourg), *ARCS* (Austria) and *LFE-URBO-FUNDP* (Belgium), coordinator of the "diatom group".

The PAEQANN Diatom Database comprises 2847 records in total. To make a list of sufficiently representative and/or significant taxa and to guarantee sufficient homogeneity among samples, only records comprising at least 380 counted objects, and originating from sampling carried out on stony substrates were selected for further analyses. Consequently, 2147 records were finally available for further analysis, among which 467 were identified as reference, according to their IPS[‡] value (equal or higher than 16). 1719 different taxa names were recorded for the whole database, among which 1255 different taxa could be identified. After grouping, 1051 taxa were potentially available for further analysis. A selection of taxa was nevertheless made prior to analysis, in order to remove occasional taxa; this led to a list of 123 taxa for the reference data matrix, and 283 for analysis of the 2147 records constituting the whole database data matrix.

Finally, it must be mentioned that in order to implement the PAEQANN tool, a simplified MS Access Diatom database was produced.

---

[‡] IPS = Index of Pollution Sensitivity

# 8 General conclusions and perspectives

Scardi M[*]

Understanding the way community structure is affected by environmental conditions is a key issue in modern ecology, especially since the advent of ordination techniques and, in particular, of multivariate statistical methods for direct gradient analysis. New statistical tools have been developed by ecologists to address this problem more effectively, and they have been widely accepted and used in many studies. Canonical Correspondence Analysis (ter Braak 1987), which constrains coenocline analysis to physical, chemical and other environmental variables, is an example of such an ecologically-inspired data analysis procedure.

Although data analysis techniques made it possible to infer relationships between community structure and environmental variables, they are mainly useful for descriptive purposes or, to a limited extent, to test ecological hypotheses. Of course, some results of data analysis procedures can be used as a basis for assessing species' responses to environmental variables, but this is only possible to a limited extent and from a strictly qualitative point of view.

Therefore, if predictions about community structure are needed, suitable modelling techniques have to be used. In particular, statistical methods that directly relate environmental variables to species presence and/or abundance can play a role, ranging from very simple regression models to more complex ones (e.g. Partial Least Squares). However, only a few applications have proven to be useful, such as those based on logistic regression for predicting species presence or absence given adequate environmental data.

Modelling single species distributions as a function of environmental (mostly abiotic) information is certainly an interesting task, but predicting community structure is a more complex problem that usually cannot be solved just by assembling single species models into a more complex composite model. In fact, in many cases the available information about the environmental relationships (both biotic and abiotic) that determine species distributions is too limited, and it cannot support the development of reliable models.

Therefore, it is obvious that the efficiency of the modelling approach plays a fundamental role in predicting community structure in such data-limited situations. This is the reason why the number of ecological applications involving Artificial Intelligence (A.I.) techniques and Machine Learning methods has grown significantly during the last ten years.

These new modelling methods rely on computing power that is now easily available to extract as much useful information as possible from the existing - and usually insufficient - data. Sometimes these approaches do not provide significant advantages over conventional methods, but they are often much more effective than the latter and some applications among those that are presented in this book provide clear evidence for their superiority.

About ten years since the first attempts, ecological applications of A.I. and Machine Learning modelling methods are now mature and, in many cases, they are presented without comparisons to conventional counterparts, as their improved performance is accepted. Readers who are interested in understanding to what an extent these methods may be beneficial can find complete, yet easy introductions and examples in Fielding (1999) as well as in Lek and Guegan (2000).

A common factor in many applications aimed at modelling freshwater community structure is that the number of field records is usually limited with respect to the ecological

---

[*] Correspondence: mscardi@mclink.it

complexity of the problem, thus making it very difficult to reconstruct the causal relationships that link species distribution to environmental variables. However, in such data-poor situations other sources of information are often available, although deeply embedded in the data sets, and they can be exploited by appropriate modelling methods.

For instance, many species tend to occur in association with others, while other species seldom coexist at the same site. Obviously, this can be due to the biotic interactions that make a set of species a community, but in some cases species assemblages are modelled, rather than real communities, and therefore the relevance of biotic interactions cannot be taken for granted. Nevertheless, relationships between species are still useful from the modeller's point of view as they narrow the number of independent combinations of species to be predicted. Basically, if two species respond in the same (or in opposite) way to a given set of environmental conditions, then the actual dimensionality of the modelling problem to be solved is lower than expected and less data are needed to build a good model.

Moreover, morphodynamic features play a fundamental role in defining the ecological characteristics of freshwater ecosystems, specifically in streams and rivers. In other words, given a set of physical constraints (e.g. elevation, slope, etc.) only a very limited range of ecological conditions is likely to be observed, thus reducing the "degrees of freedom" of models aimed at predicting the structure of communities or other biotic assemblages, at least within a given ecoregion.

Such ecological relationships, however, are usually complex, and linear or unimodal responses that are the basis for most statistical approaches are seldom observed. On the contrary, complex non-linear relationships are often involved, and this is probably the more important reason for the success of new modelling strategies. In particular, Artificial Neural Networks (ANNs) have been successfully applied in many cases, both for predictive modelling (usually via supervised methods) and for descriptive modelling, i.e. for revealing underlying patterns in data sets (usually by means of unsupervised methods). The majority of the case studies that are presented in this book are based on ANN applications and this is clear - although not unbiased - evidence for the role they play in modelling community structure.

Independent of the modelling technique, however, there are limits to the predictability of species distributions that strictly depend on the intrinsic nature of ecological data sets. In fact, different species may occur with very different frequencies in data sets, according to their actual density or as a consequence of the sampling strategy or spatial scale. Since species presence (as well as abundance) depends on environmental conditions, rare species, as well as nearly ubiquitous ones, are usually unpredictable, because in both cases it is virtually impossible to detect significant correlations between environmental variables and species presence (or abundance).

This limitation cannot be overcome by improving modelling algorithms or by introducing new techniques. The only viable solution is to modify sampling strategies, making them more suited to the modelling needs. As a matter of fact, most attempts at modelling community structure are usually carried out on the basis of data sets collected for other purposes, e.g. for mapping species distributions via GIS tools or for applying multivariate statistical methods for indirect gradient analysis. A typical feature in such data sets is a regular or random sampling design that is certainly adequate when no prior information is available, but that often fails to reveal essential information if small scale coenoclines also play a role. In these cases, sampling strategies that address variable spatial scales would provide much more relevant information, especially when previous data or pilot surveys are available and the sampling design can be effectively stratified. In this framework, modelling community structure, independently of the accuracy of the results obtained in the first attempts, may also induce both a significant optimisation in sampling strategies and a better understanding of the factors that control species distributions.

New modelling techniques that are able to exploit existing information as efficiently as possible (such as ANNs) will certainly catalyze a better ecological understanding of factors controlling community structure because of their ability to reproduce complex non-linear responses. In particular, sensitivity analysis of such models may provide useful insights into the ecological relationships that control species distributions and biotic interactions, thus offering clues to improve sampling designs and sampling scales to resolve relevant biotic signals. This task is not trivial, of course, and, while several methods have been developed (as presented in this book), no well established standards are available at present. However, this topic is certainly among the most stimulating and new approaches to sensitivity analysis are emerging. The successful application of methods that elucidate second and possibly higher order interactions between abiotic factors in determining species distributions, as well as complex interactions among species, will probably be the main goal for the next decade in modelling community structure.

Sensitivity analysis is an example of indirect use of models as tools for stimulating advances not only in computational methods, but also in the ecological background of the modelling applications, thus involving the way data are collected, coded, analyzed, etc. On the other hand, ecological issues may play a significant role in improving modelling techniques and especially in adapting them to properly handle the peculiar characteristics of ecological data. For instance, using Mean Square Error (MSE) as a measure for goodness of fit is a common practice in modelling, and it is certainly adequate for many quantitative variables. Therefore, it is usually adopted as a default choice and, in the case of many software packages, is the only available option. It is obvious, however, that this way of measuring the distance between observed and modelled data is seldom appropriate when species abundance data (not to mention species presence data, i.e. binary data) are taken into account. The vast amount of similarity and distance coefficient that have been developed for measuring differences in community structure are a clear evidence for the inadequacy of Euclidean distances and related coefficients from an ecological viewpoint. In particular, when community structure is concerned, it is obvious that the role of each species must be interpreted in the light of its ecological context.

A very simple example of this need is in the different weight that should be given to the same error in predicting the presence or the abundance of a given species in the case of a very simple community (low species richness) and in the case of a more complex one (high species richness): MSE obviously fails in this task, because it does not scale the errors with respect to the complexity of the community, whereas other coefficients, such as Jaccard or Bray-Curtis similarity, do. Thus, significant improvements in community structure modelling could be achieved by adapting existing algorithms to their ecological framework, e.g. by adopting procedures for measuring modelling errors that are based on appropriate metrics. Adaptations of modelling algorithms are probably beyond the capabilities of most ecologists, but some are actively working in this field, developing new strategies and methods for ecological modelling that more closely match their specific needs.

The need for ecologically sound metrics is not only a problem in predictive modelling, of course. As many community ecologists already know, the outcomes of multivariate analyses are very deeply influenced by the selected metrics. Not only the results of quantitative analyses often are quite different from those of qualitative ones, but even among the results of analyses based on the same type of data there might be significant differences. The most obvious example is that the meaning of absence data (zeroes) is not the same in all the cases: it can express real absence of a species in a given site, but it can also depend on the frequency of its occurrence (i.e. by its density) with respect to the characteristics of the sampling design and devices. Thus, the selection of appropriate metrics is a key issue even in descriptive models, such as those based on Self-Organizing Maps (see the applications in this book) as well as those based on conventional ordination techniques.

Given the available methods and data, community structure modelling is really feasible only in case of "toy" problems or when complexity is somehow reduced. Good results can be achieved by focusing on subsets of the whole community (e.g. on assemblages of species that are not too complex and that can be sampled in an effective and straightforward way, like fish), or by simplifying the way the community structure to be modelled is represented (e.g. by using a few trophic guilds instead of a complex list of species). In any case, the number of successful applications is rapidly increasing, and it is clear that the ability of these models to address complex problems is only limited by the availability of adequate information, i.e. by the lack of field data, either in general or with respect to the spatial scale that is relevant to the distribution of the species to be modelled. Obviously, expectations for species distribution and community structure modelling are also growing, but potential users have to bear in mind that the reliability of mathematical models depends on the adequacy of the data bases that support their development, e.g. no reliable weather forecasts would be available if the underlying models were not supported by large meteorological observation networks and data bases.

Despite the difficulties in developing community structure models, the demand for such tools in applied ecology is certainly growing, as comparing observed community structure with some reference conditions may form the basis for assessing environmental quality. This will be the case, for instance, in those European countries that will follow guidelines indicated by the EU Water Framework Directive (also known as Directive 2000/60/EC), which sets restoration targets and clearly points out that changes in community structure with respect to reference conditions are related to changes in the ecological status of a water body. However, in many cases natural reference conditions are not available because of the lack of genuinely unperturbed water bodies, and only models can provide estimates about the expected community structure in such pristine conditions.

In this rapidly evolving scenario, mathematical models aimed at predicting community structure will certainly play a key role in many applied and basic research tasks. Conventional models (e.g. statistical models) will continue to be widely used, but ecological modellers who want to be on the leading edge will explore new approaches. This book is both a showcase of successful applications and a useful reference for those who want to get started in this field.

# References

Aboal M, Puig, MA. Soler G (1996) Diatoms assemblages in some Mediterranean temporary streams in Southeastern Spain. Arch Hydrobiol 136: 509-527

Adkison MD, Peterman RM (1996) Results of Bayesian methods depend on details of implementation: an example of estimating salmon escapement goals. Fish Res 25: 155-170

AFNOR (1992) Qualité de l'eau - Détermination de l'Indice Biologique Global Normalisé (IBGN). Norme NF T 90-350. AFNOR, Paris

AFNOR (2000) Détermination de l'Indice Biologique Diatomées (IBD). Norme NFT 90-354. AFNOR, Paris

AFNOR (2004) Qualité de l'eau - Détermination de l'indice poisons rivière (IPR). NF T90-344. AFNOR, Paris

Agences de l'eau (1998) SEQ milieu physique: système d'évaluation de la qualité du milieu physique des cours d'eau. - Les études des Agences de l'Eau (testing version 0). Agences de l'eau

Aguilar Ibarra A (2004) Les peuplements de poissons comme outil pour la gestion de la qualité environnementale du réseau hydrographique de la Garonne. PhD Thesis, Ecole National Supérieur Agronomique de Toulouse, Institut National Polytechnique, Toulouse, France

Aguilar Ibarra A, Gevrey M, Park YS, Lim P, Lek S (2003) Modelling the factors that influence fish guilds composition using a back-propagation network: assessment of metrics for indices of biotic integrity. Ecol Model 160: 281-290

Aguilera PA, Frenich AG, Torres JA, Castro H, Vidal JLM, Canton M (2001) Application of the Kohonen neural network in coastal water management: methodological development for the assessment and prediction of water quality. Water Research 35: 4053–4062

Alba-Tercedor J, Sanchez-Ortega A (1988) Un metodo rapido y simple para evaluar la calidad biologica de las aguas corrientes basado en el de Hellawell (1978). Limnetica 4: 51-56

Alhoniemi E, Himberg J, Parviainen J, Vesanto J (1999) SOM Toolbox 2.0, a software library for Matlab 5 implementing the Self-Organizing Map algorithm. [online] http://www.cis.hut.fi/projects/somtoolbox/

Allan JD (1995) Stream ecology: structure and function of running waters. Chapman & Hall, London

Allen TFH, Hoekstra TW (1992) Toward a unified ecology. Columbia University Press, New York

Allen TFH, Starr TB (1982) Hierarchy: perspectives for ecological complexity. The University of Chicago Press, Chicago

Almeida SFP, Pereira MJ, Gil MC, Rino JM (1999) Freshwater algae in Portugal and their use for environmental monitoring. In: Prygiel J, Whitton BA, Bukowska J (eds) Use of algae for monitoring rivers III, Agence de l'Eau Artois-Picardie, Douai pp 10-16

Andersen MM, Riget FF, Sparholt H (1994) A modification of the Trent index for use in Denmark. Water Research 18: 145-151

Anderson A (1935) The irises of the Gaspe Peninsula. Bull Am Iris Soc 69: 2-5

Angelier E (2001) Ecologie des Eaux Courantes. Tec & Doc, Paris