assemblages can be predicted by means of SOM, ANN-BP and environmental variables in headwater streams of Luxembourg.

Perspectives and further developments

The typology of assemblages for benthic diatoms is a fundamental goal already studied since several years. For instance Descy and Coste (1991) developed a table for determining the CEC index. This table gives groups with taxa ranked following their pollution tolerance, and subgroups ranked following the natural diatom successions according to Strahler order. The European Water Framework Directive (European Parliament and The Council of the European Union 2000) formalised this concept of typology and introduced the concept of reference condition. It requires the characterisation of different surface water bodies with hydromorphological and physical and chemical variables and to define their reference conditions (which correspond to its highest ecological status, i.e. to undisturbed conditions). According to the European Parliament (2000) and the REFCOND working group (2002), the "reference conditions establishment can be either spatially based or based on modelling or a combination of these".

Our study shows that it is possible to predict with a high accuracy diatom assemblages of headwater streams in different ecoregions and different water qualities using physical and chemical variables. Artificial neural networks seem to be, at this stage, good techniques to meet the requirements of the European Water Framework Directive. Next developments could be the integration of the REFCOND working group requirements (2002) in the methodology presented here in order to correctly define reference conditions and then to develop new river quality assessment tools.

5.8 Use of neural network models to predict diatom assemblages in the Loire-Bretagne basin (France)*

Di Dato P†, Rimet F, Tudesque L, Ector L, Scardi M

Introduction

The aim of our work was to test the accuracy of the artificial neural networks (ANN) as predictive tools for benthic diatom taxa presence starting from a set of environmental variables. The river basin we studied is characterized by a huge complexity both in terms of spatial heterogeneity, as far as the environmental information is concerned, and in terms of biotic information, because of the large number of taxa that have been identified. In particular, this study focused on the application of different approaches to the reduction of the complexity of the data set and of the ANN models. In the meantime, it was also aimed at showing, as already pointed out for other kind of organisms (Scardi et. al. in # 3.8, Di Dato et al. in # 4.3), that too frequent or too rare species usually provide trivial information about the relationships of environmental variables with their presence or absence, thus affecting the accuracy of the models. The taxa that were selected according to the different approaches we tried including only those that can be actually modelled on the basis of the available environmental information. The limits and the perspectives of these species selection approaches are then thoroughly discussed.

Materials and methods

This study was carried out on a data set collected in the Loire-Bretagne (Loire-Bretain) basin, which is located in western France. In this basin a wide variety of climates and landscapes can be found, but for the sake of simplicity, it can be divided into 3 different sections. The mid and high Loire basin is rather hilly and mountainous with the Massif-Central mountains in the south-eastern part. The low Loire basin is characterised by lowland and some large rivers. The Bretagne region is a rather flat region, composed by lentic rivers and is deeply influenced by the Atlantic Ocean (Fig 5.8.1).

The diatom database has been assembled collecting 641 samples from 1996 to 2000 in the framework of national river survey, organised and funded by the Loire-Bretagne Water Agency. For each sampling, physical, chemical and geomorphological descriptors were recorded. Among the latter, some were measured in the field (width, shading, sampled substrate and current velocity), others on maps and on Geographical Information Systems (geology, altitude, distance from source, slope, catchment area, discharge). Average values computed over the last 3 months were assumed for physical and chemical descriptors (temperature, pH, dissolved oxygen, dissolved organic carbon, HCO_3^-, CO_3^{2-}, NO_3^-, NO_2^-, NH_4^+,

* This work is part of the PAEQANN project supported by the European Commission under the 5th Framework Programme (contract n°: EVK1-CT1999-00026). We thank Mr J. Durocher of the "Loire-Bretagne" Water Agency and Ms M. Leitao of the Bi-Eau society for their valuable collaboration.

† Correspondence: pdidato@mclink.it

PO_4^{2-}, Ca^{2+}, Cl^-, Na^+, biological oxygen demand, chemical oxygen demand, suspended matter, NKJ, P_{tot}). The environmental variables that were used to train the models are presented in the Table 5.8.1.

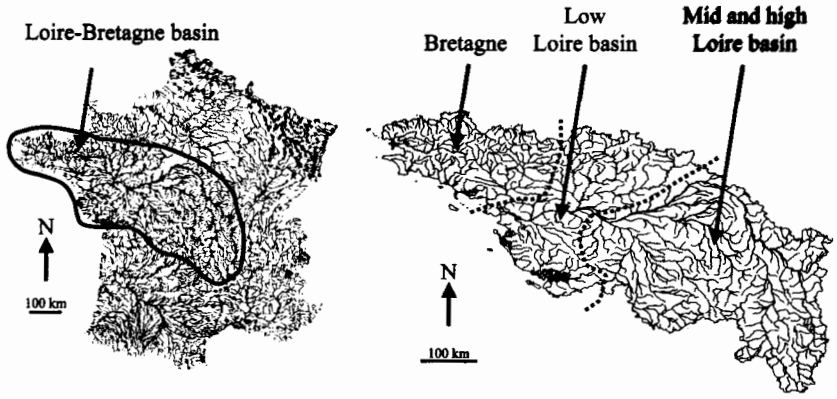

Fig. 5.8.1. The study area. Black dots in the map on the left show the sampling sites.

Table 5.8.1. List of the environmental variables measured or assessed for each sampling site.

Altitude (m)	**HCO_3^- (mg L^{-1})**
Source distance (km)	NO_3^- (mg L^{-1})
Slope (m km^{-1})	NO_2^- (mg L^{-1})
Catchment area (km^2)	NH_4^+ (mg L^{-1})
Electric conductivity 20°C (□S cm^{-1})	PO_4^{2-} (mg L^{-1})
pH	Ca^{2+} (mg L^{-1})
Dissolved oxygen (DO) (mg L^{-1})	Cl^- (mg L^{-1})
Dissolved organic carbon (DOC) (mg L^{-1})	Na^+ (mg L^{-1})
Biological Oxygen Demand (BOD) (mg L^{-1})	NKJ (Kejdahl nitrogen) (mg L^{-1})
Suspended matter (mg L^{-1})	P_{TOT} (mg L^{-1})
Shading (1:closed - 3:open)	Current velocity (1:lotic - 3:semi-lotic)
Geology (limestone: 1/0)	Geology (quaternary sediment: 1/0)
Geology (sandstone: 1/0)	Geology (other:1/0)
Geology (granitic: 1/0)	Geology (mudstone-schiste-schale:1/0)
Geology (volcanic:1/0)	

The sampling of benthic diatoms was carried out by means of a brush. The epilithon was collected by brushing at least 5 stones found in lotic zones of the sampling site. In laboratory, the samples were cleaned with hydrogen peroxide and hydrochloric acid to dissolve calcium carbonates. Then the cleaned frustules were mounted in a resin (*Naphrax*).

A total of 930 diatoms taxa were determined. Therefore, the data set included 641 patterns (samples), for 32 predictive environmental variables (input variables) and 930 diatom taxa (output variables). Two subsets of taxa were then extracted from this data set. In the first subset (A), only taxa that were present in more than 5% and less than 95% of the 641 samples were included (202 taxa met this condition). In the second subset (B) all the taxa that were either present or absent in less than 100 samples were excluded, thus considering

only those taxa whose frequency of occurrence ranged from 15% to 85% (91 taxa met this condition). The latter subset was also used with two different modelling strategies. The first strategy was based on the development of a single model for all the species, i.e. a single model with 91 outputs, whereas the second involved developing 91 separate models, each one with only one output. In all cases 3-layer feed-forward ANNs were used, i.e. perceptrons with a single hidden layer.

In order to evaluate how much information was retained after selecting species subsets, we compared the results of a Principal Coordinate Analysis (Gower 1966) based on the whole set of species with the ones of Principal Coordinate Analyses based on the species subsets (i.e. on 202 and 91 taxa, respectively). The Jaccard's coefficient (Jaccard 1908) was used to compute dissimilarity among samples. This asymmetrical coefficient was selected because it only takes into account the presence of species, ignoring absence data, which are not always completely reliable.

The 641 available patterns were randomly distributed among three subsets for training, validation and test. The training set included 50% of the available patterns, whereas the validation set as well as the test set included 25% of the available patterns. The validation set was used to compute the Mean Square Error (MSE) of the ANN outputs after each training epoch. The test set allowed obtaining unbiased estimates of the accuracy of the trained ANN models.

Floristic information was expressed at the lowest information level, i.e. as binary presence/absence data. The predictive environmental variables were normalized by rescaling their range of variation within the [0,1] interval.

The optimal structure of the ANN models was defined after empirical tests. In practice, the number of nodes in the hidden layer was selected by comparing the MSE for different ANNs, with up to 60 nodes in the hidden layer. The ANN structure that provided the best performance was the one with 51 hidden nodes for the subset A (202 taxa), whereas 11 hidden nodes were used for both subset B (91 taxa) modelling strategies. In the latter case the selection of the ANN structure was based on the average MSE.

During the training procedure only a random subset of training patterns (50% of the available patterns) was submitted to the ANN at each epoch in order to prevent overtraining due to the memorization of the pattern sequence. In order to better generalize the ANN learning, white noise in the [-0.01,0.01] range was also added to each input value (Györgyi 1990).

In all the nodes of the hidden and output layers of the ANN sigmoid activation functions were used. The error back propagation algorithm was selected to adjust the weights during the training procedure. In particular, we applied an early stopping procedure based on the validation set MSE. The learning rate and the momentum were respectively set to 0.90 and 0.10 and never modified during the training.

The continuous outputs values of the ANN were converted to binary using a threshold function and then compared with the observed data to obtain the percentage of Correctly Classified Instances (CCI).

For each species subset (A and B) and for each modeling strategy (1 and 2) we analyzed the accuracy of the models by testing the independence of the modeled data wit respect to observed data. Therefore, we computed the K statistic (Cohen 1960, Kraemer 1982) on the basis of contingency tables in which presence and absence data in the model output and in the target patterns (i.e. in observed data) were cross-tabulated as shown in Table 5.8.2.

Table 5.8.2. Contingency table for K statistics computation.

		Model output	
		Presence	*Absence*
Target	*Presence*	1-1 A	1-0 B
	Absence	0-1 C	0-0 D

The K statistic was then obtained as:

$$K = \frac{O_a - E_a}{N - E_a} \tag{5.8.1}$$

where O_a is the observed count of CCI ($O_a = A+D$), E_a is the count of CCI that are expected if the model is independent of the observed data ($E_a=[A+B][A+C]/N+[C+D][B+D]/N$) and N is total number of cases ($N=A+B+C+D$).

Results

Reducing the number of species in a floristic data set has a cost in terms of information about the overall structure of the species assemblages. In other words, it is not possible to preserve the whole amount of information about the relationships among samples when only a subset of species is considered. However, a smart selection of the most relevant species might reduce the information loss to a minimum. In order to provide a very rough estimate of the degree of approximation about the diatom assemblage structure that we accepted when we decided to reduce the number of species to be modelled, the results of a Principal Coordinates Analysis performed on the whole data set and on the two subset of species were compared (Fig 5.8.2).

The two scatter plots show that the distortion of the first Principal Coordinate for each sample due to the reduction of the number of species is quite limited both in the case of subset A (202 taxa, left) and subset B (91 taxa, right). In fact, in both cases the correlation between the Principal Coordinates is very high (Spearman's r= 0.969 for Subset A and r=0.992 for subset B). Of course, the Mantel statistics between the dissimilarity matrices is highly significant in both cases, so we rejected the null hypothesis of independence of the subsets from the whole data set. These evidences suggest that reducing the number of species does not affect significantly the ability to represent the main features of the diatom assemblage and to reproduce the relationships among samples.

The results of the ANN models were evaluated by taking into account the test data set only (i.e. 25% of the available patterns). In other words, the accuracy of the predictions was estimated on the basis of information that is completely independent of that on which the ANN models were developed. This strategy, that is often overlooked when conventional statistical models are considered, allows both to obtain unbiased estimates of the modelling errors and to effectively compare different modelling approaches.

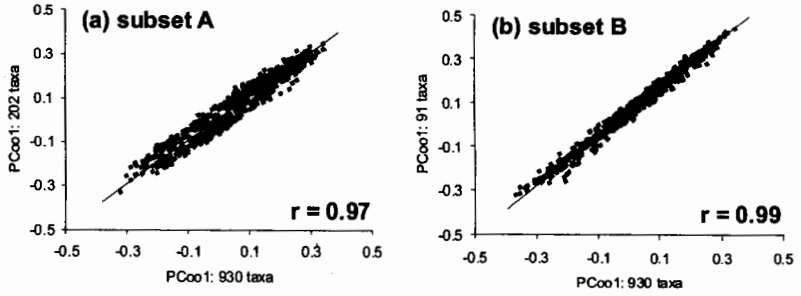

Fig. 5.8.2. First Principal Coordinate of the diatom assemblages: all the species vs. reduced species set. Subset A, 202 taxa, on the left (a), and subset B, 91 taxa, on the right (b).

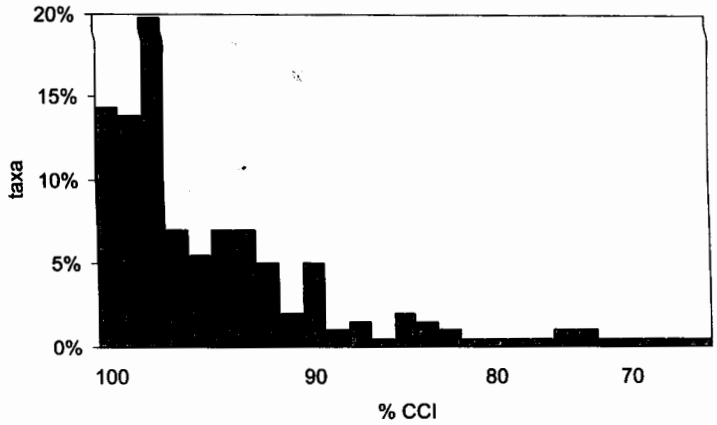

Fig. 5.8.3. Percentage of correctly classified instances (CCI) for the subset A (202 taxa).

The percentage of correctly classified instances (CCI) for the larger subset of species (subset A, 202 taxa) was quite high, as the average value was 94.5% (Fig 5.8.3). More than 10% of the taxa had a percentage of CCI of 100%, whereas almost 20% had a percentage higher than 97%. In Figure 5.8.3 the frequency distribution of the CCI percentages is shown, and it is very clear that almost 90% of the taxa had CCI values larger than 90%.

Although the model performed very well according to this criterion, the values of the K statistic did not confirm that result. In fact, only 5 taxa out of 202 were effectively predicted by the model, i.e. in only 5 cases the prediction about species presence or absence was significantly different from random according to the K statistics (Fig. 5.8.4a).

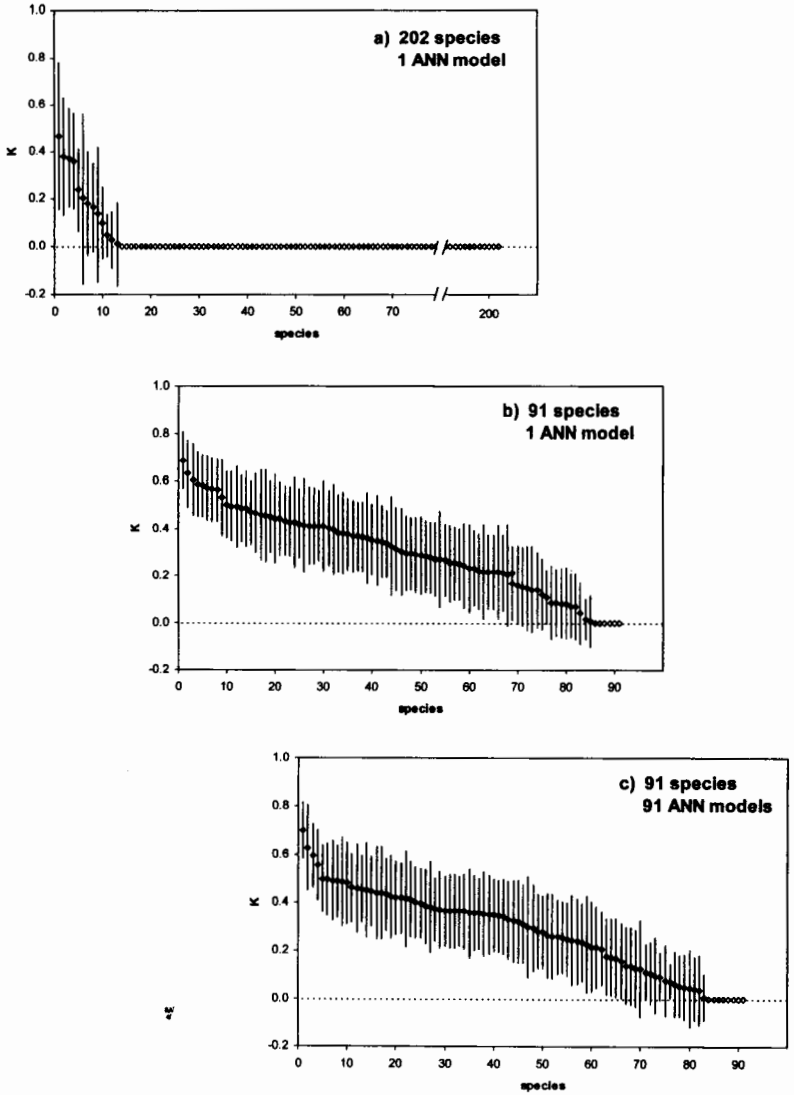

Fig. 5.8.4. K statistics values with confidence intervals for the different modelling strategies: a) species subset A (202 species, 1 ANN model); b) species subset B, strategy 1 (91 species, 1 ANN model); c) species subset B, strategy 2 (91 species, 91 ANN models). Solid diamonds indicate significant K values (i.e. cases in which the lower P=0.95 confidence bar does not intersect the K=0 line). White diamonds at K=0 with no bars stand for species that were never predicted by the model. In this case the K value was actually undetermined.

The disagreement between CCI percentages and K statistics results might seem surprising at first glance, but it can be easily explained if the errors in model prediction are carefully analysed. In fact, most species are present only in a very limited number of samples (on the average 6.6 samples out of 125, i.e. slightly more than 5%, in the test set) and it is very difficult for the ANN model to correctly predict their presence on the basis of a handful of known cases. Therefore, when dealing with rare species, ANN models learn to predict only species absence independently of their inputs, easily attaining low mean square errors and very high CCI percentages. Needless to say, the predictions of those models are useless from a practical point of view, and the high CCI percentages they attain are meaningless.

On the contrary, the K statistics is able to provide a reliable estimate of the predictive ability of a model, as it takes into account the relative frequency of presence and absence records. Of course, predictions about rare species are inherently unreliable because of the lack of "examples" about the relationships between species presence and environmental variables, but a "false" model cannot obtain an high (and significant) K statistics value if it is not able to predict enough instances of presence with respect to the real frequency of the species. It is not surprising that the five species that were associated to significant K statistics were more frequently found than the average (21.4 cases out of 125, i.e. 17.1%).

Reducing the number of species to be modelled by excluding the most rare ones seems therefore a sound choice, especially considering that the relationships among samples are well described even when rare species are excluded. Therefore using a smaller subset of species, selected on the basis of their frequency and excluding the rarest ones, makes definitely sense and that is the reason why the subset B (91 taxa) was selected. In this case, however, two modelling strategies were adopted, respectively involving a single model with 91 outputs (strategy 1) and 91 models with a single output (strategy 2). The results for the two strategies are shown in Figure 5.8.4b,c as far as the K statistics is concerned.

The CCI percentages, although not relevant in the light of the evaluation of the predictive capabilities of the models, are lower than in the model for species subset A (on the average 75.2% and 74.2%, whereas the first model attained 79.9%). In particular, 68 out of 91 species had significant K statistics values with strategy 1 and 67 out of 91 with strategy 2. Thus, the two strategies for the species subset B returned similar results, although the average K value for strategy 1 was slightly larger than the one for strategy 2 (0.30 and 0.28, respectively).

A more detailed comparison of the K statistics values obtained for strategies 1 and 2 is shown in Figure 5.8.5. The unit slope line corresponds to a perfect agreement between strategy 1 and strategy 2 K statistics values, and it is evident that the overall agreement between the two series is rather good (r=0.77). This implies that species that some species are intrinsically more predictable than others, as they tend to have high K statistics values independently of the modelling strategy.

However, there are also differences in the accuracy of the predictions that depend on the modelling strategy. In fact, points below the unit slope line correspond to species that were more accurately predicted using strategy 1 (i.e. $K_1>K_2$), whereas strategy 2 was a better option for species represented by points above that line (i.e. $K_2>K_1$). Since 50 points are located below the unit slope line and only 36 above it, strategy 1 seems more effective than strategy 2, but the difference between the two is not dramatic. The K values are exactly the same for both strategies in 5 cases.

Finally, the case of *Fragilaria capucina* var. *vaucheriae*, the outlier in the lower right corner of the plot, is worth commenting, as it was the only case in which there was a very large difference between the K statistics values for the two modelling strategies. In particular, this taxon was accurately predicted in the strategy 1 model, whereas the strategy 2 model completely failed.

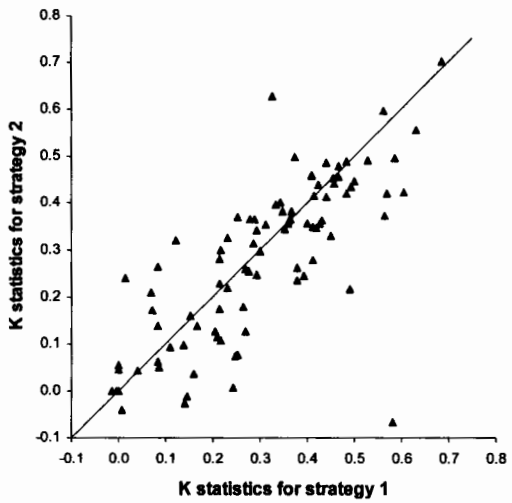

Fig. 5.8.5. K statistics values for the modelling strategy 1 (one model for all the species) are compared to those for modelling strategy 2 (a different model for each species). The overall agreement between the two strategies (r=0.77) indicates that some species are intrinsically more predictable than others. The unit slope line represents the perfect agreement between the two strategies that was observed in 5 cases. Strategy 1 provided larger K statistics values in 50 cases (points below the unit slope line), while strategy 2 was more effective in 36 cases (points above the unit slope line). The outlier in the lower-right corner represents Fragilaria capucina var. vaucheriae, which was accurately predicted only by strategy 1.

The only likely reason for this difference is that information about species interactions (or association) that is implicitly embedded in the strategy 1 model (that predicts all the species simultaneously and has as a much more complex structure) may play a role in cases in which the relationships between environmental variables and species distribution are so weak that the latter cannot be reliably modelled by a single species model (strategy 2). The alternate hypothesis, of course, is that strategy 2 modelling failed by chance, and a more effective model could have been obtained by further iterating the ANN training procedure.

Another comparison between the two modelling strategies for species subset B was carried out on the basis of the results of Principal Coordinate Analyses performed on Jaccard's dissimilarity matrices. In particular, the first Principal Coordinates for the predicted species composition of the samples in the test set were plotted against the first Principal Coordinates for the observed species composition of the same samples for both strategies (Fig. 5.8.6). The agreement between Principal Coordinates obtained from analyses involving observed and predicted diatom assemblage compositions is a proxy for the agreement between predicted and observed dissimilarity matrices, which, in turn, is a proxy for the resemblance of the predicted and observed assemblage composition.

The best strategy, according to this comparison criterion, was the one based on single species model, i.e. strategy 2 (see Fig. 5.8.6b). The rank correlation between Principal Coordinates based on predicted and observed data (Spearman's r=0.726) was higher than in that case of strategy 1 (Spearman's r=0.812), but the difference between the two values of the correlation coefficient was not significant (n=148, p=0.070).

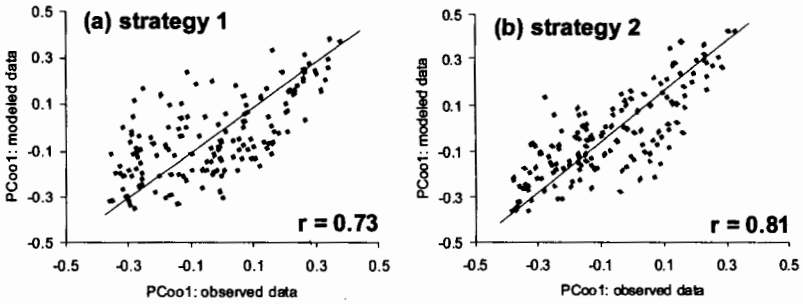

Fig. 5.8.6. Principal Coordinates obtained from Jaccard's dissimilarity matrices computed from predicted data were compared to those obtained from observed data according to the same procedure. Results based on modelling strategy 1, i.e. one model for all the species is shown on the left (a), whereas those based on strategy 2 (a separate model for each species) are shown on the right (b).

These results are not in agreement with those based on the K statistics, which but in both cases the differences between the two modeling strategies were minor. In fact, the Mantel test allowed rejecting the null hypothesis of independence between the dissimilarity matrices based on predicted and observed data in the case of both modeling strategies.

Discussion and conclusions

This study provided further evidence about the problems in modelling rare species, as well as ubiquitous, if any, that has been pointed out elsewhere (e.g. Scardi et. al, this book). In fact, even in the case of diatom assemblages such species significantly affect the performance of ANN models.

In particular, very high percentages of correctly classified instances (CCI) are often a false indication of model accuracy. As CCI only take into account the number of cases in which the model predictions match observed data, they are strongly affected by the relative frequency of presence and absence records. Therefore they are biased indicators of model accuracy when dealing with rare or very common species and other approaches, like the one based on K statistics, should be selected.

Excluding rare species from a data set may cause a significant reduction in the number of species, and the ability to correctly represent the complex relationships among samples might be seriously impaired. However, if properly selected, even a subset of taxa that represent a significant part of the assemblage structure may preserve enough information. Our results showed that 91 taxa out of 930, selected in a way that excluded the less frequent ones, were able to provide an adequate representation of the differences in the diatom assemblage structure among many different sites.

Most of the taxa that were well predicted by the model are found in eutrophic waters: *Cocconeis placentula* var. *placentula, Fistulifera saprophila, Gomphonema minutum, Gyrosigma attenuatum, Mayamaea atomus* var. *permitis, Navicula tripunctata* according to van Dam et al. (1994), *Diatoma vulgaris* according to Hoffman (1994), *Navicula antonii, N. capitatoradiata* according to Lange-Bertalot (2001). No taxa characterizing oligotrophic waters were well predicted by the model.

The most accurately predicted species, independently of the model, was *Navicula tripuncata,* which is considered as good indicator specie for eutrophic waters with average to high electrolyte content (Lange-Bertalot 2001). *Navicula antonii* is also a species consid-

ered as good indicator for waters that are often affected by anthropic sources of perturbation (Lange-Bertalot 2001). The high efficiency of the ANN models in predicting the above-mentioned species on the basis of physical and chemical variables, including pollution indicators, supports the hypotheses about an environmental control of their distribution.

Taxa as *Fistulifera saprophila* and *Mayamaea atomus* var. *permitis* have particular ecological characteristics. In fact, they are found in heavily polluted water and are α mesopolysaprobic taxa according to van Dam et al. (1994). As environmental variables that are linked to pollution parameters play an important role among the models inputs, it is not surprising that the models are able to accurately predict the distribution of these species.

Nitzschia cf. *tropica* is an invasive species (Coste and Ector 2000) and its presence is associated to a well-defined area within the studied region, i.e. the upper part of the mid and high Loire basin. The accuracy of the ANN models in predicting its presence is probably related to the use of geographical predictive variables (e.g. elevation, distance form source) that allow recognizing sites within the boundaries of the area where this species can be found.

In practice, however, in many cases the model had not enough information to learn the taxa response to environmental variables. This is probably the case especially for pollution sensitive taxa, since the sampling network of this study is adapted to the assessment of pollution in the Loire-Bretagne basin. The database is composed only by 7% of very good biological quality samples according to the diatom index SPI (Specific Pollution sensitivity Index; Coste 1982).

A similar case is the one involving *Bacillaria paradoxa, Nitzschia clausii, N. filiformis.* These taxa, which are uncommon in the Mid and High Loire basin, indicate brackish waters (van Dam et al. 1994) even if their abundance in the assemblage is low. Despite this remarkable stenoecy, ANN models are not able to accurately predict their presence in such particular environments because of the relative rarity of the occurrence of brackish waters (less than 1% of the samples were collected in waters with conductivity above 1000 µS/cm).

On the other hand, several very common taxa were not well predicted. For instance, this was the case of *Achnanthidium minutissimum*, which is absent in only 9% of the samples and is considered as pollution sensitive specie in the diatom indices (Coste 1982, Descy 1979, Leclercq and Maquet 1987a, b). However, this taxon has also an opportunistic behaviour (Ivorra 2000) and develops rapidly in the biofilm when clear space is available (Sabater 2000) and when inter-specific competition has decreased (Rodríguez 1994). As biotic factors play a relevant role in controlling the abundance of this taxon, its behaviour is rather difficult to predict. Moreover, the lack of an adequate number of records in which this species is absent made it even more difficult for the ANN models to predict its distribution.

In fact, the results improved by reducing the number of taxa and by defining subsets of species that accounted for a very large part of the variation within the ecological data set. Starting from this reduced set of suitable data, we obtained very similar results with two different modeling strategies. However, strategy 1 (a single model predicting all the species) can be considered theoretically more effective, because it can take profit of the information that is provided by the inter-specific relationships, even though they are not known.

An alternate approach to the reduction of the number of taxa to be modelled might involve the reduction of the taxonomic resolution to the genus level. Some diatom indices based on information at the genus level were developed for bioindication (Coste and Ayphassorho 1991). However, this approach is probably not the best one, and it was not included in our tests. In fact, pollution tolerant and pollution sensitive taxa often occured in the same genus in the dataset of the Loire basin.

An interesting perspective for reducing the complexity of diatom databases could be in the use of life forms that have been described in several papers (e.g. Hoagland et al. 1982), but the applicability of this approach in biomonitoring studies is still to be evaluated.

In conclusion, our results showed that accurate models for the prediction of diatom assemblages on the basis of environmental variables can be developed using ANNs. Unfortunately, these models are not able to predict all the species, because in many cases the information about the relationships that link species distribution to environmental variables or to other species distribution (e.g. via competitive interactions) is not sufficient. Rare species and almost ubiquitous ones belong to this category, and the only way to improve the accuracy of the predictions about their distribution is to develop models *ad hoc*, using very focused data sets. To reduce the complexity of the data, dividing the dataset of the Loire basin in sub-datasets by use of an adequate typology should be tested. Then developing models for each sub-dataset corresponding to precise river types should improve the efficiency of the prediction of diatom species abundances.

However, even a subset of species can be very effective in reproducing the ecological relationships among sampling sites and in pointing out sites in which the diatom assemblage structure is impaired because of various types of disturbance. Models that are able to predict the main features of the expected diatom assemblages will effectively support the detection of ecological perturbations and the evaluation of the environmental quality in freshwater ecosystems, and ANNs will certainly play a major role in this scenario.