PERGAMON

Soil Biology & Biochemistry 32 (2000) 197–209

Soil Biology & Biochemistry

www.elsevier.com/locate/soilbio

Application of machine learning techniques to the analysis of soil ecological data bases: relationships between habitat features and Collembolan community characteristics

Christian Kampichler[a,*], Sašo Džeroski[b], Ralf Wieland[c]

[a]GSF National Research Centre for Environment and Health, Institute of Soil Ecology, Ingolstädter Landstrasse 1, D-85758 Neuherberg, Germany
[b]Department of Intelligent Systems, Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia
[c]ZALF Centre for Agricultural Landscape and Land Use Research, Institute of Landscape Modelling, Eberswalder Strasse 84, D-15374 Müncheberg/Mark, Germany

Accepted 31 July 1999

Abstract

We applied novel modelling techniques (neural networks, tree-based models) to relate total abundance and species number of Collembola as well as abundances of dominant species to habitat characteristics and compared their predictive power with simple statistical models (multiple regression, linear regression, land-use-specific means). The data used consisted of soil biological, chemical and physical measurements in soil cores taken at 396 points distributed over a 50 × 50 m sampling grid in an agricultural landscape in southern Germany. Neural networks appeared to be most efficient in reflecting the nonlinearities of the habitat–Collembola relationships. The underlying functional relations, however, are hidden within the network connections and cannot be analyzed easily. Model trees — next in predictive power to neural networks — are much more transparent and give an explicit picture of the functional relationships. Both modelling approaches perform significantly better than traditional statistical models and decrease the mean absolute error between prediction and observation by about 16–38%. Total carbon content and measurements highly correlated with it (e.g. total nitrogen content, microbial biomass and respiration) were the most important factors influencing the Collembolan community. This is in broad agreement with existing knowledge. Apparent limitations to predicting Collembolan abundance and species number by habitat quality alone are discussed.

1. Introduction

Collembola are by far the most abundant insects in soil and attain high densities of up to several 100,000 individuals m^{-2}. They can exert a significant influence on mineralization processes and nutrient cycling via trophic interactions with decomposer microorganisms (Verhoef and Brussaard, 1990; Lussenhop, 1992). Soil zoologists have been trying to discern the factors governing the distribution of Collembola in various spatial scales, spanning from biogeographic to local and microhabital scales (Hopkin 1997, p. 174). Characteristically, these analyses are performed by various ordination techniques. The sampling sites are represented by points in two-dimensional space, whereby sites similar in species composition are arranged closer together than sites that are dissimilar. The resulting ordination diagram can be interpreted by whatever is known about the environmental characteristics of the sampling sites (indirect gradient analysis), or, if environmental data have been collected, the interpretation is performed in a formal way (Jongman et al., 1995). Collembolan communities have been analyzed in agroecosystems (e.g. Dekkers et al., 1994; Kovác and Miklisová, 1997) as well as in natural ecosystems (e.g. Ponge, 1993; Ponge et al., 1993). A number of

* Corresponding author. Present address: Free University Berlin, Institute of Biology, Grunewaldstrasse 34, D-12165 Berlin, Germany. Tel.: +49-30-838-3948; fax: +49-30-838-3886.
E-mail address: kampichl@zedat.fu-berlin.de (C. Kampichler).

0038-0717/00/$ - see front matter
PII: S0038-0717(99)00147-9

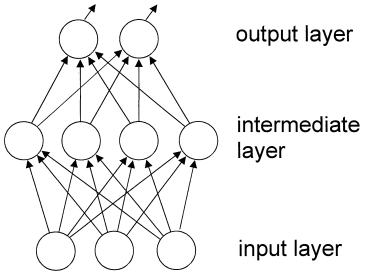

Fig. 1. Topology of a sample backpropagation network with three input nodes, four hidden nodes and two output nodes.

factors have been found to be potentially responsible for Collembolan distribution patterns and include, among others, temperature, soil acidity, soil moisture, as well as characteristics of the leaf–litter layer or of the fungal community (Hopkin, 1997, p 175). The degree of explanation of distribution patterns by these analyses is typically moderate (Klironomos and Kendrick, 1995). Vague relationships between environmental characteristics and characteristics of the Collembolan community (abundance, species composition) are most probably a consequence of the highly aggregated small-scale distribution patterns of the Collembola; a fact which, even at macroscopically uniform sites, leads to a high sample variance (Butcher et al., 1971; Ekschmitt, unpubl. PhD thesis, 1993).

Ordination is certainly a suitable instrument for exploratory analysis and interpretation of relationships between variables in soil ecological datasets, but it does not offer a possibility of creating predictive models. We think, however, that predictive models are an important step towards understanding abundance and distribution patterns of animals. For this purpose, it is necessary to explicitly relate the values of the dependent variable (e.g. characteristics of the Collembola community) to the values of the independent variables (e.g. environmental features); the values of the dependent variables may then be inferred from datasets of the independent variables. The correspondence of predicted with observed values of the dependent variables must be the criterion used for deciding whether there is a generalizable relationship between the independent and dependent variables (=predictive validation *sensu* Rykiel, 1997). This is a major task if large databases with a high dimensionality are to be analyzed and ecologists are becoming increasingly aware of the advantage of novel and advanced machine learning techniques, such as artificial neural networks or the induction of tree-based models.

Briefly, neural networks (NN) consist of a number of computing units, which are termed *cells* or *nodes* (see Gallant (1993) for an introduction to NNs). In the most popular type of NN, the *backpropagation network*, the nodes are typically organized in a layered structure (Fig. 1): an input layer, whose nodes represent the independent variables; an intermediate layer (or more) of so-called *hidden nodes*; and an output layer, whose nodes represent the dependent variables. All nodes are unidirectionally joined and each connection has a numerical *weight*. Along these connections, the input values are propagated through the network. A NN can be trained by providing it with training patterns, e.g. patterns with known outputs for a given input. The NN begins with a randomly chosen set of connection weights, compares the calculated output of the first training pattern with the desired output and, by applying the *backpropagation algorithm*, propagates the error backwards through the net. In the end of a backpropagation step, the connection weights are slightly altered and the output approaches by a small step the desired output. NNs are universal approximators since, by providing a sufficiently complex network iteratively with a large number of training patterns, it is able to model any differentiable relationship (Warner and Misra, 1996). This is particularly important for ecological modelling since standard methods of relating independent and dependent variables, such as multiple regression, can only insufficiently cope with the nonlinearities of ecological systems. Recently, NNs have successfully been used in ecology, e.g. for the modelling of trout spawning sites in small rivers (Lek et al., 1996), the modelling of algal blooms (Recknagel et al., 1997), or the prediction of water balance factors in soil (Schaap and Leij, 1998). These studies have shown that NNs can fit the complexity and nonlinearity of ecological phenomena to a high degree. On the other hand, NNs are not transparent, especially when many independent variables are used, since the relationship between independent and dependent variables is not explicitly stated anywhere, but represented in the matrix of connection weights of the network. Thus, if the actual relationships between variables are of interest, a trained NN must be analyzed by keeping one input variable constant and scanning over the possible values of the others. This is a long-lasting process and the limits of feasibility of an analysis of this kind are easily reached since the number of possible permutations is m^n with n being the number of input nodes and m being the number of levels of each input variable (but see Dimoupoulos et al. 1995 for specific algorithms to detect the influences of input variables).

Tree-based models offer a more transparent means of analyzing high-dimensional datasets (see Breiman et al. (1984) for an introduction to tree-based models). Briefly, they are based on the assumption that the relationship between independent and dependent variables is not constant over the entire range of possible variable values, but can be approximated in smaller subdomains. Tree-based models are thus constructed by splitting the dataset into subsets based on the mini-

Table 1
Analytical methods and data range of the environmental variables from the Scheyern experimental farm (n = 195)

Variable	Method	Range
Microbial respiration	automatic IRGA[a]	0.76–10.03 μg CO_2 g^{-1} soil dry weight h^{-1}
Microbial biomass	automatic IRGA[a]	4.97–80.98 μg CO_2 g^{-1} soil dry weight h^{-1}
Soil acidity	$CaCl_2$	4.72–7.04 $-\log[H^+]$
Soil moisture	drying at 105°C	5.1–48.13% soil dry weight
C_t	elemental analyzer[b]	0.96–5.86% soil dry weight
N_t	elemental analyzer[b]	0.079–0.833% soil dry weight
Median particle diameter	sieving[c], sedimentation[d], laser diffraction[e]	6.3–630 μm[f]

[a] After Heinemeyer et al. (1989).
[b] Oxidization at 1020°C in a Carlo Erba NA 1500 and subsequent determination of the gas concentrations of CO_2 and N_2 by gas chromatography and heat conductivimetry.
[c] After Hartge and Horn (1989).
[d] After Köhn (1928).
[e] After Heuer and Leschonski (1985).
[f] Class limits at 12.5, 20, 28, 40, 50, 63, 100 and 200 μm.

mization of a variance criterion for the subsets. This procedure is iteratively repeated for each subset and results in a tree-like structure where each branch is defined by a certain range of values of the independent variables. The endpoints of the branches, the *leaves*, show average values of the dependent variable or relate the dependent variable to the independent variables by multivariate linear models. In an ecological context, tree-based models were used for the prediction of algal blooms by Kompare and Džeroski (1995).

In this paper, we present the analysis of a comprehensive dataset on Collembola and on soil chemical, physical and biological variables from the FAM Research Network on Agroecosystems in southern Germany. The data provided a unique opportunity to test the predictive potential of soil habitat features for selected measurements of the Collembolan community: (1) we model total abundance of Collembola, species number of Collembola and abundance of the dominant species of the community by means of neural networks and tree-based models; (2) we compare the predictive power of these models to simpler predictive models generated by standard statistical methods (univariate linear regression, multiple linear regression, land-use-specific means); and (3) we discuss the limits of models explaining features of local Collembolan assemblages in terms of habitat characteristics.

2. Material and methods

2.1. Study site, sampling design and data treatment

The FAM Research Network on Agroecosystems runs a 153 ha experimental farm in Scheyern, approx. 40 km N of Munich, southern Germany. It is located at an elevation of 450–490 m above sea level; mean annual temperature and mean annual precipitation are 7.5°C and 833 mm, respectively. In April 1991, one soil core was taken at each intersection of a 50 × 50 m mesh-size grid (7.8 cm dia, 5 cm depth) and yielded a total of 396 cores. The majority of these points were situated in arable fields (n = 302), the remainder in pastures, meadows and arable fields on former hop fields. All arable land (except grassland) was uniformly grown with winter wheat. Microarthropods were counted and Collembola identified by species (Fromm, 1997). For the measurement of the following environmental factors, cores were taken from the same sampling points at a distance of approximately 25 cm from the first cores: microbial biomass, microbial respiration, soil moisture, soil acidity, carbon content (C_t) and nitrogen content (N_t) (Winter, 1998). Soil texture at the sampling points was determined by Sinowski (1994) and expressed by the $_{10}$logarithm of the median particle diameter (log mpd). From the 396 cores, only those that had no missing values for any of these variables were included in the model development, leaving a dataset of n = 195 (155 of which to be found in arable fields). Table 1 summarizes the methods and observed values for the independent variables (= soil habitat features). All data were drawn from the FAM data base at the GSF National Research Centre for Environment and Health in Neuherberg, Germany (URL: http://www.gsf.de/FAM/adis.html). Each soil core was also characterized by its location on the experimental farm (position on the axes of the 50 × 50 m sampling grid) and by the type of land use (agricultural fields, meadows, pastures and agricultural fields on former hop fields).

2.2. Data used for modelling

Only data of euedaphic and hemiedaphic Collembola were included in the analysis; epigeic and atmobiotic Collembola were not used in this study.

Species number of euedaphic and hemiedaphic Collembola — throughout this paper simply termed 'species number' — ranged from 0 to 10 (median 3) per core, total abundance of euedaphic and hemiedaphic Collembola — throughout this paper simply termed 'total abundance' — ranged from 0 to 169 (median 10) individuals per core and abundance of the two dominant species, *Onychiurus armatus* (40.1% of total abundance) and *Folsomia quadrioculata* (17.1% of total abundance), ranged from 0 to 153 (median 3) and from 0 to 95 (median 0) individuals per core, respectively. The frequency distribution of abundance was highly skewed, with very few cores containing very high numbers of individuals, a familiar pattern typical for the distribution of soil microarthropods (e.g. Debauche, 1962). We assume that the outliers were due to factors other than habitat quality, for example, aggregation behavior at the microscale or large numbers of juveniles in a core after hatching from an egg batch. We thus established a threshold at the 95% quantile of the abundance data distribution. The threshold value was assigned to all cores with abundance values beyond the 95% quantile (45 for total abundance, 16 for the abundance of *F. quadrioculata* and 23 for the abundance of *O. armatus*). There were no other data transformations.

The dataset of 195 cores was randomly split into ten subsets. Subsequently one subset was left aside and the model was generated with the remaining nine subsets. The subset left aside was used for testing the model for its predictive power. Machine learning and statistics have different terminologies. To prevent any ambiguity, the terms we use are defined as follows:

- the values of the soil–habitat variables and the corresponding values of the Collembolan community characteristics at a single sampling point constitute a *pattern*;
- the patterns of the subsets used for *model development* (least-squares fitting in the case of linear regression, backpropagation training in the case of the neural network, etc.) are termed *development patterns*;
- the patterns of the subsets used for *testing the models* are termed *test patterns*;
- the soil–habitat characteristics are termed *input variables*;
- the community characteristics of the Collembola are termed *output variables*.

2.3. Model development

2.3.1. Neural networks

For designing, training and testing the networks, we used the SNNS Stuttgart Neural Network Simulator (Zell et al., 1995). Prior to neural network modelling, all data were projected into the interval [0 1] by applying the transformation $x_i' = (x_i - x_{\min})/(x_{\max} - x_{\min})$, with $x_{\min}$ being the smallest and $x_{\max}$ being the largest value of a variable in the development patterns. The network topology consisted of one output node and one intermediate layer of hidden nodes, with the number of hidden nodes equalling the number of input nodes (ten hidden nodes for models with ten input variables, three hidden nodes for models with three input variables, see below). We used backpropagation learning, currently the most widely used algorithm in neural-network learning (see Appendix A for a description). The learning parameter ρ was chosen between 0.1 and 0.35. SNNS allows the change of an error term of the test patterns to be followed while training the network with the development patterns. Learning was stopped when the error of the test set was minimal. The number of learning cycles varied between 5 and 100,000. For every output variable (total abundance, species number, abundance of *F. quadrioculata* and of *O. armatus*), a separate network was trained.

2.3.2. Tree induction

For the development of tree-based models, we used the program M5 by Quinlan (1992, 1993). A commercial version of M5, the program Cubist 1.05, can be downloaded from the URL http://www.rulequest.com. It can work in three different ways: (1) using only instance-based learning, (2) using only regression and (3) using instance-based learning and regression for tree construction. We used all three possibilities, thus creating models with instances only (MI), regression trees (RT) and model trees (MT). A closer description of the tree-induction approach and the functioning of the M5 algorithm is presented in Appendix B.

2.3.3. Statistical models

We compared the machine learning models with three predictive models derived by standard statistics: multiple linear regression (MR), univariate linear regression (LIN) and land-use-specific means (LSM). For the LIN model, first, the input variable that had the highest correlation coefficient with the output variable was determined and, second, the linear regression between these two variables was calculated. For the LSM model the average of the specific output variable in the development patterns was determined for four different types of land use: agricultural fields, meadows, pastures and agricultural fields on former hop fields. The calculated average was used as a predictor for the test patterns.

For clarity, the models and their abbreviations are listed as follows:

NN neural network model

Table 2
Mean absolute errors (MAE) of predictive models for total abundance of hemi- and euedaphic Collembola, for species number of hemi- and euedaphic Collembola and for numerical abundance of *F. quadrioculata*, based on ten selected independent variables (x and y coordinates of sampling point, land use type, microbial biomass, microbial respiration, soil moisture, pH, C_t, N_t, log mpd) and on three selected independent variables (microbial respiration, pH, log mpd). F values and significance levels of ANOVAs for dependent samples (6 and 54 degrees of freedom), computed after 10-fold crossvalidations, are shown. Models sharing lowercase letters are not statistically different at $P < 0.05$. The relative decrease of MAE as compared with the model with the weakest predictive power is also shown. NN, neural network; MT, model tree; RT, regression tree; MI, model tree with instances only; MR, multiple linear regression; LIN, simple linear regression; LSM, land-use-specific mean (see text for detailed description of models); $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$

Number of input variables	Total abundance				Species number				Abundance of *F. quadrioculata*			
	Model	F	MAE	decrease in MAE (%)	Model	F	MAE	decrease in MAE (%)	Model	F	MAE	decrease in MAE (%)
10	NN	10.47***	9.02a	−19.19	NN	3.00*	1.33a	−15.52	NN	8.40***	1.64a	−38.24
	MT		9.31a,b	−16.55	MT		1.51b	−4.41	MT		2.24b	−15.48
	RT		9.60a,b	−13.97	RT		1.51b	−4.03	RT		2.26b	−14.84
	MI		9.83b,c	−11.94	MI		1.51b	−3.90	MI		2.28b	−13.93
	MR		10.54c,d	−5.58	MR		1.51b	−3.89	MR		2.35b,c	−11.57
	LSM		10.92d	−2.11	LSM		1.55b	−1.79	LSM		2.54b,c	−4.39
	LIN		11.16d		LIN		1.58b		LIN		2.65c	
3	NN	7.64***	9.40a	−15.76	NN	2.79*	1.40a	−10.92	NN	15.30***	1.78a	−33.00
	MT		9.72a	−12.92	RT		1.50a,b	−4.68	MI		2.17b	−18.05
	RT		9.80a,b	−12.19	LSM		1.51b	−3.89	MT		2.24b,c	−15.51
	MI		10.05a,b	−9.97	MI		1.53b	−2.60	LSM		2.35b,c	−11.57
	MR		10.45b,c	−6.39	MR		1.57b	−0.16	RT		2.35b,c	−11.32
	LSM		10.92c,d	−2.11	MT		1.57b	−0.11	MR		2.46c	−7.45
	LIN		11.16d		LIN		1.58b		LIN		2.65c	

MI M5 model with instances only
RT M5 regression tree
MT M5 model tree
MR multiple linear regression
LIN univariate linear regression
LSM land-use-specific mean

All models — except LSM (only input variable: land use type) and LIN (only input variable: variable with the highest correlation coefficient with the output variable) — were generated with ten input variables: x and y coordinate of the sampling point, land use type, microbial biomass, microbial respiration, soil moisture, pH, C_t, N_t and log mpd.

Assuming that, for the estimation of a function, at least five data points are necessary, a NN with ten input variables would roughly require 10^5 patterns or more for training in order to cover the entire variable space. With a restricted number of training patterns, a parsimonious approach to the design of the NN must be followed and the number of input variables must be kept as small as possible. Since several of the input variables — microbial biomass, microbial respiration, C_t, N_t, soil moisture — demonstrate strong correlations among each other with $r > 0.95$ (Kampichler, 1998a), a second series of models was generated. These models have only three uncorrelated input variables: microbial respiration (which turned out to be the most important variable among the five highly correlated variables listed above), soil acidity and log mpd. The same development and test subsets as above were used. A subscript figure will denote the number of input variables throughout the paper (e.g. NN_{10} is the neural network model with ten input variables, MT_3 is the model-tree model with three input variables).

2.4. Comparison of predictive power

In a 10-fold crossvalidation, all models (NN, MT, RT, MI, MR, LSM, LIN) were generated by use of the ten subsets of development patterns. These subsets were identical for each model type. The models were then applied to the respective subsets of test patterns and the mean absolute error (MAE) between the model outputs (=predictions) and the observed values — a deviance measure recommended by Mayer and Butler (1993) for model validation — was calculated for each test subset. MAEs of each model for each test subset were tested for normality by means of the nullklassen test (a test particularly designed for small samples with $n < 20$: Zöfel, 1992), for homogeneity of variances (Bartlett test) and for additivity of effects (Tukey test). Differences in predictive power of the different model types were tested by an ANOVA for

Table 3
Increase in mean absolute error (MAE) of predictive models for total abundance of hemi- and euedaphic Collembola, for species number of hemi- and euedaphic Collembola and for numerical abundance of *F. quadrioculata* following the reduction of the number of independent variables from 10 to 3. *t* values and significance level of *t*-tests for dependent samples (9 degrees of freedom) are shown

Variable	Model	MAE		*t*	*P*
		ten input variables	three input variables		
Total abundance	RT	9.60	9.80	2.06	0.07
	NN	9.02	9.40	2.08	0.07
	MT	9.31	9.72	1.22	0.25
	MI	9.83	10.05	0.61	0.56
	MR	10.54	10.45	0.32	0.76
Species number	MT	1.51	1.70	1.78	0.11
	NN	1.33	1.40	1.25	0.24
	MI	1.51	1.53	0.65	0.53
	MR	1.55	1.57	0.55	0.60
	RT	1.51	1.50	0.47	0.65
Abundance of *F. quadrioculata*	RT	2.28	2.35	1.76	0.11
	MR	2.54	2.46	1.05	0.32
	NN	1.64	1.78	0.72	0.49
	MI	2.24	2.17	0.59	0.57
	MT	2.26	2.24	0.21	0.84

dependent samples and subsequent Duncan tests. Differences between models with ten and three input variables within the same model type were tested by *t*-tests for dependent samples.

3. Results

3.1. Predictive power of models with ten input variables

Generally, linear regression (LIN_{10}) models had the weakest predictive power for total abundance, abundance of *F. quadrioculata* and species number (Table 2). Thus, the relative gain in predictive power of the other models are expressed as the relative decrease of MAE compared with LIN_{10}. Land-use-specific mean (LSM_{10}) models performed slightly better and MAE decreased by about 2–4%. Multiple regression (MR_{10}) models decreased the modelling error by about 4–12%. The model types capable of representing non-linear relationships showed the best predictive power, with tree-based models (MT_{10}, RT_{10}, MI_{10}) and neural network (NN_{10}) models having MAEs of about 4–17 and 16–38%, respectively, lower than LIN_{10}.

The gain in predictive power was moderate for total abundance (Table 2; ten input variables). NN_{10}, MT_{10} and RT_{10} models performed similarly well and decreased MAEs by about 14–19% compared with the weakest model, LIN_{10}. For species number, only NN_{10} models showed an increase in predictive power (MAE decrease of about 15.5%); all other models were statistically not significantly different from the LIN_{10} models. The best results of NN_{10} and tree-based models were achieved for abundance of *F. quadrioculata*: MT_{10}, RT_{10} and MI_{10} models (MAE decrease of about 14–15.5%) are slightly better than the MR_{10} models (MAE decrease of about 11.5%), NN_{10} models decreased MAE by more than a third (38%).

Abundance of *O. armatus* resisted any modelling attempt. It could not be related to any environmental characteristic and even NN_{10} models — which had turned out to be the most powerful modelling approach for total abundance, species number and abundance of *F. quadrioculata* — yielded predictions that remained uncorrelated with the observations. Thus, any further attempts to model abundance of *O. armatus* were ceased.

3.2. Predictive power of models with three input variables

Generally, the reduction of input variables led to a slight increase in MAE, although some models (e.g. MR_3, MR_3, MI_3) had lower MAEs than the corresponding models with ten input variables. None of these changes, however, were statistically significant at $P < 0.05$, with only RT_3 and NN_3 for total abundance showing a considerable decrease in predictive power ($P = 0.07$) (Table 3).

LIN_3 models were still least successful in predicting total abundance, species number and abundance of *F. quadrioculata* (Table 2). NN_3 models again showed the best predictive power; their MAEs were about 11–33% lower than the MAEs of the LIN_3 models. Tree-based

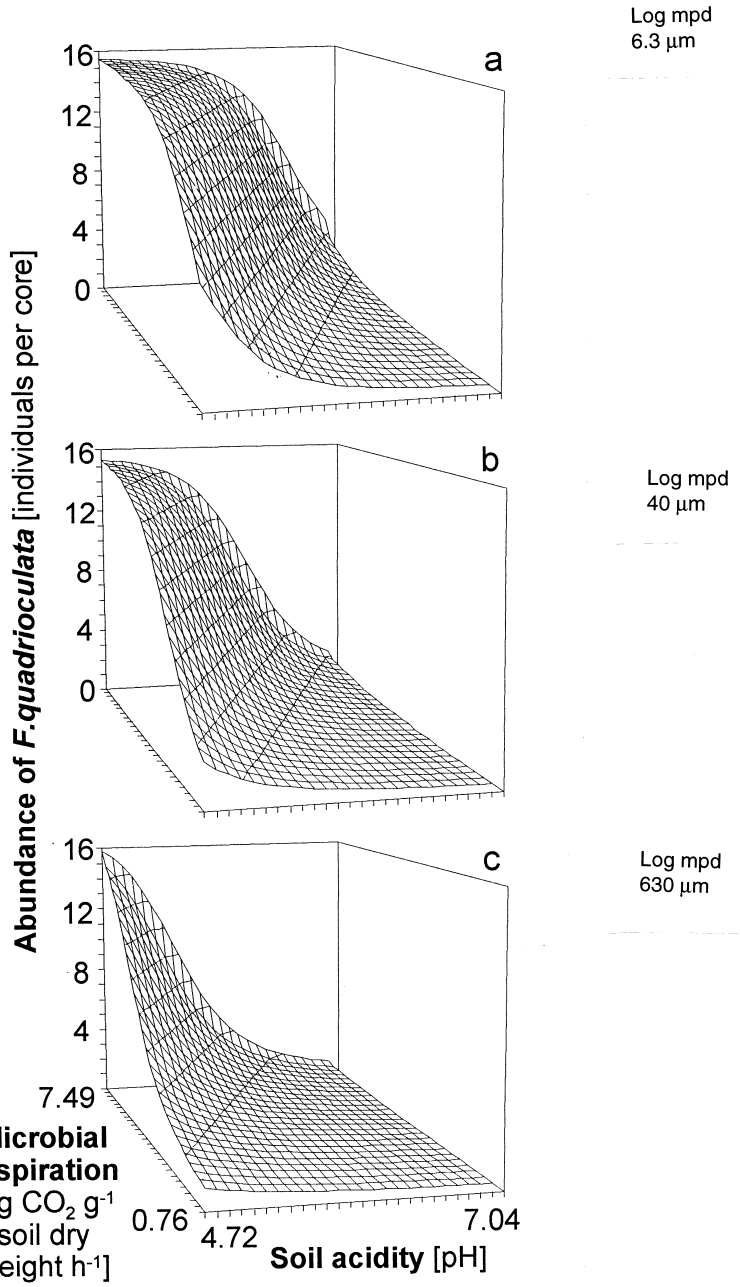

Fig. 2. Response surface of a NN_3 model for abundance of *F. quadrioculata* on the FAM experimental farm at Scheyern with input values for log mpd 6.3 µm (a), 40 µm (b) and 630 µm (c). Endpoints of *x*- and *y*-axis are the minimum and maximum values of microbial respiration and soil acidity in the development pattern set.

models appeared to be next best in predictive power: MT_3 for total abundance had a MAE of 13% lower than LIN_3, MI_3 for abundance of *F. quadrioculata* led to a MAE reduction of 18%.

Again, NN and tree-based models could efficiently represent the abundance of *F. quadrioculata*, but were less successful in modelling total abundance. For species number, only NN_3 led to at least a moderate increase in predictive power.

3.3. Relationships between environmental characteristics and Collembola

NN_{10} models gave the best predictions, but, for a complete sensitivity analysis of the network, approximately 10,000,000 patterns would have been necessary if each input variable had been divided into only five input levels (cf. Introduction). Sensitivity analysis of NN_3 models, in contrast, was feasible and yielded re-

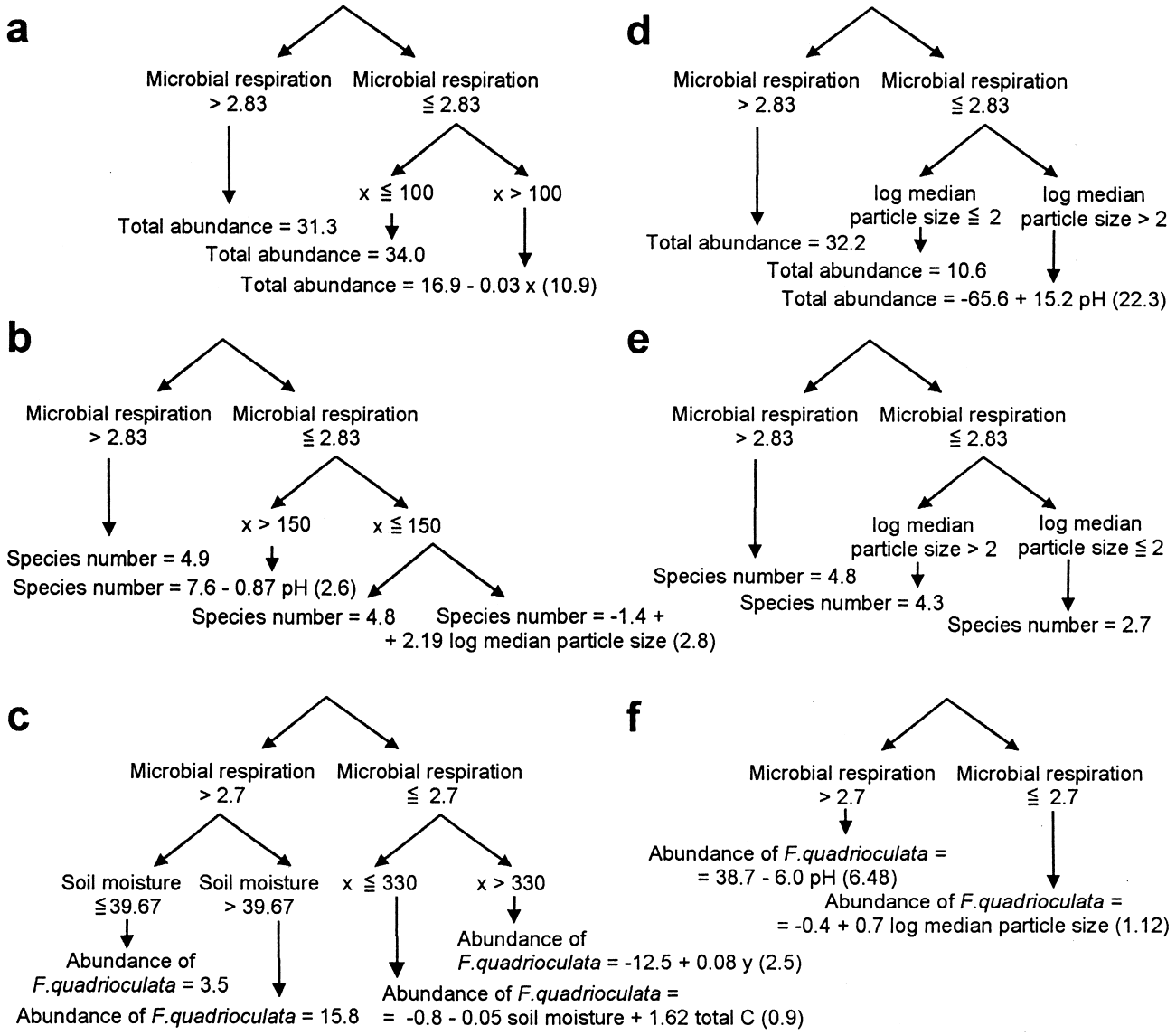

Fig. 3. Selected MT_{10} trees for (a) total abundance of hemi- and euedaphic Collembola, (b) species number of hemi- and euedaphic Collembola and (c) abundance of *F. quadrioculata*, as well as selected MT_3 trees for (d) total abundance of hemi- and euedaphic Collembola, (e) species number of hemi- and euedaphic Collembola and (f) abundance of *F. quadrioculata* on the FAM experimental farm at Scheyern. Mean values of the output variables for regression models in leaves of the model tree are given in parentheses.

sponse surfaces that could cover the entire variable space. Fig. 2 illustrates the dependence of *F. quadrioculata* on microbial respiration and soil acidity at low, medium and high values of log mpd.

The tree-based models gave an explicit and illustrative picture of the relationship between input and output variables. For total abundance, the MT_{10} models yielded trees of the form in Fig. 3a: in all of the ten crossvalidation sets, microbial respiration was the input variable determining the main bifurcation of the tree and this was the reason, why microbial respiration was included in the models with only three input variables (cf. Material and Methods). In eight of the ten crossvalidation sets, the position of the soil cores along the east–west axis of the farm was responsible for the second ramification. Microbial respiration was also responsible for the main bifurcations in the MT_{10} model trees for species number and abundance of *F. quadrioculata* (Fig. 3b,c). For species number, the position along the east–west axis of the farm was again important for further ramifications of the model trees (six of ten crossvalidation sets).

For total abundance, all of the ten MT_3 models yielded trees with two bifurcations; the first determined by microbial respiration and the second by log mpd (Fig. 3d). For species number, the model trees tended to be simpler with only one bifurcation, again determined by microbial respiration. In only two of the ten

crossvalidation sets was log mpd responsible for another ramification (Fig. 3e). One half of the ten MT_3 trees for abundance of *F. quadrioculata* showed only one bifurcation, one branch with a linear regression including soil acidity and the other branch with a linear regression including log mpd (Fig. 3f). The remaining trees had a second bifurcation determined by soil acidity or log mpd.

In general, missing data of input variables cause difficulties in training NNs, since the missing data must be substituted by arbitrary values, e.g. by the mean value of the respective variable. In contrast, the M5 algorithm can easily handle datasets with missing data. In contrast to NNs, M5 also has no upper limit for the number of input variables for a given sample size. Since MTs appeared to be the best models in predictive power next to NNs, we additionally created MTs using all 396 cores of the sample (see Introduction) and using all available habitat features as input variables (see the FAM database at the URL http://www.gsf.de/FAM/adis.html for details). We again performed a 10-fold crossvalidation. MTs with all available habitat features as input variables will be abbreviated MT_{all}.

The main bifurcations of the ten MT_{all} models were caused by field type (five trees), microbial respiration (two trees), N_t (two trees) and percent cover of herbs (one tree). This diversity of trees demonstrates a strong dependence of the individual models on the random assignment of patterns to the crossvalidation data subsets. C_t, N_t and microbial respiration were closely correlated with field type; thus, all trees showed an overall effect of organic matter content on total abundance. When total abundance was not thresholded at 45 individuals per core, MT_{all} models become more regular with all trees showing an identical main bifurcation at $C_t = 2.55\%$. Further ramifications, however, showed no regularities.

The MT_{all} models for species number also showed irregularities: the main bifurcations were caused by the cover of herbs in arable fields (five trees), N_t (four trees) and soil moisture (one tree). The trees were intensely ramified and showed no similarities among each other.

The MT_{all} models for abundance of *F. quadrioculata* were dominated by N_t (nine trees) with a ramification point at 0.20%. In six of ten trees the $N_t > 0.20$ branch was further ramified at soil moisture = 37.5%. Abundance of *F. quadrioculata* averaged at 13–15 ($N_t > 0.20$ and soil moisture > 37.5% branch), at 2–9 ($N_t > 0.20$ and soil moisture $\leq$ 37.5% branch) and at 1.1–1.2 ($N_t \leq 0.20$ branch) individuals per core.

Average MAEs of the MT_{all} models were slightly higher (10.54 for total abundance, 1.60 for species number, 2.33 for abundance of *F. quadrioculata*) than the average MAEs of the MT_{10} models, thus yielding no increase in predictive power.

4. Discussion

4.1. Predictive power of models

The higher predictive power of NN models and tree-based models clearly reflect the nonlinearities in the relationship between environmental variables and characteristics of the Collembolan community and confirm earlier findings of Kampichler (1998a, b). Although less successful in predicting abundance and species numbers, MT and RT models have the appealing quality of being transparent and providing explicit information about the quantitative relationships between the variables. Thus, if prediction is the task, NNs are the proper tools for achieving that goal; if, however, the understanding of abundance and diversity patterns is also desired, tree-based models appear to serve better.

The predictive power of models with three input variables was slightly weaker than that of models with ten input variables. Measured against the computational complexity and requirements, however, predictive power went up considerably with less variables for NNs and tree-based models.

A striking finding is that variables that predicted abundance for one species (*F. quadrioculata*) fairly well were useless for another (*O. armatus*). If species are indeed that different in their requirements for habitat characteristics, it is unlikely that predictive models for Collembolan communities can be developed.

4.2. Relationships between environmental characteristics and Collembola

Abundance and distribution of Collembola depend on a multivariate array of environmental characteristics, such as soil structure, soil pH, soil moisture, crop rotation, tillage and soil microbial characteristics (e.g. Jagers op Akkerhuis et al., 1988; Dekkers et al., 1994; Klironomos and Kendrick, 1995). A soil property repeatedly addressed as being important for determining Collembolan numerical abundance or biomass is organic matter content (Alejnikova, 1965; Ghilarov, 1975; Andrén and Lagerlöf, 1983; Kovác and Miklisová, 1997). Vreeken-Buijs et al. (1998), in contrast, could not detect any relationship between Collembola biomass and soil organic matter when comparing Dutch forest, agricultural and grassland soils. They concluded that large differences in land use and soil type caused the absence of any significant correlation with organic matter (and other soil characteristics, such as pH or N mineralization). At the FAM

experimental farm at Scheyern, soils vary considerably at small distances (cf. Table 1). Due to the hilly landscape and resulting translocation processes, there are colluvial silty and clayey soils at the foot of the slopes and eroded sandy and gravelly soils at the hilltops (Hantschel et al., 1993). Despite this heterogeneity of soil types and the variety of land-use systems, soil organic matter turned out to have a considerable potential for predicting Collembolan total abundance at the farm. Indeed, microbial respiration, C_t and N_t were the variables with the highest predictive potential using the MT_3, MT_{10} and MT_{all} models. However, among the highly correlated soil characteristics microbial biomass, microbial respiration, soil moisture, C_t and N_t (cf. Material and Methods), the latter two were the variables that could be most highly considered 'independent' (even disregarding feedback processes in soil). Carbon content determines the size of the microbial biomass attainable at a site: the proportion of soil C within the microbial biomass is remarkably constant and normally ranges from 2 to 3% (Anderson and Domsch 1989). Soil organic matter also has a high water capacity and, thus, can profoundly determine soil moisture characteristics (Scheffer and Schachtschabel, 1989).

Kovác and Miklisová (1997) reported that none of the abundances of the dominant species in agricultural fields in east Slovakia was related to any edaphic factor, although total abundance of Collembola was correlated with the content of organic C, N and P. Similarly, *O. armatus* also did not demonstrate a relationship to any of the soil characteristics at the FAM experimental farm. The abundance of *F. quadrioculata*, in contrast, could be well explained by habitat features. NN modelling was particularly successful in this case and the MT_{all} models highlighted the dependence of *F. quadrioculata* abundance on N_t and — when N_t is low — on soil moisture. This led to the assumption that an apparent lack of statistical relationships between a single species and soil characteristics is potentially due to the nonlinearity of these relationhips, which can hardly be detected by standard linear methods (correlation analysis, PCA), if indeed there is any real relationship. In the case of *O. armatus* — a species known to be euryoecious and ubiquitous — any identifiable relationship to certain habitat characteristics in fact seemed to be absent.

Species number of Collembola was the least predictable quantity. This is in accord with the observation by Kovác and Miklisová (1997) that average species richness in arable fields at twelve study sites in east Slovakia was not correlated with edaphic factors. At the FAM experimental farm, this was most likely due to the low accuracy of estimating species number by taking single cores at the sampling points. The presence of less abundant species in a core is a matter of chance and, thus, the relationship between soil characteristics and species number is necessarily very weak. Insofar, it is remarkable that NN models could decrease the mean absolute error of the prediction by about 16% (NN_{10}) and 11% (NN_3) (Table 2). This highlights the fact that there actually is a certain relation between soil characteristics and species number, which, as in the cases of total abundance and abundance of *F. quadrioculata*, can be better resolved by nonlinear methods. MT_{10} and MT_3 models suggest that, again, C_t and N_t as well as the microbial variables correlated with it exert a major influence, whereas MT_{all} also points at the importance of the herb cover in arable fields (probably influencing soil moisture).

4.3. Limits of the models

In an earlier modelling attempt on the dataset from the Scheyern experimental farm, Kampichler (1998a) compared the modelling success of neural networks and multiple regressions and concluded that the NN models are well able to characterize the potential of a site to provide a habitat for *F. quadrioculata*. Since the actual density is, however, dependent on a variety of processes acting at different spatial and temporal scales (for example, microscale aggregation behavior, interspecific interactions, immigration from adjacent habitats, local abundance of predators, or stochastic disturbances), it is not surprising that pure habitat models fail to predict local abundance beyond a certain point of precision (e.g. for total abundance, species number and abundance of *F. quadroculata*) or totally (e.g. for abundance of *O. armatus*). NN_{10} and MT_{10} models clearly illustrate the fact that processes acting on a regional scale influence community characteristics of Collembola: the position along the *x*-axis of the sampling grid, representing the west–east extension of the experimental farm for about 1.5 km, was, besides microbial respiration, the most important factor influencing total abundance, species number and abundance of *F. quadrioculata* (Fig. 3a–c). Raimondi (1990) suggests that differences in local habitat quality can only be expected to be reflected in animal density–distribution patterns when abundance is very high and approaches the carrying capacity of the habitat.

All models we compared in this study also suffer from a common fundamental problem: they assume causal unidirectionality from 'independent' to 'dependent' variables, disregarding feedback loops in soil processes. Microbial biomass not only provides food for a certain number of grazing organisms, but may itself respond to grazing by compensatory growth, thus again altering food supply (Lussenhop, 1992). Advanced methods, such as backpropagation NNs and tree-based models, may be better suited to model non-

linear relationships than conventional statistics, but even they cannot take into account feedback loops and dynamic relationships between variables.

Besides this general problem of pure habitat–quality models, their predictive power also heavily depends on the characteristics and quality of the data used for model development and for model validation, since simulation models cannot be expected to provide results that are more accurate and precise than the available data (Rykiel, 1997). First, the data used here originated from only one point in time (spring 1991). Possibly, the factors that set a limit for local abundance are effective at another time in the year (e.g. the amount of soil moisture during summer), leading to only weak relationships between habitat variables and density of Collembola at the time of sampling. Species number is probably less affected since, in spring, 90% of all species known from the Scheyern experimental farm could be recorded (J. Filser, pers. comm.). Second, at each sampling point, only one single core was taken, potentially leading to erratic relationships between soil characteristics and abundance and species number of Collembola. Moreover, cores for species extraction and for measuring environmental variables were taken separately. Third, only the upper 5 cm of soil were sampled, thus missing an unknown number of individuals in other horizons and possibly giving rise to errors in abundance and species number estimates. Filser and Fromm (1995) found that several euedaphic species known to be present at the Scheyern experimental farm were not detected by the grid sampling in April 1991. A sampling depth of 5 cm may yield reasonable estimates for Collembolan abundance in grasslands at the farm, but not for the euedaphic species in arable fields (Filser and Fromm, 1995). This may also be an explanation for the modelling failure for abundance of *O. armatus*, which is known to be an euedaphic species. Fourth, microflora in the grid sample was characterized only by total biomass and respiration; no further distinction between certain groups of microflora had been undertaken (Winter, 1998). Collembola, however, may react differently to different microbial variables. Klironomos and Kendrick (1995), for instance, found total length of fungal hyphae and diversity of darkly pigmented fungi to be important variables influencing microarthropod community structure in a maple-forest soil. In organically managed agricultural fields in Denmark, the abundance of Collembola showed significant correlations with yeasts and cellulose-degrading fungi (J.A. Axelsen, pers. comm.).

All these drawbacks of data quality may have potentially veiled the actual relationship between Collembola and environmental variables. We assume that this dataset, which is distinctly larger than a typical soil biological sample, illustrates the limits of habitat models for Collembola. Additional expenditure of time and manpower — if at all feasible — could possibly overcome the data quality drawbacks listed above, but cannot solve the fundamental problems of: (1) various ecological processes superimposing the pattern of habitat quality and of (2) feedback loops prevailing in functional relationships, both of which hamper the predictive power of habitat features for the local abundance and species number of Collembola. We have shown, however, that novel modelling approaches that are able to reflect the nonlinearities of habitat–Collembola relationships may distinctively improve the predictive potential of habitat characteristics.

Acknowledgements

The scientific activities of the research network Forschungsverbund Agrarökosysteme München (FAM) are financially supported by the Federal Ministry of Culture and Science, Research and Technology. Rent and operating expenses are paid by the Bavarian State Ministry for Education and Culture, Science and Art. J.A. Axelsen, J. Filser, C.G. Jones and G. Weigmann commented on the manuscript, D. Russell corrected the English. We are grateful for their help.

Appendix A

The following description of the functioning of backpropagation networks follows the introduction by Gallant (1993).

The input into a node u_i (except for the input nodes) is a weighted sum S of the outputs from all nodes u_j connected to it:

$$S_i = \sum_j w_{i,j} u_j \tag{A.1}$$

where $w_{i,j}$ is the weight assigned to a connection from node u_j to node u_i. The node u_i computes its output — its *activation* — as a nonlinear and differentiable function $f(S_i)$ of the weighted sum of the inputs to that node. Most commonly — and also in this paper — the logistic activation function is used:

$$u_i = f(S_i) = \frac{1}{1 + e^{-S_i}} \tag{A.2}$$

The backpropagation learning algorithm is a gradient-descent algorithm.

The backpropagation update rule is

$$\Delta w_{i,j} = \rho \ \delta_i u_j \tag{A.3}$$

with

$$\delta_i = (C_i - u_i)f'(S_i), \tag{A.4a}$$

if u_i is an output node,

$$\delta i = (\sum w_{m,i}\delta m)f\,'(S_i), \tag{A.4b}$$

if u_i is a hidden node, and

$$f\,'(S_i) = u_i(1 - u_i) \tag{A.4c}$$

where $\Delta w_{i,j}$ is the change of the connection weight between u_j and u_i in a backpropagation step, ρ a constant learning parameter (termed η in SNNS), C_i the teaching output, u_i the actual output, $f'(S_i)$ the derivative of the logistic activation function $f(S_i)$ and $w_{m,i}$ the weight assigned to the connection from u_i to u_m.

In the beginning of the learning process, a small positive value for ρ is chosen and small initial weights are randomly assigned to all connections. Then, the network is provided with a *training pattern*, corresponding to the term *development pattern* used in this paper. In a *forward propagation step*, the input is passed through the network: the weighted sums of inputs, S_i and the activations, $u_i = f(S_i)$, are calculated for all nodes according to Eqs. (A.1) and (A.2). Subsequently, a backward pass through the network is performed, starting with the output node, computing Eqs. (A.3) and (A.4) and, finally, leading to the update of the connection weights:

$$w_{i,j}(t+1) = w_{i,j}(t) + \Delta w_{i,j} \tag{A.5}$$

with $w_{i,j}(t)$ denoting connection weights at learning step t and $w_{i,j}(t+1)$ denoting connection weights at learning step $t + 1$. This procedure is iteratively performed, until a stop criteria is reached, e.g. at minimum error for a set of test patterns.

Appendix B

The following descriptions of tree-based models and of the M5 algorithm follow the papers by Breiman et al. (1984) and Quinlan (1992, 1993).

The problem of regression analysis is a problem of searching for the dependencies between a dependent variable y and independent variables x_i. Tree-structured regression is based on the assumption that the functional dependence is not uniform throughout the entire domain, but can be approximated as such on smaller subdomains. The tools developed for model tree induction search for these subdomains automatically and characterize them with regression functions or constants of the dependent variable. A tree consists of branches, internal nodes (the branching points) and leaves (the terminal nodes). Given an example for which the value of the dependent variable should be estimated, knowing the values of the independent variables, the tree is interpreted from the root, i.e. its uppermost node. In each internal node, a test is performed and, according to its result, the corresponding left or right subtree is selected until a leaf is reached. At that point, a value is computed according to the model in that leaf. This value represents an answer by the tree and is assigned to the example as the value of the dependent variable.

The M5 algorithm belongs to the TDIDT (top-down induction of decision trees) family of algorithms. This means, it splits the dataset at each node of a tree into subsets, from which it recursively forms subtrees. The M5 algorithm maximizes the expected error reduction when splitting the data into several sets depending on various potential tests. Let L denote a learning set of examples. At each step of tree construction, M5 tests whether the set contains just a few examples or examples with class values that vary only slightly. In this case, a leaf is constructed and tree construction is terminated. Otherwise, the learning set L is split according to the outcome of a test. A list of potential tests is evaluated by determining the subset of cases, L_i, associated with each outcome i. Standard deviation $\mathrm{sd}(L_i)$ of the target values of the examples in L_i is selected as the measure of error. The expected reduction of error as a result of this test can be written as

$$\Delta\mathrm{error} = \mathrm{sd}(L) - \sum_i \frac{|L_i|}{|L|}\mathrm{sd}(L_i) \tag{B.1}$$

After examining all possible tests, M5 chooses one that maximizes this expected error reduction.

The M5 program can build trees in three different ways: (1) using only regression (regression trees), (2) using only instance-based learning (models with only instances) and (3) using regression and instance-based learning (model trees). Regression tries to fit a linear model through the remaining examples from the learning set L. Instance-based learning tries to classify an example according to the neighbors in close surroundings. These close surroundings should be understood in a multidimensional space, i.e. one dimension for each of the attributes. Model trees combine both approaches.

References

Alejnikova, M.M., 1965. Die Bodenfauna des Mittleren Wolgalandes und ihre regionalen Besonderheiten. Pedobiologia 5, 17–49.

Anderson, T.-H., Domsch, K.H., 1989. Ratios of microbial biomass carbon to total organic carbon in arable soils. Soil Biology & Biochemistry 21, 471–479.

Andrén, O., Lagerlöf, J., 1983. Soil fauna (microarthropods, enchytraeids, nematodes) in Swedish agricultural cropping systems. Acta Agricultura Scandinavica 33, 32–52.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, CA, Belmont.

Butcher, J.W., Snider, R., Snider, R.J., 1971. Bioecology of edaphic Collembola and Acarina. Annual Review of Entomology 16, 249–288.

Debauche, H.R., 1962. The structural analysis of animal communities of the soil. In: Murphy, P.W. (Ed.), Progress in Soil Zoology. Butterworths, London, pp. 10–25.

Dekkers, T.B.M., van der Werff, P.A., van Amelsvoort, P.A.M., 1994. Soil Collembola and Acari related to farming systems and crop rotations in organic farming. Acta Zoologica Fennica 195, 28–31.

Dimoupoulos, Y., Bourret, P., Lek, S., 1995. Use of some sensitivity criteria for choosing networks with good generalization ability. Neural Processing Letters 2, 1–4.

Filser, J., Fromm, H., 1995. The vertical distribution of Collembola in an agricultural landscape. Polskie Pismo Entomologiczne 64, 99–112.

Fromm, H., 1997. Räumliche und zeitliche Variabilität der Collembolenfauna und ihre Bedeutung für C- und N-Umsatz in einer Agrarlandschaft. FAM-Bericht 26. Shaker, Aachen.

Gallant, S.I., 1993. Neural Network Learning and Expert Systems. MIT Press, Cambridge, MA.

Ghilarov, M.S., 1975. General trends of changes in soil animal population of arable land. In: Vanek, J. (Ed.), Progress in Soil Zoology. Academia, Prague, pp. 31–39.

Hantschel, R.E., Lenz, R.J.M., Kainz, M., Beese, F., 1993. Ziele, Hypothesen und Arbeitsschritte des Forschungverbundes Agrarökosysteme München. In: Hantschel, R., Kainz, M. (Eds.), Abschlussbericht Aufbauphase (1990–1992). GSF, Oberschleissheim, pp. 1–11.

Hartge, K.H., Horn, R., 1989. Die physikalische Untersuchung von Böden, 2nd ed. Enke, Stuttgart.

Heinemeyer, O., Insam, H., Kaiser, E.-A., Walenzik, G., 1989. Soil microbial biomass and respiration measurements: an automated technique based on infrared gas analysis. Plant and Soil 116, 191–195.

Heuer, M., Leschonski, K., 1985. Erfahrungen mit einem neuen Gerät zur Messung von Partikelgrössenverteilungen aus Beugungsspektren. In: Leschonski, K. (Ed.), Proceedings of the 3rd European Symposium on Partikelmesstechnik. Nürnberger Messe- und Ausstellungsgesellschaft, Nürnberg, pp. 515–538.

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

Jagers op Akkerhuis, G.A.J.M., de Ley, F., Zwetsloot, H.J.C., Ponge, J.-F., Brussaard, L., 1988. Soil microarthropods (Acari and Collembola) in two crop rotations on a heavy marine clay soil. Revue d'Écologie et de Biologie du Sol 25, 175–202.

Jongman, R.H.G., Braak, C.J.F.t., Tongeren, O.F.R.v., 1995. Data Analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge.

Kampichler, C., 1998a. The potential of soil habitat features for the modelling of numerical abundance of Collembola: a neural network approach. Verhandlungen der Gesellschaft für Ökologie 28, 151–159.

Kampichler, C., 1998b. Modelling abundance of edaphic Collembola in an agricultural landscape by means of artificial neural networks. In: Pizl, V., Tajovsky, K. (Eds.), Soil Zoological Problems in Central Europe. Proceedings of the 4th Central European Workshop on Soil Zoology. ISB ASCR, Ceske Budejovice, pp. 79–86.

Klironomos, J.N., Kendrick, B., 1995. Relationships among microarthropods, fungi and their environment. Plant and Soil 170, 183–197.

Kompare, B., Džeroski, S., 1995. Getting more out of data: automated modelling of algal growth with machine learning. In: Matsuzaki, C., Miyaji, Y. (Eds.), Proceedings of the International Conference on Coastal Ocean Space Utilization. University of Hawaii, Honolulu, pp. 209–220.

Kovác, L., Miklisová, D., 1997. Collembolan communities (Hexapoda, Collembola) in arable soils of east Slovakia. Pedobiologia 41, 62–68.

Köhn, M., 1928. Bemerkungen zur mechanischen Bodenanalyse, III. Ein neuer Pipettapparat. Zeitschrift für Pflanzenernährung, Düngung und Bodenkunde 11, 50–54.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling 90, 39–52.

Lussenhop, J., 1992. Mechanisms of microarthropod–microbial interactions in soil. Advances in Ecological Research 23, 1–33.

Mayer, D.G., Butler, D.G., 1993. Statistical validation. Ecological Modelling 68, 21–32.

Ponge, J.-F., 1993. Biocoenoses of Collembola in atlantic temperate grass–woodland ecosystems. Pedobiologia 37, 223–244.

Ponge, J.-F., Arpin, P., Vannier, G., 1993. Collembolan response to experimental perturbations of litter supply in a temperate forest ecosystem. European Journal of Soil Biology 29, 141–153.

Quinlan, J.R., 1992. Learning with continuous classes. In: Adams, A., Sterling, L. (Eds.), Proceedings AI'92. World Scientific, Singapore, pp. 343–348.

Quinlan, J.R., 1993. Combining instance-based and model-based learning. In: Proceedings of the tenth International Conference on Machine Learning. Morgan Kaufman, San Mateo, pp. 236–243.

Raimondi, P.T., 1990. Patterns, mechanisms, consequences of variability in settlement and recruitment of an intertidal barnacle. Ecological Monographs 60, 283–309.

Recknagel, F., French, M., Harkonen, P., Yabunaka, K.-I., 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecological Modelling 96, 11–28.

Rykiel Jr., E.J., 1997. Testing ecological models: the meaning of validation. Ecological Modelling 90, 229–244.

Schaap, M.G., Leij, F.J., 1998. Using neural networks to predict soil water retention and soil hydraulic conductivity. Soil and Tillage Research 47, 37–42.

Scheffer, F., Schachtschabel, P., 1989. Lehrbuch der Bodenkunde, 12th ed. Enke, Stuttgart.

Sinowski, W., 1994. Die dreidimensionale Variabilität von Bodeneigenschaften: Ausmass, Ursachen und Interpolation. FAM-Bericht 7. Shaker, Aachen.

Verhoef, H.A., Brussaard, L., 1990. Decomposition and nitrogen mineralization in natural and agro-ecosystems: the contribution of soil animals. Biogeochemistry 11, 175–211.

Vreeken-Buijs, M.J., Hassink, J., Brussaard, L., 1998. Relationships of soil microarthropod biomass with organic matter and pore size distribution in soils under different land use. Soil Biology & Biochemistry 30, 97–106.

Warner, B., Misra, M., 1996. Understanding neural networks as statistical tools. American Statistician 50, 284–293.

Winter, K., 1998. Räumliche und zeitliche Variabilität der mikrobiellen Biomasse und ihrer Aktivität in einer heterogenen Agrarlandschaft. FAM-Bericht 20. Shaker, Aachen.

Zell, A., Mamier, G., Vogt, M., Mache, N., Hübner, R., Döring, S., Herrmann, K.-U., Soyez, T., Schmalzl, M., Sommer, T., Hatzigeorgiou, A., Posselt, D., Schreiner, T., Kett, B., Clemente, G., Wieland, G., 1995. SNNS Stuttgart Neural Network Simulator, User Manual. Report 6/95. University of Stuttgart, Institute for Parallel and Distributed High Performance Systems, Stuttgart.

Zöfel, P., 1992. Statistik in der Praxis, 3rd ed. Gustav Fischer, Stuttgart.