Available online at www.sciencedirect.com
SCIENCE DIRECT®
Ecological Modelling 191 (2006) 131–143
ECOLOGICAL MODELLING
www.elsevier.com/locate/ecolmodel

Using multi-objective classification to model communities of soil microarthropods

Damjan Demšar[a,*], Sašo Džeroski[a], Thomas Larsen[b], Jan Struyf[c], Jørgen Axelsen[b], Marianne Bruus Pedersen[b], Paul Henning Krogh[b]

[a] *Department of Knowledge Technologies, Jožef Stefan Institute, Jamova Ljubljana, Slovenia*
[b] *Department of Terrestrial Ecology, National Environmental Research Institute, Roskilde, Denmark*
[c] *Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium*

Available online 7 October 2005

Abstract

In agricultural soil, a suite of anthropogenic events shape the ecosystem processes and populations. However, the impact from anthropogenic sources on the soil environment is almost exclusively assessed for chemicals, although other factors like crop and tillage practices have an important impact as well. Thus, the farming system as a whole should be evaluated and ranked according to its environmental benefits and impacts. Our starting point is a data set describing agricultural events and soil biological parameters. Using machine learning methods for inducing regression and model trees, we produce empirical models able to predict the soil quality from agricultural measures in terms of quantities describing the soil microarthropod community. We are also interested in discovering additional higher level knowledge. In particular, we have identified the most important factors influencing the population densities of springtails and mites and their biodiversity. We also identify to which agricultural actions different microarthropods react distinctly. To obtain this higher level knowledge, we employ multi-objective regression trees.

Keywords: Multi-objective classification; Modelling; Soil microarthropods

1. Introduction

The impact of anthropogenic sources on the soil environment is almost exclusively assessed for chemical factors only, although in agriculture mechanical factors like tillage and biological factors such as crops have a large impact as well (Steen, 1983). Since farming systems consist of a certain temporal sequence of interdependent events of different types and durations it is necessary to handle the farming system as a whole in order to accurately rank its environmental benefits and impacts. Based on data about the agricultural events and the soil biological parameters reflecting

* Corresponding author.
E-mail addresses: damjan.demsar@ijs.si (D. Demšar), saso.dzeroski@ijs.si (S. Džeroski), thl@dmu.dk (T. Larsen), jan.struyf@cs.kuleuven.be (J. Struyf), phk@dmu.dk (P.H. Krogh).

0304-3800/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2005.08.017

these events, we build empirical models that relate the sequence of agricultural events to the biological parameters. More specifically, we employ machine learning algorithms that build regression and model trees to induce models able to predict the soil quality in terms of quantities describing the microarthropod community, given historical data about sequences of crops, tillage, fertilisation and other agricultural measures.

Besides building accurate models, we are also interested in discovering higher level knowledge. In particular, we identify the most important factors influencing the biodiversity and the population densities of different microarthropods. Such knowledge can guide us in further experiments and in more focused data collection.

The long-term goal of this work is to design a decision support system for managing farms, which can take into account both the economical and ecological consequences of agricultural actions. The knowledge discovered in this study can be later incorporated into the ecological part of this system. For this reason the purpose of the present modelling exercise is not only to discover new knowledge, but also to "rediscover" the knowledge in a quantitative form and therefore operational form. Obtaining knowledge from domain experts can be hard because it is generally difficult to put down in writing or may be simply too obvious to mention from the expert's point of view. Therefore, we prefer machine learning tools that produce descriptive models from datasets, which can be used as a source of questions to domain experts. Such questions would otherwise be impossible to pose without a significant amount of background knowledge. The answers from the experts can then be used in combination with the discovered knowledge to construct the decision support system.

To obtain interpretable models, we employ machine learning tools capable of multi-objective regression. Such tools allow us to produce one model predicting several biological variables at once. This one model is not only simpler compared to a set of models, one for each individual variable, but can also help us to understand different effects of the same agricultural actions on different aspects of the soil microarthropod community. The tools that we use moreover allow us to constrain the size of the models. In this way, we can easily trade off interpretability for predictive accuracy.

This paper is organized as follows. Section 2 describes the data: the data sources, the observed variables and the transformations that are used. In Section 3, we discuss the modelling techniques that we apply namely single and multi-objective regression trees and model trees. Section 4 describes the modelling phase: the experimental setup, the obtained models and the knowledge that can be derived from these models. In Section 5, we state the main conclusions.

2. Data

The data that is used in this study originates from two data sets. The first data set (Krogh, 1994) describes four experimental farming systems (all located at the Foulum experimental station, Denmark) over the period 1989–1993. Two systems are conventional systems with pesticide use; the other two are organic ones with no pesticide use. Five hundred and thirty microarthropod samples are available in this data set (Demšar et al., 2003). The second data set describes a number of organic farms (Foulum and Flakkebjerg experimental stations and a number of farms in Jutland) over the period 2002–2003. One thousand four hundred and fifteen samples are collected.

The combined data set has 1945 records in total (while our approach works also with significantly less records, larger data sets usually improve accuracy and reduce overfitting). Each record describes one microarthropod sample. A record consists of 145 attributes, of which 142 are input attributes and 3 are target attributes (the abundances of Acari and Collembolans as well as Shannon–Wiener biodiversity). Table 1 lists each attribute together with a short description.

The input attributes describe the field where the microarthropod sample was taken and mainly include agricultural measures (e.g., crops planted, packing, tillage, fertilizer and pesticide use, etc.). For several measures a history of 3 years is recorded, i.e., there is an attribute for the year in which the sample was collected, one for the past year, and one for 2 and 3 years ago.

The type of crop planted on the field is represented with a set of binary attributes, one for each possible crop. The attribute corresponding to the actual crop

Table 1
The available attributes: 142 variables as identified when characterising the fields and 3 target variables

Attribute	Explanation
actsit_mo	Age of the current situation (time in months since current crop was sown or last crop was harvested)
soil_JB	Soil classification number
samp_time	Sample time (1 = March–April, 2 = May–June, 3 = July–August, 4 = September–November)
Current_crop = X	A set of binary attributes describing the current crop (i.e., the data contains one binary attribute for crops X listed in Table 2—not all values from Table 2 appear)
crop1=X	A set of binary attributes describing last year's crop (possible values are listed in Table 2—not all values from Table 2 appear)
crop2 = X	A set of binary attributes describing crop 2 years ago (possible values are listed in Table 2—not all values from Table 2 appear)
crop3 = X	A set of binary attributes describing crop 3 years ago (possible values are listed in Table 2—not all values from Table 2 appear)
AC, AC_Y	A set of binary attributes indicating a crop of type 'annual crop'. The attributes describe current year (AC) and the previous 3 years (AC_1 to AC_3)
MC, MC_Y	A set of binary attributes indicating a crop of type 'multi crop' (with undersown crop). The attributes describe current year (MC) and the previous 3 years (MC_1 to MC_3)
CAC, CAC_Y	A set of binary attributes indicating a crop of type 'catch crop'. The attributes describe current year (CAC) and the previous 3 years (CAC_1 to CAC_3)
WIC, WIC_Y	A set of binary attributes indicating a crop of type 'winter crop'. The attributes describe current year (WIC) and the previous 3 years (WIC_1 to WIC_3)
PC, PC_Y	A set of binary attributes indicating a crop of type 'permanent crop'. The attributes describe current year (PC) and the previous 3 years (PC_1 to PC_3)
ca, ca_Y	A set of attributes describing that cattle are/were grazing on the field (ca_1, ca_2 and ca_3 describe the grazing in previous years)
sh	Sheep are grazing on the field
grazing	Animals are grazing on the field
si, si_Y	A set of attributes indicating that the current/past crop is/was intended for silage or hay (si_1, si_2, si_3 describe the previous years)
sf	Stubble field (current field condition)
o	Seed bed: bare field, seeds planted less than 1 month ago (current field condition)
seha	Seed bed harrowed (current field condition)
sepl	Seedbed ploughed current field condition)
soha	Bare field harrowed (current field condition)
sopl	Bare field ploughed (current field condition)
tr_packing	Months since packing transformed using (1) with $i = 1$ to obtain a positive correlation with the impact
tr_shal_till	Months since shallow (0–5 cm layer) tillage (weed harrowing etc.) transformed using (1) with $i = 4$
tr_subshal_till	Months since subshallow (5–10 cm layer) tillage transformed using (1) with $i = 2$
tr_deep_till	Months since deep (>10 cm layer) tillage (ploughing, rotovation, etc.) transformed using (1) with $i = 2$
fert_lev	Fertilizer level (low = 0, normal = 1, high = 2)
fert_type	Fertilizer type (no = 0, solid = 1, liquid = 2)
Tr_fert_time	Months since fertilization transformed using (1) with $i = 1$
sotr_Y	Soil treatment (tillage and similar) in past year ($Y = 1$), 2 years ago ($Y = 2$) or 3 years ago ($Y = 3$): 0 = none, 1 = in spring or autumn, 2 = in spring and autumn
Target variables	
Acari	Abundance of acari species
coll	Abundance of Collembolan species
H	Shannon biodiversity

Table 2
Possible crops

Abbreviation	Crop	Abbreviation	Crop	Abbreviation	Crop
ba	Winter barley	fa-gr	Fallow, grass	ry-clgr	Rye, clover, grass
ba-ch	Winter barly, chicory	gr	Grass	sba	Spring barley
ba-clgr	Winter barly, clover, grass	le	Leeks	sba-clgr	Spring barely, clover, grass
ba-gr	Winter barley, grass	lu	Lupin	sba-gr	Spring barely, grass
ba-pe	Winter barly, peas	lu-gr	Lupin, grass	swh	Spring wheat
be	Beets/carrots	oa	Oates	tc	Triticale
cc	Catch crop	oa-clgr	Oates, clover, grass	wc	Whole crop
ch	Chicory	oa-gr	Oates, grass	wc-gr	whole crop, grass
chgr	Chicory, grass	pe	Peas	wh	Winter wheat
clgr-wc	Clover, grass, wholecrop	po	Potatoes	wh-chgr	winter wheat, clover, grass
clgr	Clover, grass	ra	Rape	wh-gr	Winter wheat, grass
fa	Fallow	rd	Radish		
fa-clgr	Fallow, clover, grass	ry	Rye		

takes the value 1, and all others are set to 0. The list of possible crops can be found in Table 2.

The effect of tillage on the microarthropod community is thought to exhibit a non-linear decay over time. Therefore, we apply the transformation

$$\text{tillage}' = \left(\frac{10 - \text{months since tillage}}{10}\right)^{i} \quad (1)$$

to the variables representing tillage. The parameter i depends on the type of tillage: $i = 2$ for deep to subshallow tillage and $i = 4$ for shallow tillage.

The target attributes describe the observed microarthropod community, which is quantified by measuring the abundance of 43 species. Of these, 4 belong to the Acari group (mites) and 39 belong to the Collembola group (springtails). The species included in both groups are listed in Table 3.

To measure the abundance of each species, soil samples were collected within a 20 m × 20 m area of the field, with a distance of 5 m between the individual samples. Sampling was performed in the upper 5.5 cm soil layer and the sampling containers measured 6 cm in diameter. Sampling was done using a split soil corer and extraction was performed using a MacFadyen high gradient heat extractor.

Based on the data describing the microarthropod community, three target attributes are constructed: the total abundance of the Acari group, the total abundance of the Collembolan group, and the Shannon-Wiener biodiversity (2).

$$H = -\sum_{i=1}^{S} p_i \log_2 p_i \quad (2)$$

where p_i is the proportion of species i in the sample and S the total number of species.

3. Regression and Model trees

The models that we present in this paper are single and multi-objective regression trees and model trees. The following two sections briefly describe the theory behind such models and the systems that we have used for constructing them.

3.1. Regression trees, multi-objective regression trees and the Clus system

Regression trees are predictive models capable of modelling a numeric target (Breiman et al., 1984). Examples of regression trees can be found in Figs. 5–7. The internal nodes of a regression tree contain tests on the input attributes and the leaves store the predictions. The prediction for a new data record is obtained by sorting it down the tree, starting from the root (the top of the tree). For each internal node encountered on the path, the test stored in the node is applied to the given record, and if it succeeds, the record is sorted down the left subtree; if it fails, the path continues along the right

Table 3
The observed species (Acari group—mites and Collembola group—springtails)

Abbreviation	Species	Abbreviation	Species
Acari group (mites)			
Crypt	*Cryptostigmata*	Ast	*Astigmata*
Prost	*Prostigmata*	Meso	*Mesostigmata*
Collembola group (springtails)			
Iang	*Isotoma anglicana*	Hniti	*Heteromurus nitidus*
Ipalu	*Isotomurus palustris*	Tquad	*Stenaphorura quadrispina*
Hdent	*Ceratophysella denticulata*	Nmini	*Neelus minimus*
Hsuc	*Ceratophysella succinea*	Saure	*Sminthurinus aureus*
Xarma	*Hypogastrua* sp.	Fspino	*Folsomia spinosa*
Llanu	*Lepidocyrtus lanuginosus*	Cterm	*Cryptopygus thermophilus*
Lcyan	*Lepidocyrtus cyaneus*	Will	*Willemia* sp.
Seleg	*Sminthurinus elegans*	Ocinct	*Orchesella cincta*
Onych	*Protaphorura* sp.	Owillo	*Orchesella villosa*
Sviri	*Sminthurus viridis*	Nmusco	*Neanura*
Sminsp	*Smint*. sp.	Psexoc	*Pseudosinella sexoculata*
Tull	*Mesaphorura* sp.	Iprod	*Isotomodes productus*
Inot	*Isotoma notabilis*	Iarma	*Isotomodes armata*
Entosp	*Entomobrya* sp.	IBiset	*Isotomodes bisetosus*
Fmirab	*Friesea mirabilis*	Fquad	*Folsomia quadrioculata*
Ffim	*Folsomia fimetaria*	Icilia	*Isotomurus* sp.
Palba	*Pseudosinella alba*	Tomosp	*Tomocerus* sp.
Bparv	*Brachystomelle parvula*	Tflav	*Tomocerus flavescens*
Apygm	*Anurida pygmaea*	Tminor	*Tomocerus minor*
Iminor	*Isotomiella minor*		

subtree. The resulting prediction is the value stored in the leaf where the path ends.

Multi-objective regression trees (Blockeel et al., 1998) generalize regression trees in the sense that they can predict a value for more than one target attribute. Therefore, instead of storing a single numeric value, the leaves of a multi-objective tree store a vector. Each component of this vector is a prediction for one of the target attributes. Fig. 4 shows an example of a multi-objective regression tree predicting the target attributes Acari abundance, Collembola abundance and biodiversity.

A (multi-objective) regression tree is usually constructed with a recursive partitioning algorithm from a training set of records, i.e., records that include measured values for the target attributes. Such an algorithm starts by selecting a test for the root node. Based on this test it partitions the data into a training set for the left (records for which the test succeeds) and right (records for which the test fails) subtree, and then recursively repeats the same procedure to construct the left and right subtree. The partitioning process stops if the number of records in the induced subsets is smaller than some predefined value *minrec*. In that case, a leaf is generated storing a vector with as components the mean of the target attributes (over the records stored in the leaf).

The test selected for a given node is the one that minimizes a heuristic computed on the training data. The goal of the heuristic is to guide the algorithm to small trees with good predictive performance. In this paper, we apply the system Clus (Blockeel and Struyf, 2002) for constructing (multi-objective) regression trees. In Clus, the heuristic is the sum of the variations in the induced subsets, where variation is measured as $\sum_j^T \sum_i^N \left(x_{i,j} - \overline{x_j}\right)^2$, with T the number of target attributes, N the number of records in the subset, $x_{i,j}$ the value of target attribute j of the ith record in the subset, and $\overline{x_j}$ the subset mean of attribute j. A low intra-subset variation results in accurate predictions.

After a regression tree is constructed, it is common to prune it, i.e., to replace some subtrees by leaves, in order to improve predictive accuracy and/or interpretability. We choose the pruning method that is

proposed by Garofalakis et al., 2003. Essentially, this is a dynamic programming optimization method that selects a subtree from the constructed tree with at most *maxsize* nodes and minimum training set error (mean squared error, summed over all target attributes). We employ this particular method because we are interested in obtaining small and interpretable trees, i.e., we set *maxsize* to a manageable value and the algorithm then returns the best subtree satisfying this size constraint.

3.2. Model trees and the M5′ system

We compare the regression trees built by Clus to model trees (Quinlan, 1992). Model trees differ from regression trees in the sense that the leaves do not contain numeric values, but linear regression models. In order to obtain a prediction with a model tree, the given record is sorted into a leaf and then the corresponding linear model is applied to obtain the actual prediction. Model trees are generally more accurate than regression trees, but more difficult to interpret because of the linear models. In the experiments, we apply the M5′ (Wang and Witten, 1997) system for inducing model trees, which is available in the Weka (Witten and Frank, 1999) data mining toolkit. Note that M5′ can only generate single-objective trees and that it uses a heuristic and pruning method that differs from the ones employed by Clus.

4. Experiments

In this section we discuss the models that have been obtained by applying the modelling techniques presented in Section 3 to the available data. We first describe the experimental setup. Next, we compare the models obtained with Clus and M5′. The section ends with a discussion of the knowledge entailed by the models.

4.1. Setup

As discussed in Section 2, the data set used in this study contains three target attributes: Acari abundance, Collembola abundance and biodiversity. We compare two settings: single-objective regression and multi-objective regression. In the single-objective setting, three regression trees are constructed: one predicting Acari abundance, one Collembola abundance and one biodiversity. In the multi-objective setting a single tree predicts all of these three target attributes at once. While multi-objective trees can yield a lower predictive performance, they have the important advantage that they are easier to interpret. Obviously interpreting one single tree is less difficult than three different trees. Moreover, the multi-objective model allows one to identify conditions that have different effects on target attributes, e.g., if a particular leaf predicts an Acari abundance above average and a biodiversity below average, then one can conclude that the conditions describing the leaf have a positive effect on Acari species, but a negative effect on other species and biodiversity.

Since we are interested in obtaining simple and understandable trees, we constrain the number of nodes in a tree to be less than *maxsize*. To be able to quantify the possible performance loss incurred by smaller trees we experiment with different values of this parameter: 400, 200, 100, 50, 20 and 10. In all experiments, the *minrec* parameter of Clus was set to 5. For the multi-objective trees, we also enabled normalization, which internally transforms the target attributes by subtracting the mean and dividing them by their standard deviation. In this way, each target attribute has a similar contribution in the computation of the heuristic and in the error estimate used by the pruning method. All other parameters are set to their default values. We also perform a number of experiments with M5′ where we vary the *minrec* parameter (in order to find a model tree with an acceptable compromise between accuracy and size).

The predictive performance of each of the models is estimated with ten-fold cross validation. The error measures used are: relative mean absolute error RMAE, relative root mean squared error RRMSE and the Pearson correlation coefficient *r*. The relative measures are obtained by dividing the error of the model by the error of a baseline model that always predicts the mean.

5. Results

Table 4 presents the results. In order to be able to better compare the different settings, Figs. 1–3 show the correlation coefficients for single and multi-objective regression, for each of the target attributes. The results

Table 4
The error rate and size for multi-objective and singe objective regression trees, and for model trees (RMAE—relative mean absolute error, RRMSE—relative root mean squared error, *r*—correlation)

Max tree size (number of nodes)	Measure	Multi-objective regression			Single-objective regression		
		Acari	Collembola	Biodiversity	Acari	Collembola	Biodiversity
400	RMAE	0.602	0.599	0.729	0.592	0.617	0.732
	RRMSE	0.701	0.701	0.733	0.693	0.714	0.736
	r	0.716	0.715	0.686	0.724	0.704	0.684
	#Leaves		197		200	200	197
200	RMAE	0.631	0.612	0.731	0.600	0.621	0.734
	RRMSE	0.714	0.707	0.740	0.694	0.713	0.742
	r	0.703	0.709	0.677	0.722	0.704	0.678
	#Leaves		100		100	100	100
100	RMAE	0.668	0.654	0.769	0.632	0.639	0.743
	RRMSE	0.731	0.730	0.772	0.713	0.718	0.751
	r	0.683	0.684	0.637	0.705	0.698	0.664
	#Leaves		50		50	50	50
50	RMAE	0.703	0.686	0.829	0.682	0.687	0.789
	RRMSE	0.758	0.738	0.828	0.735	0.738	0.788
	r	0.653	0.675	0.562	0.680	0.678	0.617
	#Leaves		25		25	25	25
20	RMAE	0.791	0.770	0.906	0.792	0.733	0.855
	RRMSE	0.819	0.777	0.913	0.800	0.755	0.847
	r	0.572	0.629	0.410	0.600	0.655	0.531
	#Leaves		10		10	10	10
10	RMAE	0.890	0.811	0.925	0.878	0.801	0.921
	RRMSE	0.874	0.791	0.936	0.877	0.793	0.918
	r	0.484	0.611	0.351	0.482	0.610	0.399
	#Leaves		5		5	5	5
M5' Best model tree	RMAE				0.733	0.647	0.740
	RRMSE				0.641	0.712	0.751
	r				0.680	0.701	0.668
	#Leaves				13	17	27
M5' Best regression tree	RMAE				0.700	0.718	0.776
	RRMSE				0.787	0.774	0.784
	r				0.618	0.637	0.625
	#Leaves				31	27	45

show that the performance of multi-objective regression is comparable to that of single-objective regression, especially for large trees. The difference increases if the trees are heavily pruned (*maxsize* < 50). This effect is most noticeable for biodiversity. The results furthermore confirm that the error of both methods increases if *maxsize* is decreased, especially for biodiversity and Acari abundance.

If we compare the regression trees constructed by Clus to the regression trees of M5′, then we observe that Clus performs better for trees of comparable size. For example, the tree of M5′ predicting Acari abundance with 31 leaves has a correlation of 0.618, which is in between the 0.680 obtained by Clus for a tree with 25 leaves and the 0.600 obtained for a tree with only 10 leaves. This effect is probably caused by the pruning method employed by Clus: the tree with 10 leaves is the 'best' possible subtree of that size. On the other hand, the model trees of M5′ perform better than the regression trees of both systems (when comparing trees

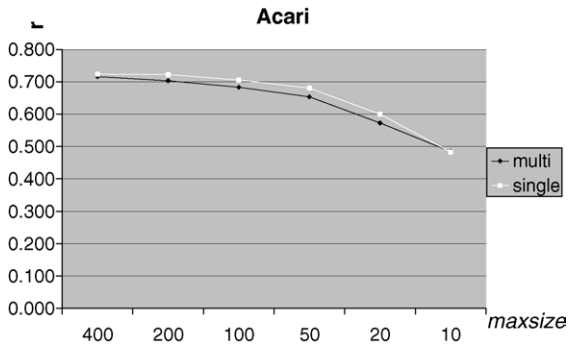

Fig. 1. The correlation coefficient *r* for multi-objective and single-objective trees predicting Acari abundance for different values of the pruning parameter maxsize.

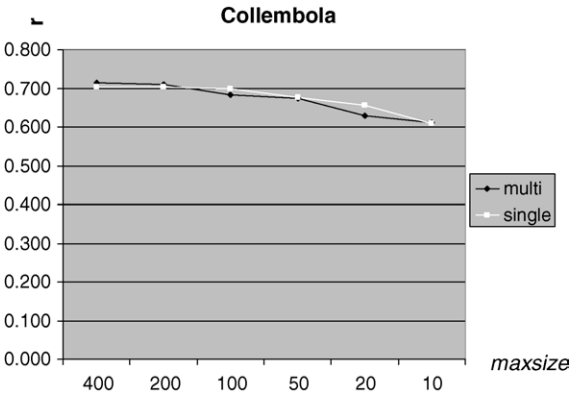

Fig. 2. The correlation coefficient *r* for multi-objective and single-objective trees predicting Collembola abundance for different values of the pruning parameter maxsize.

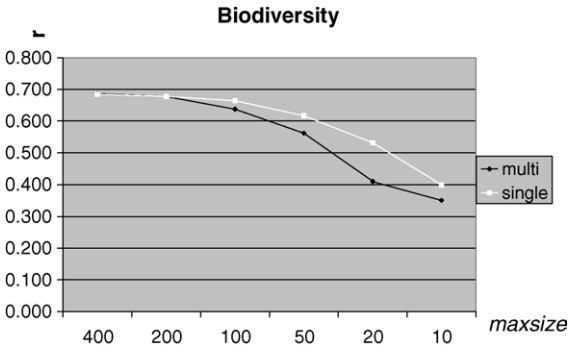

Fig. 3. The correlation coefficient *r* for multi-objective and single-objective trees predicting biodiversity for different values of the pruning parameter maxsize.

of similar size). Model trees are however more difficult to interpret because of the linear models in the leaves.

5.1. Interpretation of the obtained models

In the previous section we have shown that almost the same predictive performance is obtained with a single multi-objective tree as with three separate single-objective trees. In this section, we study the structure of the trees to identify important factors influencing the microarthropod community. Consider the multi-objective tree depicted in Fig. 4 (created with the pruning parameter *maxsize* = 50 nodes). It shows for example the following.

- Soil type and the age of the current situation are the most important factors for all three modelled measures: sandy soils and an old age of the current situation provide the best conditions for the soil microarthropods.
- While the age of the current situation (in not extremely sandy soils and summer sown crops) strongly influences the abundances of both Acari and Collembola it does not influence biodiversity, which means that most species profit in similar amounts. The same is true for the sampling time (spring samples have lower abundances but about the same biodiversity than later samples when other conditions do not change).
- On the other hand in sandy soils and in a young situation (less than 1 month after sowing or harvesting) the Acari thrive, while the Collembola struggle.
- Fertilization can have a strong negative impact on both Acari and Collembola abundance, while it has only a medium negative effect on their biodiversity.
- Rye with undersown clover-grass (at least 4 months ago and in not extremely sandy soils) has a strong positive effect on the Acari abundance and at the same time a strong negative effect on the Collembolan abundance and biodiversity.

We can furthermore compare the multi-objective regression tree (Fig. 4) with the three single-objective regression trees shown in Figs. 5–7. We observe that the multi-objective regression tree does not closely resemble any of the single-objective trees. The root node tests on the same attribute as in the Acari tree (Fig. 5), however the test condition is not identical. Such differences

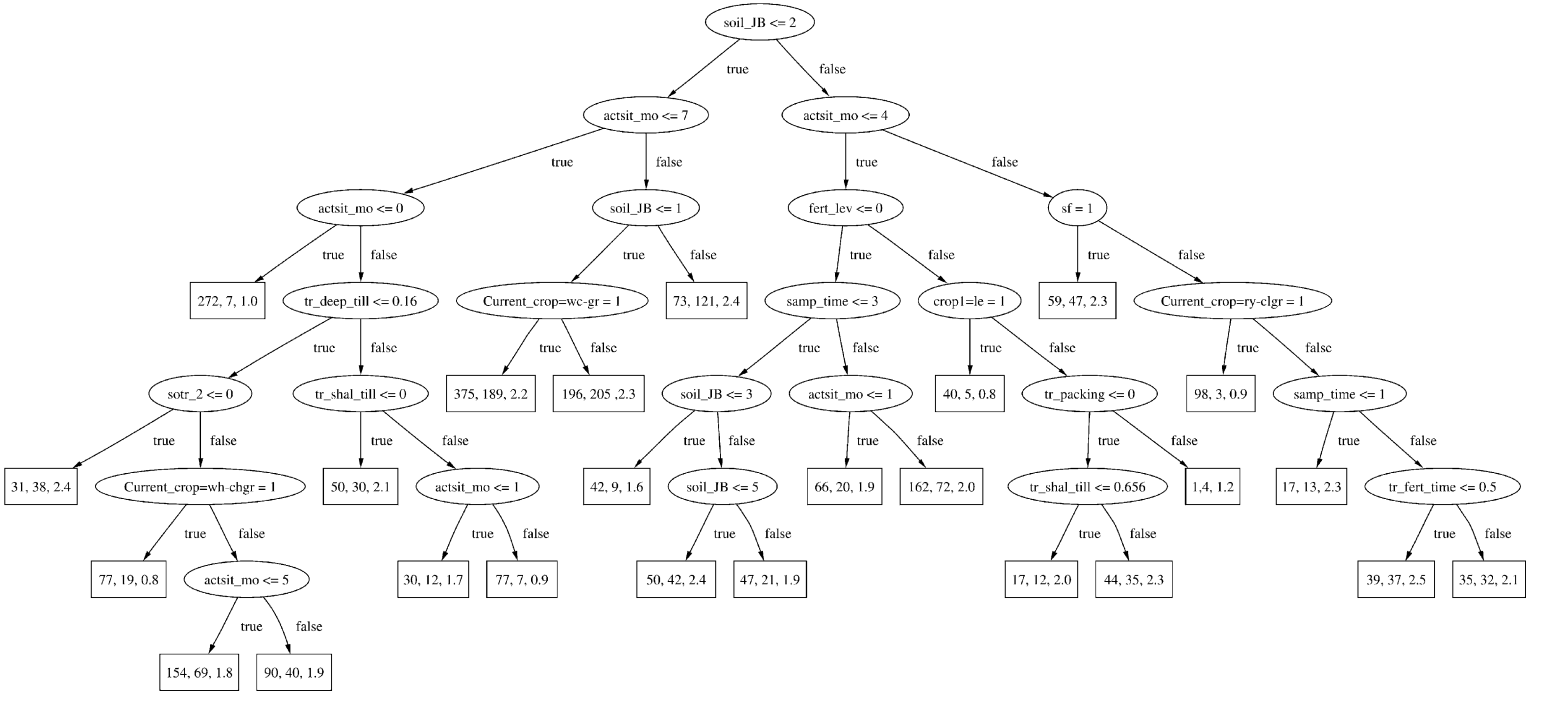

Fig. 4. The multi-objective regression tree modelling Acari abundance, Collembola abundance and biodiversity, created with pruning setting maxsize = 50 nodes. The numbers in the leaves are the number of Acari divided by 1000, number of Collembola divided by 1000 and biodiversity, respectively. The average values of the target attributes over the entire data set are: 48724, 33030 and 2.06.

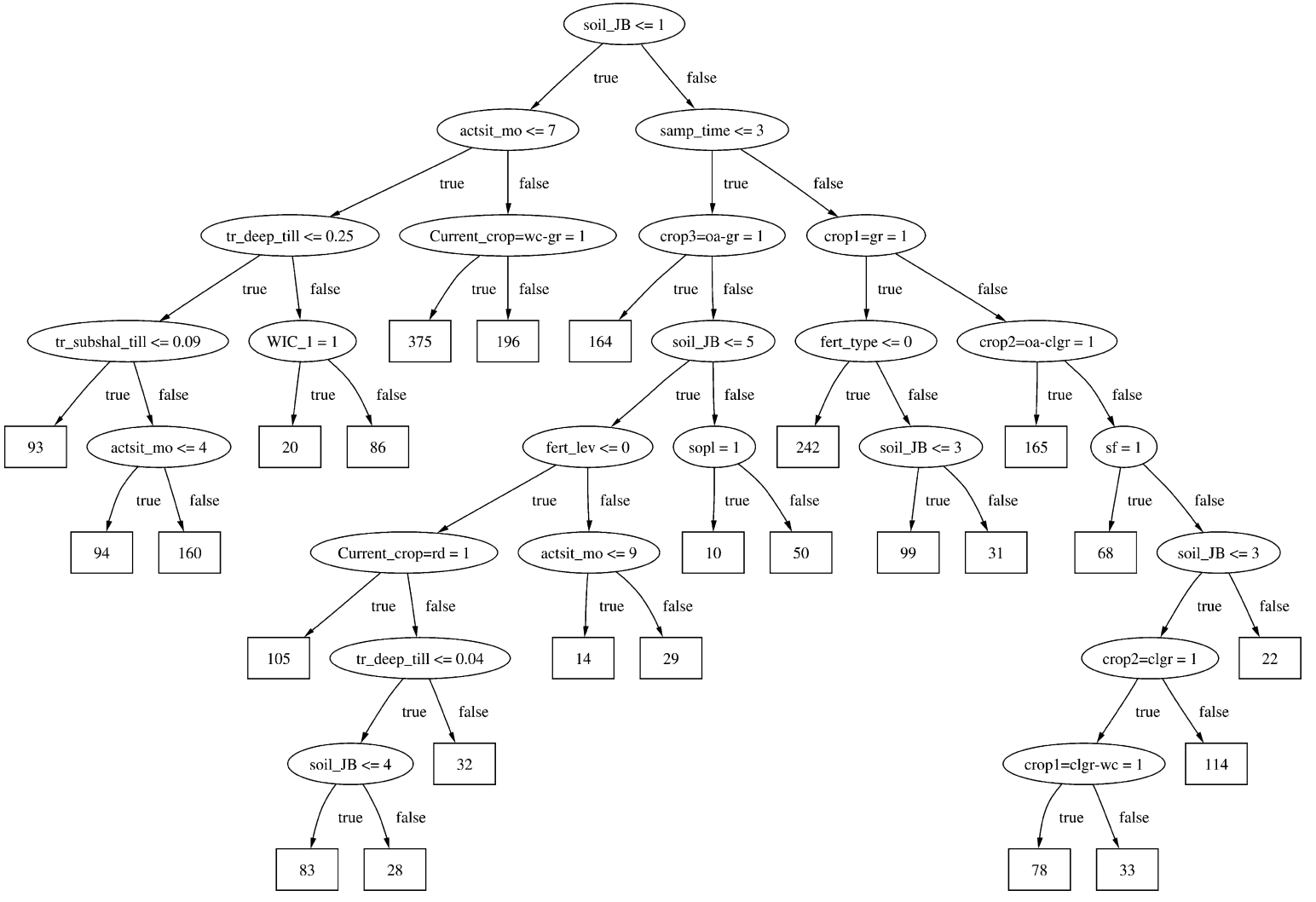

Fig. 5. The regression tree modelling Acari abundance, created with pruning setting maxsize = 50 nodes. The numbers in leaves are number of Acari divided by 1000.

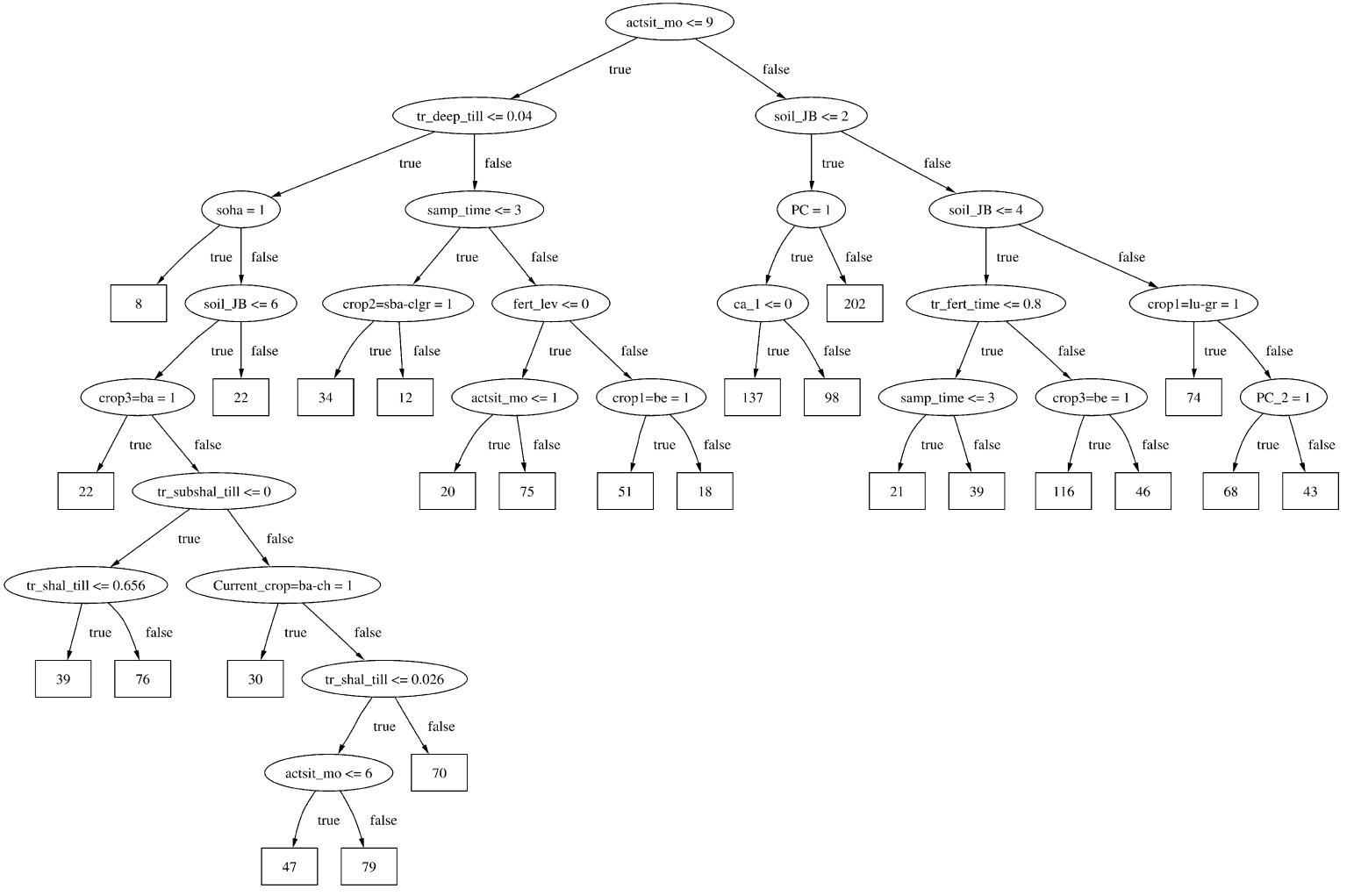

Fig. 6. The regression tree modelling Collembola abundance, created with pruning setting maxsize = 50 nodes. The numbers in leaves are number of springtails divided by 1000.

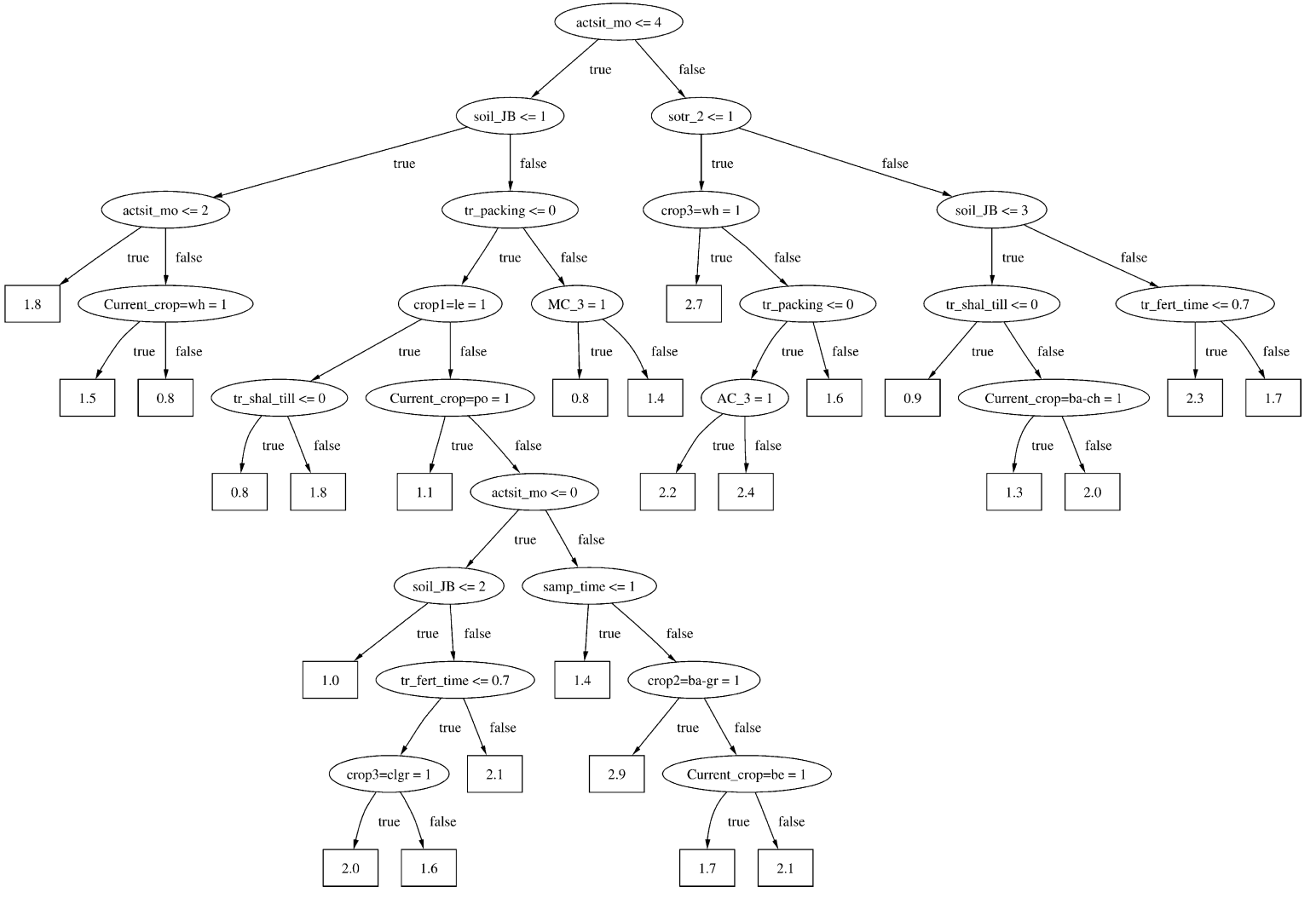

Fig. 7. The regression tree modelling biodiversity, created with pruning setting maxsize = 50 nodes.

in trees are to be expected because decision trees are known to be unstable, e.g., for a given data set, typically many trees exist that have a comparable predictive performance. Some similarities between the trees nevertheless do exist. The attributes that Clus selected as splitting criteria in the nodes are similar in all four trees. This confirms that soil type, age of the current situation, tillage, the use of crops belonging to the family of grasses, etc. are important for the community of soil microarthropods.

6. Conclusions

We have modelled the community of soil microarthropods in agricultural soil with machine learning methods based on data describing chemical, biological and mechanical actions on the fields. We used the obtained models to identify the most important parameters influencing the abundance of soil mites and springtails and the biodiversity of soil microarthropods. In particular, we show that the most important parameters are the soil type, the age of the current situation, and the different forms of tillage. We also identified the different effects of one action on several agricultural measures: some actions have a positive effect on one type of soil microarthropods and a negative effect on other types. We gained knowledge that will help us in further modelling and, in the end, in building a decision support system for the management of farms. We have also shown that machine learning models can be used in multiple ways: they can be used to predict accurate values, to gain new knowledge about the domain at hand, and to assist us in obtaining knowledge from domain experts.

Acknowledgements

This work is supported by the ECOGEN project funded by the Fifth European Community Framework Programme: Quality of Life and management of living resources contract no QLK5-CT-2002-01666 and DARCOF, Nature quality in organic farming.

References

Blockeel, H., Struyf, J., 2002. Efficient algorithms for decision tree cross-validation. J. Mach. Learn. Res. 3 (Dec), 621–650.

Blockeel, H., De Raedt, L., Ramon, J., 1998. Top-down induction of clustering trees. In: Shavlik, J. (Ed.), Proceedings of the 15th International Conference on Machine Learning, pp. 55–63.

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

Demšar, D., Džeroski, S., Krogh, P.H., Larsen, T., 2003. Identifying the most important agricultural factors for the soil community of microathropods. In: Proceedings of the International Electrotechnical and Computer Science Conference, Ljubljana, Slovenia.

Garofalakis, N., Hyun, D., Rastogi, R., Shim, K., 2003. Building decision trees with constraints. Data Mining Knowl. Discov. 7 (2), 187–214.

Krogh, P.H., 1994. Microarthropods as bioindicators. A study of disturbed populations. PhD thesis Ministry of the Environment and Energy. National Environmental Research Institute, Silkeborg.

Quinlan, J.R., 1992. Learning with continuous classes. In: Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence, pp. 343–348.

Steen, E., 1983. Soil animals in relation to agricultural practices and soil productivity. Swedish J. Agric. Res. 13, 157–165.

Wang, Y. and Witten, I.H., 1997. Induction of model trees for predicting continuous classes. Proceedings of the poster papers of the ECML 97. University of Economics, Faculty of Informatics and Statistics, Prague.

Witten, I.H., Frank, E., 1999. Data Mining: Practical machine learning tools with Java im-plementations. Morgan Kaufmann, San Francisco.