ELSEVIER

Ecological Modelling 120 (1999) 247–260

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecomodel

Predictive models of collembolan diversity and abundance in a riparian habitat

Sithan Lek-Ang [a,*], Louis Deharveng [a], Sovan Lek [b]

[a] *LET-UMR 5552, CNRS-University Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 4, France*
[b] *CESAC-UMR 5576, CNRS-University Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 4, France*

Abstract

The artificial neural network (ANN) was used in this work for modelling the abundance and diversity of hydrophilous *Collembola* on the microhabitat scale. The procedure was applied to a Collembolan assemblage of the northern Pyrenees. Six variables were retained to describe its structure: abundance of the three dominant species, species richness, overall abundance of Collembola, and Shannon index. Seven environmental variables were selected as explanatory variables: distance to water, soil temperature, water content, and proportion of mineral soil, moss, litter and rotten wood in the substrate. Correlations between observed values and values estimated by ANN models of the six dependent variables were all highly significant. The ANN models were developed from 83 samples chosen at random and were validated on the 21 remaining samples. The role of each variable was evaluated by inputting fictitious configurations of independent variables and by checking the response of the model. The resulting habitat profiles depict the complex influence of each environmental variable on the biological parameters of the assemblage, and the non-linear relationships between dependent and independent variables. The main results and the ANN potential to predict biodiversity and structural characteristics of species assemblages are discussed.

Keywords: Biodiversity; Species richness; Community structure; Artificial neural network models; Multiple linear regression; Wet habitats

1. Introduction

Biodiversity conservation is a growing concern in western environmental policies. While species and habitats are disappearing at an alarming rate, we are however unable to evaluate, even roughly, the extent of this biodiversity loss, not to mention predicting it. In fact, estimating biodiversity is a tedious task when thousands of species may inhabit the same patch of forest, so taxonomist training would be advantageously coupled here with the development of forecasting techniques based on habitat characteristics, or on a subset of the overall biodiversity. Surprisingly, attempts to predict biodiversity on such grounds are scarce in the literature, except with a few animal groups such as birds (McArthur et al., 1966). Conversely, a wealth of works deal with abundance and biomass prediction (Verner et al., 1986), obviously

* Corresponding author. Fax: +33-5-61556196.
E-mail address: ang@cict.fr (S. Lek-Ang)

0304-3800/99/$ - see front matter
PII: S0304-3800(99)00106-4

in relation to their more direct socio-economic importance. There are a-priori no specific mathematical tools for predicting biodiversity, so the techniques used for predicting abundance also should work for biodiversity or any other measurable biological variable.

A lot of theoretical models have been proposed in this respect (McArthur et al., 1966; Fretwell, 1972; Tilman, 1982; Schoener, 1983) using a wide range of multivariate techniques, including several methods of ordination, canonical analysis, univariate and multivariate linear, curvilinear, and logistic regressions. A thorough and critical review by James and McCulloch (1990) shows that these conventional models, usually based on multiple regression, assume smooth, continuous, and either linear or simple polynomial relationships between variables. They are capable of solving many problems, but also have serious shortcomings since the main processes that determine the level of biodiversity or species abundance are often non-linear, whereas the methods are based on linear principles. Such models are for example not able to adequately reproduce the behaviour of real systems when very low or high values of the variables are considered (Lek et al., 1996b). Non-linear transformation of variables (logarithmic, power or exponential functions) may improve the results only to a limited extent. The artificial neural network (ANN) approach as proposed here emerges as a different and original methodology which is not constrained by assumptions about the type of relation between the studied variables (Rumelhart et al., 1986). The number of papers using ANN methodology published in ecological sciences has grown rapidly in recent years, e.g. modelling of greenhouse climate (Seginer et al., 1994), identification of the major goals of underwater acoustics (Casselman et al., 1994), prediction of density and biomass of brown trout redds (Lek et al., 1996a), prediction of density and biomass of trout (Baran et al., 1996; Lek et al., 1996b), prediction of the penetration of wild boar into cultivated fields (Spitz et al., 1996), prediction of phytoplankton production (Scardi, 1996), prediction of production/biomass (P/B) ratio of animal populations (Brey et al., 1996), and prediction of fish species richness on a global scale (Guégan et al., 1998), etc.

In the field of soil ecology, multiple linear regression (MLR)-based models relating environmental variables to community structure have been proposed by some authors (Boudjema et al., 1991) sometimes using non-linear transformations of independent or/and dependent variables to improve results (Vegter et al., 1988; Cancela Da Fonseca, 1991). Even so, the results have often remained insufficient, with a low percentage of variance explained. On the other hand, it has been shown that ANN can efficiently model non-linear systems in ecology (Lek et al., 1996b; Scardi, 1996). In the present study, we apply this method to relate the structure and diversity of an assemblage of hydrophilous *Collembola* to microhabitat characteristics. Hydrophilous *Collembola* often constitute the most abundant and diversified arthropods in a large range of wet habitats (Deharveng and Lek, 1995). As such, and because their specific richness is relatively constant along the year as long as water is present, they may provide an interesting raw material to evaluate predictive methods in population ecology. Though it does present sound data on the structure of hydrophilous assemblages of *Collembola*, this case study should be seen first as an attempt to develop predictive tools that are urgently needed for the study and the monitoring of biodiversity.

2. Material and sampling methods

2.1. Study sites and sampling

The studies were undertaken at the site of Ruau, located in the Northern Pyrenees (Arbon, Haute Garonne, France) at an altitude of 784 m. A small permanent spring used for watering livestock flows at the foot of a steep 3-m high slope covered with small trees. Above are large meadows on deep soils. We selected four transects perpendicular to the streamlet, each with four sample points at increasing distance from the water. The distance was 1.50 m between transects and 0.20–0.40 m between sample points on a transect, with the starting points 0–5 cm from the streamlet. Sampling was carried out every 2

months at 12–16 points from December 1993 to December 1994, for a total of 104 samples. The lowest points were sampled at all sampling periods, but the distal row was only occasionally sampled. Each sample was a substrate core of 125 cm^3. Extraction by Berlese technique lasted 2 weeks, until complete drying of the substrate. The animals were preserved in alcohol and sorted under the stereo-microscope. The Collembolan specimens not directly identifiable were mounted in Marc-André II after clearing in lactic acid, and examined with a Nachet 300 microscope under interferential contrast. After identification, the adult and juvenile individuals of each species were counted. After completion of the faunistic analysis, six variables describing the community structure were retained for each sample: abundance of *Collembola* (total number of specimens), species richness (total number of species), relative abundance of the three dominant species (*Isotomurus cassagnaui* [Icas], *I. prasinus* [Ipra] and *Brachystomella parvula* [Brp]), and Shannon diversity index (Table 1). Two of these species are strictly hydrophilous, while the third one, Brp, has both hydrophilous populations (Deharveng and Lek, 1995) and merely open habitat populations (Ponge, 1993).

Seven environmental variables were selected to describe the studied habitats (Table 1), on the basis of their known or supposed biological importance. Temperature and water content, which have a strong impact on most insects, including soil species (Boudjema et al., 1991; Argyropoulou et al., 1993; Deharveng and Bedos, 1993), both showed large fluctuations during the year at Ruau, with patterns varying with the spatial location of the sample points. The relative importance of mineral soil, litter, moss and rotten wood in the substrate has rarely been investigated so far (Deharveng and Lek, 1995), although it is a long established fact that specialized assemblages occupy each of these four substrates (Linnaniemi, 1907; Ponge, 1980; Weiner, 1981).

Distance to water and soil temperature were recorded in situ at the sampling points. Water content (= fresh weight − dry weight of the sample) was measured in the laboratory. The proportion of the different elements of the substratum (mineral soil, moss, litter and rotten wood) was visually estimated and assigned to five ordinal classes defined by their upper limits: absent (0), present up to 25% in volume (1), from 25 to 50% in volume (2), from 50 to 75% in volume (3), more than 75% in volume (4). Volumes were preferred to weights in this estimation because of

Table 1
Independent (i) and dependent (d) studied variables with methods of measurement

Variable	Type	Abbreviated	Methods of measurement
Distance to water	i	WAT	In situ, with ribbon centimetre
Temperature	i	TEM	In situ, with digital thermometer
Water content	i	HUM	Fresh weight-dry weight of substratum
Miner	i	MIN	Proportion of mineral soil in the substratum (visual estimated)
Moss	i	MOS	Proportion of moss in the substratum (visual estimated)
Decaying leaves	i	LIT	Proportion of dead leaves in the substratum (visual estimated)
Decaying wood	i	WOO	Proportion of rotten wood in the substratum (visual estimated)
Isotomurus cassagnaui	d	Icas	Number of *Isotomurus cassagnaui* in the sample (counted under binocular loupe)
Isotomurus prasinus	d	Ipra	Number of *Isotomurus prasinus* in the sample (counted under stereomicroscope)
Brachystomella parvula	d	Brp	Number of *Brachystomella parvula* in the sample (by counting in stereomicroscope)
Total abundance of *Collembola*	d	Nind	Count by stereomicroscope
Species richness	d	SR	Identification by stereomicroscope
Shannon index	d	SI	$SI = -\pi^*\log(\pi)$

the very large differences in density and spatial structure (i.e. spaces available for animals) of the substrates.

2.2. Technique of modelling

We analyzed our data set with: (i) the traditional method of multiple linear regression (MLR), to obtain a predictive model of reference; (ii) optimal non-linear transformation using the SAS Transreg procedure (SAS Institute, 1988; this procedure seeks an optimal transformation of variables, using a method of alternating last squares, a B-spline transformation); (iii) an artificial neural network (ANN) method, to evaluate the performance of this recent method in non-linear modelling. To compare these three methods the whole set of available data was used. To justify the predictive capacity of ANN and MLR methods, modelling was carried out in two steps. First, to fit the models, the matrix (104 records × 7 environmental variables) was used to perform the MLR, the alternating last squares and the ANN methods. The correlation coefficient between observed and predicted values was used to quantify the capability of models to produce the right answer through the training procedure. Second, to test the ANN models, we selected at random a training set (80% of the records, i.e. 83) and a validation set (20% of the records, i.e. 21). This operation was repeated three times giving rise to test 1, test 2 and test 3 which we studied by ANN and MLR. For each of the three sets, the model was determined with the training set and then validated with the test set. The quality of the model was judged through the correlation between observed and predicted values in the validation set.

For classical statistical analysis, univariate, bivariate and multivariate analyses were performed by the SPSS Software release 6.0 (Norusis, 1993). The univariate analyses estimated the mean, standard deviation, coefficient of variation, minimum, maximum, median and quartiles. In bivariate analyses we studied the correlation between variables using Pearson correlation coefficients (values and probabilities of significance at 5 and 1% of confidence intervals). In multivariate analyses, MLR procedures were applied. Examination of studentized residuals for normality, independence and homogeneity was used to test the validity of the models.

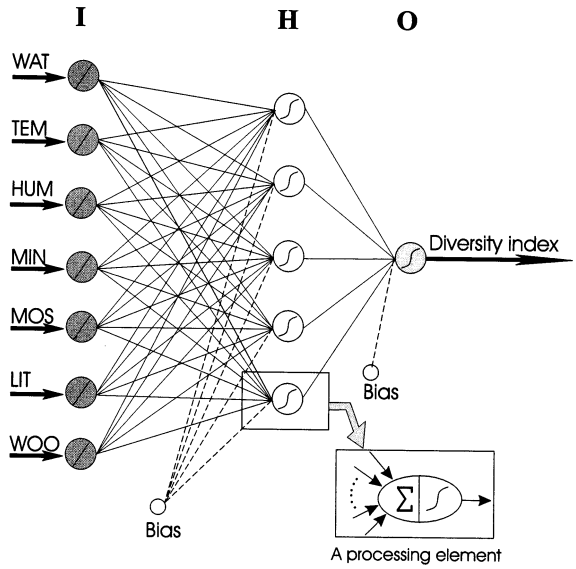

Fig. 1. Representation of the structure of the neural network used. Seven input nodes (I), five hidden layer nodes (H) and one output node (O) are shown. WAT, distance to water; TEM, soil temperature; HUM, water content in the substratum; MIN, proportion of mineral soil in the substratum; MOS, proportion of moss in the substratum; LIT, proportion of litter in the substratum; WOO, proportion of wood in the substratum.

For ANN modelling, the classic multilayer feed-forward neural network was used throughout the analyses. The processing elements in the network, called neurons are arranged in a layered structure (a typical three-layer network is shown in Fig. 1). The first layer, called the input layer, connects with the input variables. In our case, it comprises seven neurons corresponding to the seven habitat variables. The last layer, called the output layer, connects to the output variables. It comprises a single neuron which gives the value of the dependent variable to be predicted. The layers between the input and output layers are called the hidden layers. There can be one or more hidden layers and the number of neurons in each layer is an important parameter of the network. The network configuration is determined empirically by

testing various possibilities and selecting the one that provides the best compromise between bias and variance (Geman et al., 1992; Kohavi, 1995). In our study, a network with one hidden layer of five neurons was selected for each of the six dependent variables studied.

Each neuron is connected to all neurons of adjacent layers (neurons within a layer and in non-adjacent layers are not connected). Neurons receive and send signals through these connections. In feed-forward networks, signals are transmitted only in one direction: from input layer to output layer through hidden layers (no feed-back connections are permitted). Connections are given a weight which modulates the intensity of the signal they transmit.

Training the network consists in using a training data set to adjust the connection weights in order to obtain the best fit between expected and observed values. This training was performed according to the back-propagation algorithm (Rumelhart et al., 1986). The connection weights, initially taken at random in the range $[-0.3, 0.3]$, are iteratively adjusted by a method of gradient descent based on the difference between the observed and expected outgoing signals. Many iterations are necessary to guarantee the convergence of estimated values toward their expectations, without obtaining an overfit, i.e. incapability of the model to generalize (Smith, 1994). The computational program was realized in Matlab environment and computed with an Intel Pentium processor.

Input data have orders of magnitude that differ greatly according to the variables. So as to standardize the measurement scales, inputs were converted into standardized variables. The dependent variable was also scaled in the range [0…1] to adapt it to the demands of the transfer function used (sigmoid function).

2.3. Sensitivity of independent variables

A disadvantage of ANN in comparison with MLR models is their lack of explanatory power. MLR analysis can identify the contribution of each individual input in determining the output and can also give some measures of confidence for the estimated coefficients. On the other hand, there is currently no theoretical or practical way of accurately interpreting the weights in ANN. For example, weights cannot be interpreted as regression coefficients nor easily used to compute causal impacts or elasticities. Therefore, ANN are generally suited for forecasting or prediction rather than for explanatory analysis. But in ecology it is necessary to be able to explain the impact of the variables. To illustrate the importance of explanatory variables inside the ANN, Garson (1991) and Goh (1995) proposed a procedure for the partitioning of the neural network connection weights in order to determine the relative importance of the various input variables. Lek et al. (1995, 1996a,b) have built an algorithm allowing the visualization of the profiles of explanatory variables. In this work, an experimental approach has been used to determine the response of the model to each input variable separately by applying the technique described by Lek et al. (1996a,b).

3. Results

The 104 samples contained a total of 11 637 specimens of *Collembola* of which 11 312 were identified at species level, representing 55 species. Hydrophilous species were dominant in number with Icas, 2658 specimens i.e. 22.8% of the total, Ipra, 1272 specimens i.e. 10.9% and Brp, 1170 specimens i.e. 10%. However, Brp was present in a higher proportion of samples (66.35% occurrence) than the two other species (about 30% occurrence). Large variations in abundance of these three species were observed between samples (Table 2), with a high coefficient of variation (188, 237 and 309% for Brp, Icas and Ipra, respectively).

All samples contained *Collembola*. Mean species richness was 10.64 ($SD = 3.75$, $N = 104$). This is a low diversity compared to forest litter habitats in the same area where over 15 species are recorded on average for samples of the same size, but similar to values obtained in wet habitats at another Pyrenean site, the Arize mountain (9.2 with $SD = 4.50$ and $N = 60$, Deharveng and Lek,

1995). However, as the sample volume was only 125 cm^3 at Ruau, but 250 cm^3 in Arize, the former site is significantly richer, probably in relation to its lower elevation. Species richness and Shannon index were relatively stable with coefficients of variation below 37%. The abundance of *Collembola*, with a coefficient of variation of 84%, reflects fairly large seasonal and spatial fluctuations, but remains well under the variation level of the hydrophilous species of the assemblage.

Among environmental variables, much of the variation was due to the seasonal cycle (particularly temperature: 9.48°C (SD = 3.8), with a minimum of 3.9°C in February, and a maximum of 17.8°C in August). The largest variations were observed for litter and rotten wood content of the substrate (CV = 122 and 175%, respectively), independently of the season.

3.1. Correlation between assemblage characteristics and environmental variables

Among the environmental variables (Table 3), correlation coefficients are significant or highly significant in most cases but with relatively low values: only three correlations above $|0.5|$ ($P < 0.001$) were observed, involving MIN, MOS, HUM and WAT. Some correlations between independent and dependent variables are highly significant: Icas with HUM, WAT and MOS; Ipra with HUM and WAT; Nind with HUM and WAT; SI with WAT. Water content and distance to water therefore appear as major determinants of assemblage characteristics. Other correlations were significant at a lower level (Brp and TEM, Nind and MIN, SR and TEM, SI and TEM) and most (30) were not significant. In particular, species richness was very poorly related to environmental variables.

Among the dependent variables, a high correlation was found between Nind and the abundance of each of the two most abundant species ($r = 0.71$ for Icas, $r = 0.70$ for Ipra) indicating the numerical importance of these species in the community. The correlation was even higher between SI, the Shannon index, and SR, one of the measures on which it is built ($r = 0.76$, $P < 0.001$). Correlation between Icas and Ipra was relatively high ($r = 0.53$, $P < 0.001$) reflecting their strong dependence on water. Brp was conversely poorly related to Icas ($r = -0.10$, $P = 0.293$) or Ipra ($r = -0.10$, $P = 0.327$). Other highly significant correlations were between SR and Brp, between SI and Icas and between SR and Icas. Correlations were weaker with the other variables; the low value of correlation between species richness and the total number of Collembolan specimens is particularly noticeable.

Table 2
Summary statistics[a]

	Minimum	Q1	Median	Q3	Maximum	Mean	SD	CV%
TEM	3.9	6.5	8.45	11.95	17.8	9.48	3.8	40.08
HUM	3.6	25.4	34.5	46.55	89.4	36	18.86	52.39
WAT	0	2.5	20	50	100	35.77	31.89	89.15
MIN	0	2	2	2	4	1.9	0.78	41.05
MOS	0	1	1	2	3	1.21	0.89	73.55
LIT	0	0	0	1	3	0.6	0.73	121.67
WOO	0	0	0	1	2	0.28	0.49	175.00
Brp	0	0	2	12	113	11.25	21.18	188.27
Icas	0	0	0	4	296	25.56	60.69	237.44
Ipra	0	0	0	1	275	12.23	37.79	308.99
SR	1	8	11	13	21	10.64	3.75	35.24
Nind	1	53	74	149.5	561	111.89	94.23	84.22
SI	0	1.94	2.45	2.92	3.51	2.26	0.83	36.73

[a] SD, standard deviation; CV%, coefficient of variation in percentage; Q1, Q3, first and third quartile.

Table 3
Pearson correlation coefficient matrix between studied variables

	TEM	HUM	WAT	MIN	MOS	LIT	WOO	Brp	Icas	Ipra	SR	Nind
HUM	−0.41**											
WAT	0.01	−0.55**										
MIN	0.23*	−0.25*	0.36**									
MOS	−0.21*	0.47**	−0.50**	−0.63**								
LIT	−0.16	−0.18	0.16	−0.29**	−0.38**							
WOO	0.25*	−0.21*	0.11	0.02	−0.27**	−0.36**						
Brp	−0.25*	−0.10	0.08	−0.14	0.12	0.03	−0.03					
Icas	−0.06	0.46**	−0.45**	−0.15	0.26**	−0.09	−0.11	−0.10				
Ipra	−0.11	0.30**	−0.30**	−0.06	0.02	0.02	−0.11	−0.10	0.53**			
SR	−0.24*	0.00	0.14	−0.17	0.02	0.18	−0.05	0.35**	−0.28**	−0.04		
Nind	−0.15	0.36**	−0.32**	−0.24*	0.11	0.18	−0.12	0.22*	0.71**	0.70**	0.15	
SI	−0.22**	−0.14	0.30**	0.08	−0.17	0.15	−0.05	0.11	−0.52**	−0.14	0.76**	−0.16

* Significant ($P<0.05$).
** Highly significant ($P<0.01$).

Table 4
Multiple linear regression between parameters of Collembolan assemblages and environmental variables[a]x

	Brp	Icas	Ipra	SR	Nind	SI
TEM	−0.279*	0.277**	−0.083	−0.212	−0.001	−0.276*
HUM	−0.157	0.470**	0.187	−0.027	0.307*	−0.123
WAT	0.080	−0.196	−0.296*	0.186	−0.212	0.213
MIN	0.078	0.305	−2.016**	−0.917	−0.805	−0.410
MOS	0.300	0.375	−2.547**	−0.866	−0.942	−0.544
LIT	0.122	0.365	−1.903**	−0.652	−0.477	−0.379
WOO	0.120	0.174	−1.251**	−0.475	−0.437	−0.306

[a] The models shown the standard coefficients of seven independent variables with their significant level.
* Significant at 0.05.
** Significant at 0.01.

3.2. Multiple linear regression (MLR) analysis

For the 104 samples, the MLR procedure using the 7 independent variables gives the following coefficients of multiple correlation: with Brp, $R^2 = 0.10$ ($F_{7, 96} = 1.55$, $P = 0.17$); with Icas, $R^2 = 0.32$ ($F_{7, 96} = 6.53$, $P < 0.001$); with Ipra, $R^2 = 0.22$ ($F_{7, 96} = 3.91$, $P < 0.001$); with SR, $R^2 = 0.13$ ($F_{7, 96} = 2$, $P = 0.07$); with Nind, $R^2 = 0.24$ ($F_{7, 96} = 4.22$, $P < 0.001$); with SI, $R^2 = 0.16$ ($F_{7, 96} = 2.65$, $P = 0.02$). Low correlation coefficients reflect the low percentages of explained variance (less than 33% for all studied variables). With $\log(x + 1)$ transformation of variables, we obtained R^2 equal to, respectively 0.29, 0.64, 0.28, 0.31, 0.40 and 0.32 for Brp, Icas, Ipra, SR, Nind and SI. All models were highly significant ($P < 0.001$). Values of determination coefficients indicate a clear improvement of MLR models after non-linear transformation of variables. As this operation improves their linearity, we can conclude that non-linear relationships exist between the dependent and independent variables. Thus, a method based on alternating last squares was used to try to linearise the variables. With the Transreg procedure in SAS Software after maximum transformation of variables using the B-spline function, we obtained a squared multiple correlation equal to 0.41, 0.67, 0.47, 0.55, 0.49 and 0.48 for Brp, Icas, Ipra, SR, Nind and SI, respectively, i.e. a significant improvement of model quality.

Returning to the results of the MLR analysis, we give in Table 4 the standard coefficients of seven independent variables for the six dependent variables characterizing the Collembolan assemblage. Except in the SR model where none of the variables were significant, other models had at least one significant variable. The maximum was recorded for Ipra with five significant variables (Table 4).

3.3. Artificial neural network (ANN)

We used an ANN of one hidden layer of five neurons with seven independent variables, i.e. a 7-5-1 neural network (46 parameters in total: $7 \times 5 + 5 + 6$). Results after 500 iterations of the training procedure are presented in Fig. 2. The correlation coefficient (r) between observed and estimated values was close to 1 for Icas, Ipra, Brp and Nind ($r = 0.996$, $r = 0.965$, $r = 0.944$ and $r = 0.914$, respectively, $P < 0.001$). The lowest correlation coefficients were observed for SR and SI ($r = 0.847$ and $r = 0.872$, respectively, $P < 0.001$). The ANN therefore gave satisfactory results practically over the whole range of values of the dependent variables (Fig. 2). For the variables which represent species abundances (Icas, Ipra, Brp and Nind) most points were well aligned on the perfect fit diagonal (coordinates 1:1). Although poorly represented, the strong values of the output variable are clustered around this same perfect line. Only a few points lie far off, with some weak values slightly underestimated (Ipra

and Brp). For the remaining dependent variables SR and SI, which measure assemblage diversity, fitting is acceptable in spite of poorer results.

The sensitivity of the seven independent habitat variables on the six dependent variables obtained from ANN modelling is illustrated in Fig. 3. The 12 points cover the range of variation of each of the variables tested, with a class interval which was modified according to the variables. As illustrated in Fig. 3, we can distinguished seven sensitivity types:

- Exponential contribution: the independent variables contribute only at their low values. This is the case of WAT and MIN for Icas, and MOS, WAT and TEM for Ipra.

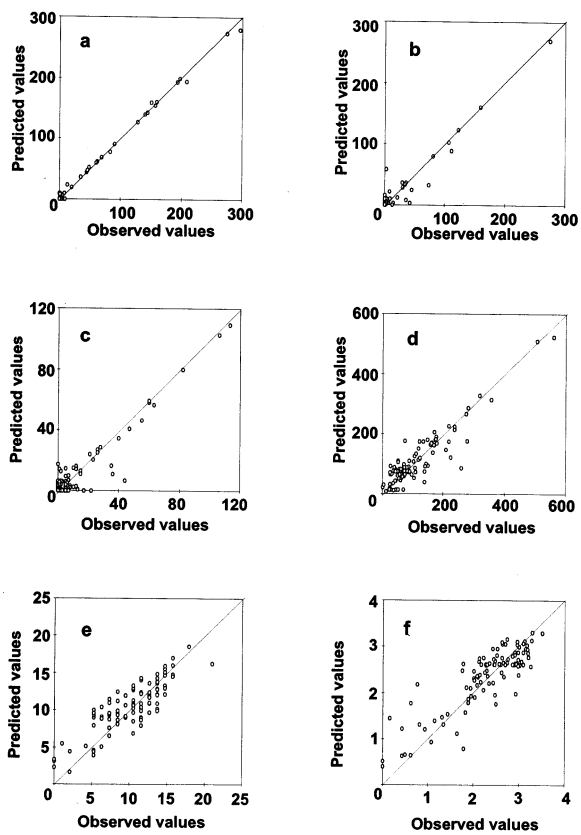

Fig. 2. Correlation graph between observed values and values estimated by the model. The solid line indicates the perfect fit line (coordinates 1:1). (a) Icas (*Isotomurus cassagnaui*); (b) Ipra (*Isotomurus prasinus*); (c) Brp (*Brachystomella parvula*); (d) Nind (total abundance of *Collembola*); (e) SR (species richness); (f) SI (Shannon index).

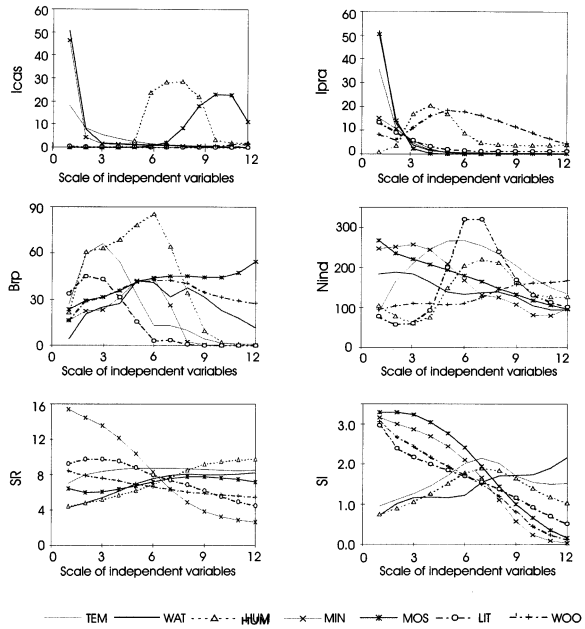

Fig. 3. Contribution profile for each independent variable to the determination of assemblage characteristics of *Collembola* fauna by ANN. Icas (*Isotomurus cassagnaui*), Ipra (*Isotomurus prasinus*), Brp (*Brachystomella parvula*), Nind (Total abundance of *Collembola*), SR (species richness), SI (Shannon index). The abscissa represents the 12 variation intervals of the independent variables between their minimum and their maximum.

- Gaussian contribution: the independent variable affects the dependent variable mostly around its average value, and has little influence at extreme values. This is the case of HUM for Icas, HUM and WOO for Ipra, WAT and WOO for Brp, LIT, TEM and HUM for Nind, and TEM and HUM for SI.
- Increasing contribution: dependent variable is low for low values of the independent variable and increases to a maximum at high values. This was observed for MOS for Brp abundance, HUM and WAT for SR, WOO for Nind, and WAT for SI.
- Decreasing contribution: dependent variable is relatively high at low values of the independent variable and decreases gradually afterwards. This was the case of TEM for Icas, MIN and LIT for SR, and MOS, MIN, WOO and LIT for SI.

- Skewed-to-the-left curve: the dependent variable is high only for high values of the independent variable. This sensitivity type was observed only for MOS to explain the abundance of Icas.
- Skewed-to-the-right curve: the dependent variable is high for low values of these independent variables; it decreases more or less rapidly afterwards to become virtually null thereafter. This contribution is present only for Brp abundance for four environmental parameters (TEM, HUM, LIT and MIN).
- Weak contribution: the contribution of the independent variable is very low, and not altered over its range, with a profile represented by a quasi-horizontal line. This is the case of WOO and LIT for Icas, and TEM, MOS and WOO for SR.

3.4. Test of the models

To test the variability, the prediction power of the different models determined from three training fractions was tested on three independent test fractions (Table 5). The lowest correlation between observed and predicted values was obtained for SR ($r = 0.66$–0.82, $P < 0.001$) and SI ($r = 0.79$, $P < 0.001$). Correlations for Nind and Brp ($r = 0.80$–0.87, $P < 0.001$ and $r = 0.84$–0.90, $P < 0.001$) were higher. The best results were obtained with Icas ($r = 0.95$–0.99, $P < 0.001$) and Ipra ($r = 0.88$–0.98, $P < 0.001$), like in the models based on the complete set of 104 samples. The same tests realized with MLR-models (Table 6) give clearly inferior results (maximum correlation coefficient equal to 0.57 for Icas in the second test set).

Table 5
Correlation coefficient between predicted and observed values by ANN models for three independent testing sets for the six studied parameters of Collembolan assemblages

	Training set			Testing set		
Set no.	1	2	3	1	2	3
Icas	0.986	0.990	0.990	0.990	0.950	0.956
Ipra	0.985	0.990	0.990	0.979	0.877	0.889
Brp	0.883	0.938	0.935	0.904	0.843	0.854
SR	0.864	0.851	0.834	0.700	0.656	0.823
Nind	0.940	0.946	0.906	0.865	0.820	0.797
SI	0.865	0.901	0.879	0.796	0.798	0.789

Table 6
Correlation coefficient between observed and predicted values by MLR-models for three independent testing sets for the six studied parameters of Collembolan assemblages

	Training set			Testing set		
Set no.	1	2	3	1	2	3
Brp	0.461	0.413	0.464	0.347	0.551	0.427
Icas	0.559	0.531	0.574	0.275	0.566	0.242
Ipra	0.630	0.582	0.610	0.161	0.515	0.287
SR	0.402	0.342	0.331	0.081	0.349	0.473
Nind	0.574	0.527	0.565	0.301	0.556	0.194
SI	0.494	0.460	0.426	0.053	0.266	0.511

On the whole, the coefficients in the training set were nearly identical to those of the models based on 104 samples. These results indicate a great stability (small standard deviations) of the prediction performance of the ANN models for different testing sets. The small decrease in performance in the test set compared to the training set can be related to the small size of the data set combined with the fact that each sample is likely to have some kind of unique information that is relevant to the model. The correlation coefficients were clearly not as low when the data were analyzed by MLR, in particular for Shannon index and specific richness.

4. Discussion

Two kinds of results emerge from this study: those related to the artificial neural network methodology and its ability to predict the characteristics of a species assemblage; and those related to the ecology of hydrophilous *Collembola*, which are of interest for Collembologists and wet habitat ecologists. MLR, spline regression and backpropagation of the ANN were applied on the same dataset with the aim to develop stochastic models of biodiversity prediction, using Collembolan assemblages and habitat features on a microhabitat scale. The backpropagation procedure of the

ANN gave much higher correlation coefficients than other methods. This may point to the predominantly non-linear relationships between the studied variables on the one hand, and on the other hand the ability of ANN to directly take into account any non-linear relationships between the dependent variables and each independent variable (Lek et al., 1996b). These results are in agreement with literature data, where performances of ANN have been repeatedly reported to overpass those of more traditional method such as MLR (Ehrman et al., 1996; Lek et al., 1996b; Scardi, 1996). However, the comparison between the predictive power of MLR and that of ANN is not quite fair, in particular as the number of parameters is different. In any case, ANN constitutes a new and powerful alternative in predictive ecological modelling, where poor fitting of biological characteristics to conventional models (mostly MLR) is often the rule.

Collembola are often the dominant group of Arthropods in wet habitats, yet literature references related to the ecology of hydrophilous species are scarce. On these grounds, it is hardly surprising that a large amount of novel information has been generated by the present study. Most characteristics of the Collembolan assemblages studied have been satisfyingly fitted to measured environmental parameters through ANN analysis. Variations in abundance among dominant species (Icas, Ipra, Brp) are in particular strongly connected to a set of environmental variables: temperature, distance to water, structure of the substratum and type of organic matter. A second important finding is the complexity of the response of Collembolan assemblages to changes in environmental parameters, so far largely overlooked in the relevant literature (e.g. Van Straalen, 1994). On the whole, emerging patterns of species abundance response to environmental parameter fluctuations appear both mostly non-linear and very heterogeneous, in spite of the high ecological similarity of the studied species. Some factors are clearly predominant, but they are not the same for the different measured biological variables. Conversely, sensitivity of species richness and Shannon index often follow similar patterns for different environmental variables, making it difficult to detect which one(s) is (are) the driving factors (Fig. 3). Nevertheless, there are some positive outcomes of this study, that are summarized below.

1. An unexpected result is that distance to free water has more impact than water content of the substrate for hydrophilous species abundance. Distance to water is weakly correlated to water content on the scale of our study because of its independence from season, local microtopography and superficial water circulation. *Isotomurus* species in particular experience an abrupt numerical decrease as soon as their distance to water increases. Their abundance peaks for medium-range water content. In contrast, *B. parvula* abundance reaches its maximum value at medium distance to water and medium water content, in agreement with empirical observations suggesting that its stenohygry is lower than that of *Isotomurus* (Deharveng and Lek, 1995). The overall abundance of Collembola follows yet another pattern which is likely to be explained by the strong impact of non-hydrophilous species, not documented in detail in this paper.
2. Distance to water has a slight but positive impact on biodiversity indices on our study scale, reflecting a more general trend of increasing species richness from water edge to mesophilous litter (Deharveng and Lek, 1995). The decreasing saturation of the mineral part of the substrate on this gradient gradually gives more micro-voids and new microhabitats for colonization by terrestrial mesofauna, and may contribute to the observed patterns.
3. Water content of the substrate is known to have an overwhelming importance for Collembolan populations (Vannier and Verhoef, 1978; Verhoef and Witteveen, 1980) but studies are lacking at the community level. According to Vegter et al. (1988), moisture heterogeneity has a strong influence on the abundance of epigeomorphic *Collembola*, but not on the assemblage structure. In our study, both abundance and assemblage structure were clearly affected by variations in water content of the substrate.

4. Surprisingly large differences were observed among species in response to variations of the studied variables (Fig. 3). The impact of temperature, the most documented environmental variable (Hopkin, 1997) is different on different ecological categories of *Collembola* as already stated in the literature (Van Straalen and Joosse, 1985; Van Straalen, 1994). In the present study, it further appears that, even among the same ecological category, species response may strongly vary between the most strictly hydrophilous species (Icas and Ipra) and those less so (Brp). The relationship of temperature to overall abundance of the hydrophilous *Collembola* assemblages still follows another pattern which is not that of these dominant species. The same comments also apply to other variables, particularly water content and mineral soil content of the substrate for which even the profiles of the two most hydrophilous species strongly diverge. A direct implication of these results is that extrapolation of ecological information from single, even dominant, species to communities may be strongly misleading: communities may be highly heterogeneous assemblages even at a relatively narrow functional level.
5. The relationships between environmental variables and biological parameters characterizing living communities have rarely been evaluated in the literature related to soil science. Boudjema et al. (1991) expresses these relations as a polynomial function, with pH and temperature as driving variables. Van Straalen (1994) reported a correlation between egg development and a measure of enzyme activity linked to temperature. But correlations, when measured, remained fairly low in all documented cases. The ecological profiles obtained from ANN models (Fig. 3) clearly exhibit the complexity and non-linearity of interacting processes, which may account for the difficulty in predicting species and community responses using traditional methods.
6. Intuitively, soil ecologists are aware of the prime importance of ligneous material for soil living assemblages, but this variable is rarely if ever taken into account in the literature. The same could be said for decaying leaves or moss content of the substrate. The classical measures of organic matter do not give any information on the relative proportion of these three elements, though it is likely to be of higher biological significance than overall amount of organic matter itself. By introducing these variables in our analysis, we expected to obtain some sound information about their influence on species abundance and assemblage structure. The results were, in fact, difficult to interpret. No influence was detected on the profiles of Icas, the most water-dependent species of the assemblage, and only a limited one on Ipra. The less strictly hydrophilous species Brp appeared more sensitive to these variables. Unexpectedly, increase in leaf litter and wood content of the substrate were associated with decreasing biodiversity, in apparent contrast to the usual (but again poorly documented) trend of increasing biodiversity from open to forested habitats. An appealing hypothesis is that leaf and wood litter is less important in wet habitat than in mesophilous habitats, because decomposition processes are less active in water or water-saturated substrate, providing a lower diversity of fungal species on which most *Collembola* feed.

Is it finally possible to predict the level of biodiversity of a living group from environmental variables? Because they largely control the presence and abundance of individual species, environmental variables necessarily contribute to the control of community structure, hence of biodiversity. Hard data, are however, lacking to support this commonplace statement in soil ecosystems and previous attempts to detect simple and linear relationships between edaphic factors and diversity have failed to give clear-cut results (for instance in tropical Collembolan assemblages, Deharveng and Bedos, 1993). Three reasons (or a combination of these) may explain this failure: (i) the pertinent variables have not been identified; (ii) interactions between species play a major role; and (iii) relationships between abiotic factors and biodiversity are non-linear. This last hypothesis was considered here. The results obtained indicate that species interaction, or consideration of addi-

tional variables, is not needed to satisfactorily predict the characteristics of the observed assemblage patterns. Several parameters relevant to biodiversity were efficiently predicted by the ANN-based models in the studied community. Additional data sets, experimental manipulations and repeated mathematical analyses would be necessary to assess this first result more firmly, but the ANN has demonstrated here a promising potential in the field of community ecology, as a tool to evaluate, understand, predict and manage biodiversity.

Acknowledgements

This work was supported by a grant from the European Community (contract DGXII níPL93-1917: High Endemism in Areas, endemic biota and the conservation of biodiversity in Western Europe).

References

Argyropoulou, M.D., Asikidis, M.D., Iatrou, G.D., Stamou, G.P., 1993. Colonization patterns of decomposing litter in a maquis ecosystem. Eur. J. Soil Biol. 29, 183–191.

Baran, P., Lek, S., Delacoste, M., Belaud, A., 1996. Stochastic models that predict trout population densities or biomass on macrohabitat scale. Hydrobiologia 337, 1–9.

Boudjema, G., Julien, J.M., Sarkar, S., Cancela Da Fonseca, J.P., 1991. Etude par analyse statistique multilinéaire de l'impact des facteurs physico-chimiques sur l'abondance des Microarthropodes édaphiques d'une forêt de mousson en Inde orientale. Rev. Ecol. Biol. Sol. 28, 303–322.

Brey, T., Jarre-Teichmann, A., Borlich, O., 1996. Artificial neural network versus multiple linear regression: predicting P/B ratios from empirical data. Mar. Ecol. Progr. Ser. 140, 251–256.

Cancela Da. Fonseca, J.P., 1991. Ecological diversity and ecological systems complexity: local or global approach? Rev. Ecol. Biol. Sol. 28, 51–66.

Casselman, F.L., Freeman, D.F., Kerrigan, D.A., Lane, S.C., Magley, D.M., Millstrom, N.H., Roy, C.R., 1994. A neural network-based underwater acoustic application. In: Proceedings of the IEEE International Conference on Neural Networks. IEEE, Orlando. pp. 3409–3414.

Deharveng, L., Bedos, A., 1993. Factors influencing diversity of soil *Collembola* in a tropical mountain forest (Doi Inthanon, Northern Thailand). In: Paoletti, M.G., Foissner, W., Coleman, D. (Eds.), Soil Biota, Nutrient Cycling and Farming Systems. Lewis, pp. 91–111.

Deharveng, L., Lek, S., 1995. High diversity and community permeability: the riparian *Collembola* (Insecta) of Pyrenean massif. Hydrobiologia 312, 59–74.

Ehrman, J.M., Clair, T.A., Bouchard, A., 1996. Using neural networks to predict pH changes in acidified Eastern Canadian lakes. Artif. Intell. Appl. 10, 1–8.

Fretwell, S.D., 1972. Populations in a Seasonal Environment. Monography. Population Biology, vol. 5. Princeton University, Princeton NJ, p. 217.

Garson, G.D., 1991. Interpreting neural-network connection weights. Artif. Intell. Expert 6, 47–51.

Geman, S., Bienenstock, E., Doursat, R., 1992. Neural networks and the bias/variance dilemna. Neural Comput. 4, 1–58.

Goh, A.T.C., 1995. Back-propagation neural networks for modelling complex systems. Artif. Intell. Eng. 9, 143–151.

Guégan, J.F., Lek, S., Oberdorff, T., 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature 391, 382–384.

Hopkin, S.P., 1997. Biology of the Springtails (Insecta: *Collembola*). Oxford University, Oxford, p. 330.

James, F.C., McCulloch, C.E., 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box? Ann. Rev. Ecol. Syst. 21, 129–166.

Kohavi, R., 1995. A study of cross-validation and bootstrap for estimation and model selection. In: Proceeding of the 14th International Joint Conference on Artificial Intelligence. Morgan Kaufmann, Montreal. pp 1137–1143.

Lek, S., Belaud, A., Lauga, J., Dimopoulos, I., Moreau, J., 1995. Improved estimation, using neural networks, of the food consumption of fish populations. Mar. Fresh Res. 46, 1229–1236.

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996a. Role of some environmental variables in trout abundance models using neural networks. Aquat. Living Resour. 9, 23–29.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulanier, S., 1996b. Application of neural networks to modelling non-linear relationships in ecology. Ecol. Mod. 90, 39–52.

Linnaniemi, W.M., 1907. Der Apterygotenfauna Finlands. I. Allgemeiner Teil. Acta Soc. Sci. Fen. 34, 1–134.

McArthur, R.H., Reicher, H., Cody, M.L., 1966. On the relation between habitat selection and bird species diversity. Am. Nat. 100, 319–332.

Norusis, M.J., 1993. SPSS for Windows. Base system user's guide release 6.0. SPSS Inc, pp. 828.

Ponge, J.F., 1980. Les biocénoses des Collemboles de la for_t de Sénart. In: Pesson (Ed.), Actualités d'Ecologie Foresti_re P. Gauthier Villard, Paris, pp. 151–176.

Ponge, J.F., 1993. Biocenose of *Collembola* in Atlantic temperate grass-woodland ecosystems. Pedobiologia 37, 223–244.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating error. Nature 323, 533–536.

SAS Institute, 1988. SAS Technical report P-179. Additional SAS/STAT procedure, release 6.03. SAS Institute, Cary, NC, pp. 255.

Scardi, M., 1996. Artificial neural networks as empirical models for estimating phytoplankton production. Mar. Ecol. Progr. Ser. 139, 289–299.

Schoener, T.W., 1983. Field experiments on interspecific competition. Am. Nat. 122, 240–285.

Seginer, I., Boulard, T., Bailey, B.J., 1994. Neural network models of the greenhouse climate. J. Agr. Eng. Res. 59, 203–216.

Smith, M., 1994. Neural networks for statistical modelling. Van Nostrand Reinhold, New York.

Spitz, F., Lek, S., Dimopoulos, I., 1996. Neural network models to predict penetration of wild boar into cultivated fields. J. Biol. Syst. 4, 433–444.

Tilman, D., 1982. Resource competition and community structure. In: Monography Population Biology, vol. 17. Princeton University, Princeton NJ, p. 296.

Van Straalen, N.M., 1994. Adaptive significant of temperature responses in *Collembola*. Acta Zool. Fenn. 195, 135–142.

Van Straalen, N.M., Joosse, E.N.G., 1985. Temperature responses of egg production and egg development in two species of *Collembola*. Pedobiologia 28, 265–273.

Vannier, G., Verhoef, H.A., 1978. Effect of starvation on transpiration and water content in the populations of two coexisting *Collembola* species. Comp. Biochem. Physiol.(A) 60, 483–489.

Vegter, J.J., Joosse, E.N.G., Ernsting, G., 1988. Community structure, distribution and population dynamics of *Entomobryidae* (*Collembola*). J. Anim. Ecol. 57, 971–981.

Verhoef, H.A., Witteveen, J., 1980. Water balance in *Collembola* and its relation to habitat selection; cuticular water loss and water uptake. J. Insect Physiol. 26, 201–208.

Verner, J., Morrison, M.L., Ralph, C.J., 1986. Wildlife 2000: Modelling Habitat Relationships of Terrestrial Vertebrates. Wisconsin University, Madison WI, p. 478.

Weiner, W.M., 1981. Collembola of the Pieniny national park in Poland. Acta Zool. Cracoviense 25, 417–500.