tion content of the data. Further optimisation of the ANN models could also be obtained by the selection of more appropriate input variables (d'Heygere et al. 2002).

Conclusion

To predict different river restoration scenarios, 'extreme' datasets were added to the original one. Therefore, ecological expert knowledge was used. The presence/absence of Asellidae in the 'extreme' validation set was predicted well when the number of 'extreme' sites in the training set increased. However, the overall predictive power of the ANN models decreased when a relatively large virtual training dataset was applied. Three case studies have shown that ANN models are in general quite robust with a rather high prediction reliability. For very extreme situations, addition of 'extreme' data to the training dataset can be very useful. However, for practical applications, the positive influence of the addition of 'extreme' data to the training dataset is negligible. Although the addition of 'extreme' data has no negative influence on the predictions, data-driven models are often sufficient to obtain good model predictions for practical applications. As a conclusion, adding extreme data improves the reliability of ANN models significantly for predictions under similar conditions. The addition of only one 'extreme' case is however insufficient to obtain a significant improvement of the predictive performance in similar cases, while the addition of too many 'extreme' cases in the training set decreases the general predictive performance of the models.

4.3 A neural network approach to the prediction of benthic macroinvertebrate fauna composition in rivers[*]

Di Dato P[†], Mancini L, Tancioni L, Scardi M

Introduction

Predicting the composition of benthic macroinvertebrate fauna in rivers is not a trivial task, both because of the number of species to be modelled and because of the complexity of biotic and abiotic relationships that determine their distribution. However, the composition of the benthic macroinvertebrate fauna usually provides very useful insights into the ecological quality of lotic systems, as these organisms are very sensitive to disturbance. Benthic macroinvertebrates are relatively sedentary and long-lived, with life cycle durations ranging from a few months to 2-3 years, and they show a wide range of adaptations to local environmental conditions. They represent a continuous monitoring system of the water body where they are living, but they are.also very easy to collect and to identify, at least at an intermediate taxonomic level. Therefore, benthic macroinvertebrates are widely used as biological indicators (Hellawell 1986) and, in particular, they have been used for many years as a source of information for computing several biotic indices that are now used worldwide to assess biological water quality (e.g., Metcalfe 1989, Resh et al. 1996, Lammert and Allan 1999). In this study, the Italian IBE index (Ghetti 1997), derived from the Extended Biotic Index proposed by Woodiwiss (1981) was used as a reference for selecting ecologically homogeneous taxa.

Several different biotic indices have been developed, as they had to be suited to ecoregional characteristics in order to provide correct diagnoses of the riverine ecosystem quality. Most indices, however, share the same rationale that is based on the identification of sensitive taxa and on the recognition of the ecological role of other taxa. The main advantage of this approach with respect to more thorough community structure analyses lies obviously in its simplicity. In fact, even people with a limited taxonomic background can be easily trained to carry out rapid surveys aimed at the computation of biotic indices. A more complex approach to the assessment of the ecological status of streams and rivers is based on the prediction of the whole community structure. In the case of benthic macroinvertebrate fauna, different modelling techniques based on ecological knowledge and monitoring data are now available. In the United Kingdom, the work by Wright et al. (1984) led to the prediction of community types on the basis of environmental data by means of a multivariate analysis procedure. This appraoch was then extended and used in the River Invertebrate Prediction and Classification System (RIVPACS) (Wright et al. 1993b), which provides estimates of the ecological quality at a given site by comparing the observed macroinvertebrate fauna composition with the expected one.

[*] Funding for this research was provided by the EU project PAEQANN (N° EVK1-CT1999-00026).

[†] Corresponding: pdidato@mclink.it

The RIVPACS approach has also been adapted to other ecoregions. For instance, the Australian River Assessment Scheme (AUSRIVAS) (Simpson and Norris 2000) is based on the RIVPACS approach, although it has been expanded and adapted to each Australian ecoregion. Another method that is closely related to RIVPACS and AUSRIVAS is the benthic assessment of sediment (BEAST) (Reynoldson et al. 1995), that is based on quantitative data about macroinvertebtare fauna instead of presence/absence data only. Even though the RIVPACS approach proved to be very effective, it has limits related to the non-linearity, complexity and dynamic nature of biotic responses to environmental characteristics. Moreover, the development of an assessment system based on the RIVPACS rationale requires a considerable amount of work and thorough statistical analyses.

A new generation of empirical techniques for analysing and modelling complex ecological data in a more simple and straightforward way is now emerging. Among these new modelling methods Artificial Neural Networks (ANNs) play a relevant role and represent a useful tool when relationships among data are unknown and/or non-linear. ANNs learn from examples and do not require *a priori* theoretical models, nevertheless they are able to model complex temporal and spatial patterns and to reproduce the behaviour of very complex systems (Recknagel and Wilson 2000). During the last 10 years, ANNs have been applied to various ecological fields (see, for instance, Lek and Guegan 2000), including studies relating community characteristics with environmental variables (e.g., Chon et al. 1996, Recknagel 1997, Recknagel et al. 1997, 1998, Guégan et al. 1998) and modelling habitat suitability (e.g., Paruelo and Tomasel 1997, Ozesmi and Ozesmi 1999). As for the particular case of macroinvertebrate fauna, Pudmenzky et al. (1998) and Walley and Fontama (2000) recently developed ANN approaches that are aimed at the same goals and ecoregions as AUSRIVAS and RIVPACS respectively.

Our study was focused on a benthic macroinvertebrate data set provided by the Latium Regional Environmental Protection Agency and it is aimed at testing different strategies for modelling the presence or absence of macroinvertebrate benthic taxa on the basis of environmental variables, using ANN models.

Materials and methods

Our data set is based on 153 sampling sites, distributed over 76 rivers in the Latium region (Central Italy), where macroinvertebrate fauna was sampled between 1998 and 2000. The hydrographic characteristics of the study area are highly variable, as a consequence of the very diverse origin and evolution of the river basin. The main river in the area is the Tiber, which is the second longest river in Italy, flowing from the north-eastern Appenine mountains through Central Italy and Rome to the Tyrrhenian Sea. All the rivers and streams in the area studied are located in the Tiber basin, with the exception of those in the Liri-Garigliano basin, which is located in the southern part of the Latium region.

The macroinvertebrate benthic fauna was collected at each sampling site by means of a small dredge. The dredge consisted of a handle, a rectangular frame (25 x 40 cm) and a cone-shaped net. The net was made of nylon and mesh size was 0.5 mm. The net had a cup-shaped detachable jar at its closed end that facilitates the collection of the organisms sampled. The sampling sites were dredged from bank to bank to cover all the microhabitats using a technique called "kick sampling". According to this technique the dredge, placed on the bottom of the river with the mouth against the water flow, was dragged along a fixed transect. At the same time the operator scrambled the substrate with his feet in order to direct the benthic organisms towards the net. The fauna collected was preliminarily sorted *in situ*, but an in-depth study by stereomicroscope was then carried out in the laboratory on material fixed in alcohol (70%). The taxonomic analyses led to the identification of 174 taxa. At each sampling site 11 environmental variables were also recorded (Table 4.3.1) some of them were derived from maps or from Geographical Information Systems (elevation, distance from source, gradient), while others were measured in the field (watershed drainage area, water flow, structure of sediment in terms of granulometric classes). The whole data set included 153 records for 11 predictive environmental variables and 174 taxa.

Four modelling strategies, based on different model structures and different complexity levels in the model outputs were selected:

- Strategy A: a single model for all the taxa that were present in more than 5% of the samples (65 taxa out of 174);
- Strategy B: a separate model for each taxon that was present in more than 5% of the samples (65 taxa out of 174);
- Strategy C: a single model for all the taxa present in more than 20% of the samples (19 taxa out of 174). Before adopting this strategy, we checked if the smaller subset of taxa preserved the information contained in the data set based on 65 taxa. A Principal Coordinate Analysis (PCO) (Gower 1966) using Jaccard's dissimilarity (Jaccard 1908) matrices and a Mantel test (1967) were carried out to compare the results obtained with 19 and 65 taxa.
- Strategy D: a single model for only 8 major taxa, which were selected on the basis of their ecological properties. In particular, we selected the taxonomic groups used for the computation of the Italian IBE index, namely Plecoptera, Ephemeroptera, Trichoptera, Gammaridae and Palaemonidae, Asellidae, Oligochaeta, the genus Leuctra, Baetidae and Caenidae.

Table 4.3.1 Environmental variables collected at each sampling site.

Environmental predictive variables	
elevation (m)	boulders (surface, %)
distance from source (km)	rocks (surface, %)
gradient (%)	pebbles (surface, %)
watershed drainage area (km^2)	gravel (surface, %)
water flow (score, 1-5)	sand (surface, %)
	silt and clay (surface, %)

The records available for both predictive and faunistic variables were divided into three subsets (training, validation and test). The training subset included 50% of the records (n=77), while the validation and test subsets contained 25% of the records each (n=38). The three subsets were defined according to a stratified procedure, using elevation as the stratification criterion. Therefore, each subset includes samples from sites at different elevations. Faunistic information was exploited at its simplest (and most reliable) level, i.e. as binary (presence/absence) data. All predictive variables, that include heterogeneous quantitative and semiquantitative environmental variables, were normalized into the [0,1] interval.

The composition of the benthic macroinvertebrate fauna was modelled using feed-forward multilayer perceptrons. The number of nodes in the hidden layer was defined after empirical tests and the structures of the ANNs that provided the best results are shown in Table 4.3.2. The validation subset was used to compute the mean square error (MSE) of the ANN after each epoch, whereas the test set was used to test the performance of the ANN after completion of the training procedure.

The learning procedure was iterated over 100 000 epochs, restarting the learning procedure each time the validation began to increase, and keeping the set of synaptic weights that provided the minimum validation error. In order to prevent overtraining, only a random subset of the training patterns (38 patterns) was submitted to the ANN at each training epoch, and white noise in the [-0.01,0.01] range was added to each input value at each epoch. Sigmoid activation functions were used in all the nodes of the hidden and output layers of the ANN, whereas the error back-propagation algorithm was selected for adjusting the ANN weights during the training procedure. The learning rate and the momentum were constant and set respectively to 0.90 and 0.10.

Table 4.3.2 Four different model outputs, corresponding to different modelling strategies were selected. The optimal ANN structure for each modelling strategy was defined after empirical tests.

MODEL OUTPUTS	MODELLING STRATEGY	ANN STRUCTURE
Only taxa present in more than 5% samples	1 model 65 outputs	11-19-65
Only taxa present in more than 5% samples	65 models 1 output each	11-5-1
Only taxa present in more than 20% samples	1 model 19 outputs	11-5-19
Only taxa involved in IBE index (Ghetti 1997) computation	1 model 8 outputs	11-14-8

The continuous ANN outputs, representing the probability of presence in a given site for each modelled taxon, were converted back to binary presence/absence estimates using a threshold function set to 0.5. The percentage of Correctly Classified Instances (CCI) was then computed for each modelling strategy and for each taxon, but a more reliable method for evaluating the accuracy of the models was needed. Therefore, the K statistic (Cohen 1960, Kraemer 1982) was also computed. In particular, this method tests the null hypothesis of independence between the modelled presence and absence data and the observed data. Finally, the modelling strategies were compared by computing Jaccard's dissimilarities (Jaccard 1908) between observed and modelled patterns (i.e. samples).

Results

Only data belonging to the independent test set, which was not used during the training and validation phases, were used for evaluating the performance and the accuracy of the different models. In the case of Strategy C the effects of the reduction of the number of modelled taxa from 65 to 19 were analysed by comparing the first Principal Coordinates (PCoo1) obtained from PCOs performed on the two data sets. The overall agreement between the two cases was very good, as shown in Fig. 4.3.1, and Spearman's rank correlation was highly significant (r=0.905, p<0.01). The Mantel test confirmed this result, as the null hypothesis of independence between the dissimilarity matrices was rejected (r=0.67, p<0.01). The results about the performances of the models trained according to the four different strategies are shown in Tables 4.3.3 to 6, in which the percentage of CCI, the values in the four confusion matrix cells and the K statistic are reported. Only taxa that were associated to significant K statistics, i.e. predicted by the ANN models in a way that was significantly different from random, have been included in the tables.

Table 4.3.3 Strategy A modelling results. Only 6 out of 65 taxa, which are associated to significant K statistics are shown.

Overall CCI: 83.8%						
Strategy A	CCI %	1-1	1-0	0-1	0-0	K
Baetis	76.32	19	4	5	10	0.50
Simuliidae	73.68	12	2	8	16	0.48
Elmidae	81.58	5	5	2	26	0.47
Hydropsychidae	78.95	6	5	3	24	0.46
Rhyacophilidae	76.32	5	7	2	24	0.38
Lumbricidae	68.42	7	8	4	19	0.31

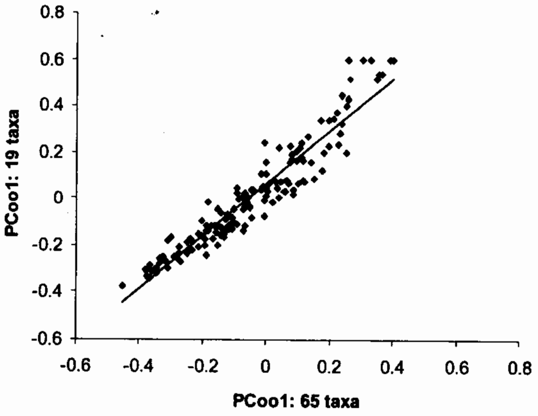

Figure 4.3.1 The overall structure of the data set including taxa that were present in more than 20% of the samples (19 taxa) was compared to that of the data set including taxa that were present in more than 5% of the samples (65 taxa). The information contained in the two data sets was similar, as shown by the agreement between the first Principal Coordinates obtained from Jaccard's dissimilarity matrices (Spearman's r=0.905, p<0.01).

Table 4.3.4 Strategy B modelling results. Only 10 out of 65 taxa, which are associated to significant K statistics are shown.

Overall CCI: 80.6%						
Strategy B	CCI %	1-1	1-0	0-1	0-0	K
Dinocras	89.47	5	1	3	29	0.65
Simuliidae	78.95	12	2	6	18	0.57
Rhyacophilidae	76.32	8	4	5	21	0.46
Baetis	73.68	22	1	9	6	0.39
Hydropsychidae	65.79	11	0	13	14	0.38
Ephemerella	71.05	7	5	6	20	0.34
Ceratopogonidae	68.42	8	6	6	18	0.32
Onychogomphus	84.21	3	3	3	29	0.41
Limoniidae	78.95	4	5	3	26	0.37
Limnephilidae	76.32	4	1	8	25	0.35

Table 4.3.5 Strategy C modelling results. Only 7 out of 19 taxa, which are associated to significant K statistics are shown.

Overall CCI: 68.3%						
Strategy C	CCI %	1-1	1-0	0-1	0-0	K
Baetis	73.7	19	4	6	9	0.44
Hydropsychidae	71.1	10	1	10	17	0.43
Elmidae	73.7	6	4	6	22	0.36
Gammaridae	68.4	9	9	3	17	0.36
Leuctra	76.3	4	8	1	25	0.35
Lumbricidae	68.4	8	7	5	18	0.32
Ceratopogonidae	68.4	8	6	6	18	0.32

Table 4.3.6 Strategy D modelling results. Only 3 out of 8 taxa, which are associated to significant K statistics are shown.

Overall CCI: 73.3%						
Strategy D	CCI %	1-1	1-0	0-1	0-0	K
Plecoptera	86.84	8	1	4	25	0.67
Ephemeroptera	73.68	23	7	3	5	0.33
Trichoptera	68.42	23	12	0	3	0.23

The comparisons between the results of the different modelling strategies, although based on a small number of efficiently predicted taxa, provided some useful hints. In particular, the comparison between Strategies A and B (Tables 4.3.3, 4.3.4) showed that the predictions were slightly more accurate when a set of models, one for each taxon, was trained instead of a single model simultaneously predicting all the taxa. In fact, with Strategy B not only was the number of taxa efficiently predicted larger (10 instead of 6 out of 65), but the average value of the K statistics for the predicted taxa was also slightly larger. This evidence, however, is not in agreement with the results for other groups of organisms, like fishes or diatoms (see sections 3.8 and 5.8), and is probably related to the smaller spatial scale at which the benthic macroinvertebrates respond to environmental conditions. Four taxa, namely Simuliidae, Rhyacophilidae, Hydropsychidae and the genus Baetis, were efficiently predicted both by Strategy A and by Strategy B, while two taxa (Elmidae and Lumbricidae) were efficiently predicted only by Strategy A, i.e. by using a single model for predicting all the species. Since information about interspecific interactions can only be embedded into this kind of model, it is possible that the success in modelling Elmidae and Lumbricidae depends on consistent association with other taxa or on the role that biotic interactions play in determining their distribution.

In Strategy C, a single model was trained for predicting the 19 taxa that were present in more than 20% of the samples. The accuracy of the predictions was not very different from the previous cases, as seven taxa had K statistics significantly different from zero (Table 4.3.5). Four of these taxa (Hydropsychidae, Elmidae, Lumbricidae and the genus Baetis) were also included among those that were efficiently predicted by the Strategy A model.

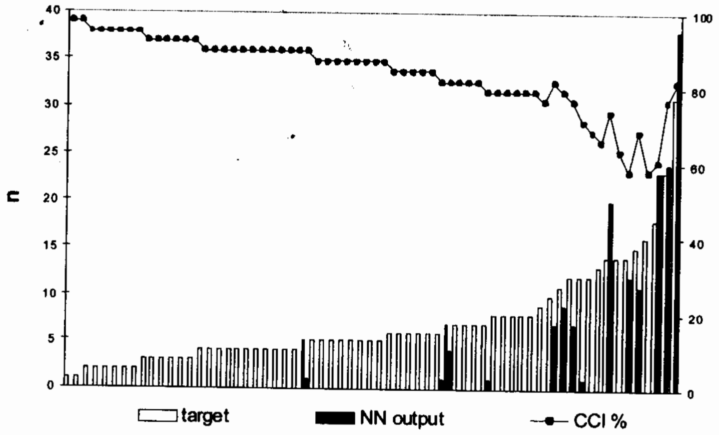

Figure 4.3.2 Strategy A: comparison between 65 ANN outputs (grey bars) and targets (white bars). The percentage of CCI is also shown (solid line with black dots).

Finally, eight taxa were modelled according to Strategy D (Table 4.3.6). In particular, these taxa are the ones that are routinely used for computing the IBE index. Obviously, these taxa have been considered for the biotic index because they have distinct ecological characteristics, and this is also the reason why they were selected as targets for ANN modelling. Three out of eight taxa were efficiently predicted by the ANN model, namely Plecoptera, Ephemeroptera and Trichoptera, but it is important to point out that these taxa are certainly the most sensitive to disturbance and pollution.

The difference between the apparently high CCI percentages, ranging from 83.8% to 68.6%, and the limited number of taxa that can be reliably predicted by the ANN models needs some explanation. In Figs. 4.3.2 to 5 the observed (target, white bar) and modelled data (ANN output, grey bar) are ranked according to the observed frequency of the taxa in the test set (n=38). It is obvious that the two bars are similar in height when the predicted values closely approximate the observed ones. The bars are not labelled to avoid clutter and because distinguishing each modelled taxon is not relevant in this case. The CCI percentage (solid line with black circles) was also plotted for each taxon.

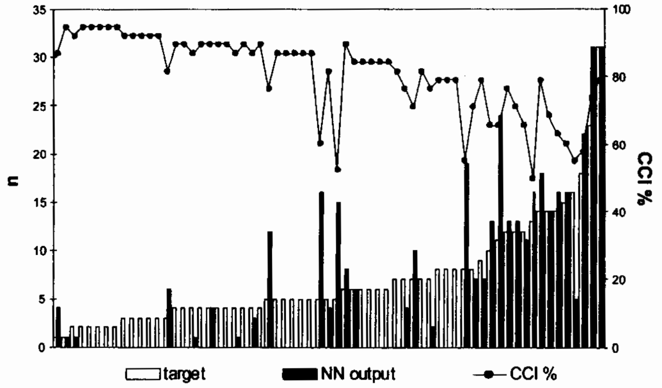

Figure 4.3.3 Strategy B: comparison between 65 ANN outputs (grey bars) and targets (white bars). The percentage of CCI is also shown (solid line with black dots).

When the number of taxa to be modelled was large (Strategies A and B, 65 taxa), the CCI percentage tended to be inversely correlated to the taxon frequency. This inverse relationship is a clear symptom of model malfunction, as the predictive ability of a model should not be related to the frequency of the taxa to be predicted. In fact, the models mainly failed in predicting the rarest taxa, and this bias was caused by the tendency of the ANNs to output only absence predictions when those taxa were considered.

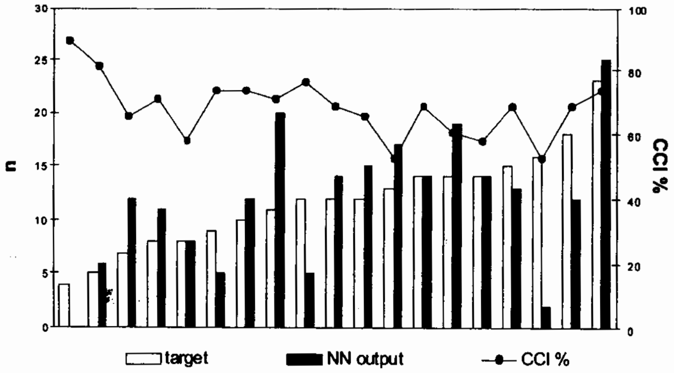

Figure 4.3.4 Strategy C: comparison between 19 ANN outputs (grey bars) and targets (white bars). The percentage of CCI is also shown (solid line with black dots).

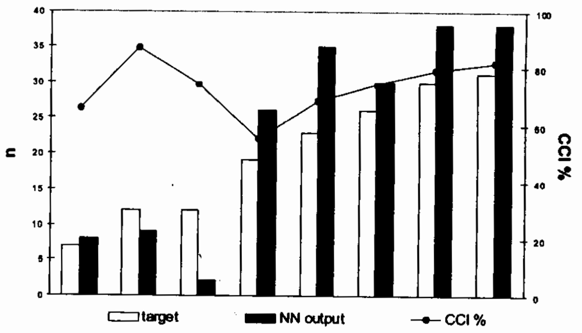

Figure 4.3.5 Strategy D: comparison between 8 ANN outputs (grey bars) and targets (white bars). The percentage of CCI is also shown (solid line with black circles).

In the case of Strategies C and D, in which only frequent taxa have been modelled, the inverse relationship between CCI percentage and taxon frequency was not observed, and the overall agreement between predicted and observed presence data was better than in the case of Strategies A and B. Therefore, these strategies were more effective in providing unbiased predictions about community structure, even though they obviously traded resolution for accuracy.

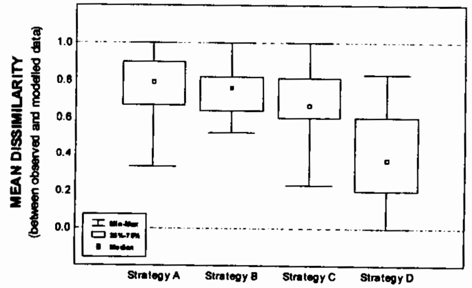

Figure 4.3.6 Comparison of the mean dissimilarities between observed and modelled samples for the four selected ANN training strategies.

Finally, observed and modelled data were compared using Jaccard's dissimilarities (Jaccard 1908) as a criterion for summarizing their resemblance. In Fig. 4.3.6 the distribution of these dissimilarities is shown, and it is evident that the closest match between modelled and observed data was obtained when the eight taxa that are used in the IBE index computations (Strategy D) were taken into account. In fact, the three lower quartiles in the dissimilarity distribution for the latter training strategy do not extend beyond the lowest quartile for the other training strategies (A, B and C). In other words, the similarity relationships that describe the structure of the test data set were closely reproduced by the ANN model when only a small number of ecologically significant taxa were selected as ANN outputs.

Conclusions

The benthic macroinvertebrate data set that was available for our study was certainly too small to support the development of accurate models. However, it provided a good opportunity for testing different training strategies and collecting useful hints for further developments. As in the case of other groups of organisms (see chapters 3 and 5), rare taxa (as well as very frequent ones, although the latter case is less likely to occur) could not be accurately predicted by the ANN models, independently of the training strategy. In fact, ANN models tend to "learn" that predicting only absence of rare taxa is the best solution for minimizing errors, even though this practice is obviously not appropriate for a real model. Obviously, the only solution to this problem would be a larger data set, but the way data are collected also plays a major role. In particular, more information is needed to model taxa that are insufficiently frequent. This goal can be attained, for instance, by planning the sampling activities at different spatial scales, thus allowing the collection of information about widely distributed taxa as well as about taxa that are only found in limited areas. It is obvious that a homogeneous spatial allocation of the sampling effort, although very convenient from a practical point of view, is not the best practice in this case. On the contrary, a multi-scale approach is needed, in which part of the samples are collected according to a ecoregional systematic sampling design, while other samples are collected in sub-areas where local maxima in beta diversity are detected. This way more information about the relationships between environmental variables and spatial distribution of benthic macroinvertebrates would probably be available.

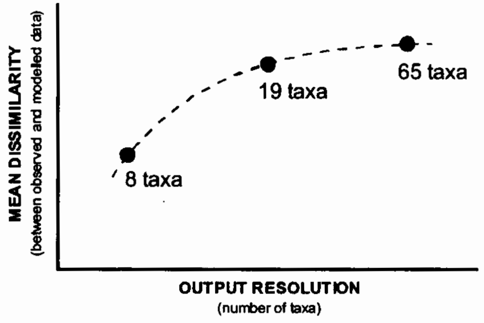

Figure 4.3.7 Three training strategies involving a single model for simultaneously predicting all the taxa are qualitatively compared. The comparison criterion is the mean dissimilarity between observed and predicted samples. The optimal efficiency of the modelling approach, corresponding to the minimum mean dissimilarity, was observed in the case of the modelling strategy based on the smallest set of taxa.

As for the different training strategies that have been tested, the macroinvertebrate fauna was predicted more efficiently when a set of single-taxon models was trained instead of a single model with multiple outputs. This result was not in agreement with previous findings obtained for other organisms (see chapters 3 & 5 in this book). Given the limited size of our data set, it is not easy to figure out whether this is a particular characteristic of benthic macroinvertebrate fauna or not. However, it is certainly possible that the lack of efficiency of the single model approach was somehow related to the complexity of underlying interspecific associations or interactions that were not adequately incorporated into a single ANN model.

Finally, the best training strategy among the ones we tested was based on very broad taxonomical units, namely on the taxa that are routinely used in the Italian IBE index computation. In particular, this approach was the one that gave the closest approximaiton of the observed structure of the dissimilarities among the samples in the test set. This result is not surprising, because the ecological characteristics of the taxa considered in the IBE index are certainly well defined. Therefore, they represent entities that are probably easier to model than others that are less closely related to the environmental variables. This result can be very useful in other ecoregions, where species that have been selected for other biotic indices could probably play a similar role in defining the structure of the macroinvertebrate assemblage. The different efficiencies, measured in terms of mean dissimilarity between observed and predicted data, of the three strategies involving a single model for the prediction of all the species is qualitatively shown in Fig. 4.3.7. It is obvious that the output resolution, i.e. the potential accuracy of the model, can be expressed as the number of modelled taxa, although the taxonomic level of the latter also plays a role. According to our results, the mean dissimilarity between observed and modelled data tends to increase with the output resolution, i.e. with the number of modelled taxa. Therefore, even though only a few cases have been considered in our study, our results support the hypothesis that the taxa to be modelled should be limited to the minimum set that provides the relevant information for correctly reproducing the relationships among observed samples.