Available online at www.sciencedirect.com
SCIENCE DIRECT®

Ecological Modelling 191 (2006) 144–158

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

# Automatic construction of concept hierarchies: The case of foliage-dwelling spiders

Martin Žnidaršič[a,*], Aleks Jakulin[a], Sašo Džeroski[a], Christian Kampichler[b]

[a] *Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia*
[b] *División Académica de Ciencias Biológicas, Universidad Juárez Autónoma de Tabasco, 86100 Villahermosa, Mexico*

Available online 29 September 2005

**Abstract**

In this paper, we present the hierarchical variable dependencies that were obtained from raw data with the use of two machine learning techniques on an ecological data set. The data set contains features of field margins and the corresponding number of spider species inhabiting them. This data set was used before by domain experts to construct a fuzzy qualitative model with hierarchical variable dependencies, which we use for comparison with our results. One of the machine learning methods constructs a hierarchical structure similar to the one in the experts' model, while revealing some additional interesting relations of environmental features with respect to the number of spider species. The other method constructs a different hierarchy from the one proposed by the experts, which, according to our classification performance experiments, might be even more appropriate.

*Keywords:* Data-based hierarchy construction; Interaction analysis; Feature construction; Hierarchical models; Spiders; Field margins

## 1. Introduction

Ecological domains are complex with interdependent variables and hidden relations that are difficult to explain. These characteristics indicate that ecology experts might benefit from the use of machine learning methods. Machine learning can be used to confirm hypotheses or to discover new relations, thus gaining insight into vast amounts of data. However, the most demanding parts, evaluation and explanation, have to be done by experts.

In this paper, we present the hierarchical structures, rules and relations that were learned from raw data with the use of two machine learning techniques. Some information is given about these techniques and the way they were used to construct hierarchies of variables from the data. The data set we used contains the measurements of variables that might influence the diversity of foliage-dwelling spiders in field margins. The meaning of the variables in this data set is described in Section 2. This data set was used before by domain experts (Kampichler et al., 2000) to construct a fuzzy qualitative model of hierarchical variable dependencies. The model was mainly constructed manually with some use of data analysis techniques.

[*] Corresponding author.
*E-mail address:* martin.znidarsic@ijs.si (M. Žnidaršič).

0304-3800/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2005.08.023

Instead, we constructed hierarchical models of this data set completely automatically in two ways, using interaction analysis and function decomposition. Experimental work with both methods was performed with the Orange (Demšar and Zupan, 2004) data mining suite. Interaction analysis (Jakulin and Bratko, 2003) comprises a set of tools for identifying interactions among the variables in data. Interactions are dependencies between variables that deserve closer investigation. In prediction tasks we are especially interested in three-way interactions between two independent variables (such as mean margin width or the number of small plants) and the outcome (number of spider species). There are two types of three-way interactions: the two variables may be synergistic in the sense that controlling for both of them unlocks an otherwise hidden pattern. On the other hand, the two variables may be redundant, if they both provide the same information. The interaction dendrogram, which summarizes the three-way interactions found in data, yields a structure similar to the model that was built by domain experts. It also provides some clues about which variables are equally appropriate to be used in the same place of the model structure, as well as additional clues about variable dependencies (some of which can be due to noise).

The second method we used, function decomposition, is a member of a larger family of constructive induction methods, which focus on discovering novel concepts in data. We employed the hierarchy induction tool HINT (Zupan et al., 1999). This method is also able to create new variables and rules to compute their values, not only the hierarchical structure. However, it tends to be sensitive to noise in the training data. The resulting hierarchy did not exactly match the hierarchy of the experts, but experiments indicate that the constructed model is valid with respect to the given data, and could therefore be interesting to domain specialists.

Interaction analysis could be useful for construction of preliminary models, thus saving experts valuable

Table 1
Variables that characterize the margins

| | |
|---|---|
| margin_density | Margin density (linear *m* of margins per ha) |
| mean_margin_width | Mean width of margins |
| margin_width | Width of the strip |
| disturbances | Number of disturbance events (ploughing, mowing, etc.) per year |
| herb_cover | Cover of herbs (%) |
| herbs | Proportion of total plant biomass (%) of herbs (estimated in the field) |
| legumes | Proportion of total plant biomass (%) of legumes (estimated in the field) |
| grasses | Proportion of total plant biomass (%) of grasses (estimated in the field) |
| sedges_brooms | Proportion of total plant biomass (%) of sedges and brooms (estimated in the field) |
| herb_legum | Herbs + legumes |
| grass | Grasses + sedges_brooms |
| grass_cover | Cover of grasses (%) |
| plant_cover | Total cover of plants (%) |
| small_plants_spp | Number of species <20 cm |
| branched_total_spp | Number of species >20 cm with branched architecture |
| linear_total_spp | Number of species >20 cm with linear architecture |
| branched_spp_persist | Number of species >20 cm with branched architecture, persistent until autumn |
| linear_spp_persist | Number of species >20 cm with linear architecture, persistent until autumn |
| branched_spp_dis | Number of species >20 cm with branched architecture, disappear by autumn |
| linear_spp_dis | Number of species >20 cm with linear architecture, disappear by autumn |
| therophyte | Proportion of species with life-form therophyte (%) |
| geophyte | Proportion of species with life-form geophyte (%) |
| hemicryptophyte | Proportion of species with life-form hemicryptophyte (%) |
| chamaephyte | Proportion of species with life-form chamaephyte (%) |
| phanerophyte | Proportion of species with life-form phanerophyte (%) |
| slope_direction | Direction of slope |
| spider_species | Number of spider species |

The variables margin_density and mean_margin_width characterize each of the seven study areas, whereas all other variables characterize single margins (plot size 50 m × 1 m).

time. It is also possible that both methods could help identify complex concepts, ones that would be hard to discover manually. The results suggest that integrating the approaches of interaction analysis and function decomposition might be a promising direction for further work: i.e., one might extract the hierarchy with interaction analysis and then use HINT to find the rules in its internal nodes.

## 2. Foliage-dwelling spiders dataset

Field margins are grassy strips between arable fields or meadows. They contain only single shrubs and trees and they are not cultivated or ploughed. Field margins support beneficial arthropods, e.g. predators of crop pests, and can be of potential value to species of conservational importance. Barthel and Placher (1996) and Anderlik-Wesinger et al. (1996) studied foliage-dwelling spider occurrence in field margins in seven agricultural areas in southern Germany (see Barthel and Placher (1996), Fig. 1) by standard visual inspection in the herbaceous vegetation. They used between 12 and 17 plots (1 × 50 m) in each area yielding a total of 96 plots and analysed the influence of margin and landscape characteristics (see list of variables in Tables 1–4) on the species number of spiders per plot. By applying correlation and regression analyses they identified margin density, margin width, percent cover of herbaceous plants and the number of mechanical perturbations as the main factors influencing spider diversity. Using these variables, Kampichler et al. (2000) elaborated a fuzzy rule-based model, increasing the predictive power for unseen field-margins in comparison with the original multiple regression model (Anderlik-Wesinger et al., 1996) which read $\text{species\_number} = 8.27 - 2.13\,(1/\text{margin\_width}) + 0.02\,\text{margin\_density} - 1.53\,\text{disturbance} + 0.05\,\text{herb\_cover}$. In the fuzzy model a rule set relates disturbance (two fuzzy sets) and margin width (three fuzzy sets) to an intermediate variable called habitat persistence (six singletons); another rule set relates habitat persistence (six fuzzy sets) and margin density (three fuzzy sets) to another intermediate variable called colonisation potential (eight singletons); finally a third rule set relating colonisation potential (seven fuzzy sets) and herb cover (three fuzzy sets) predicts species number of spiders (six singletons) (see Kampichler et al. (2000), Fig. 2). The model is ecologically plausible and reflects domain knowledge on the biology of spiders in field-margins obtained in various empirical studies (e.g., Gibson et al., 1992; Thomas et al., 1992; Baines et al., 1998).

Table 2
The following variables are the mean values of the indices by Ellenberg (Ellenberg et al., 1992)

| | |
|---|---|
| light | min. exclusively in deep shade |
| | min. exclusively in bright sunlight |
| temperature | max. exclusively in highest regions of European mountains |
| | max. exclusively on the warmest locations of central Europe |
| continentality | min. center of distribution in westernmost Europe |
| | max. only in easternmost parts of central Europe |
| moisture | min. exclusively on very dry soils |
| | max. mostly on very wet soils |
| acidity | min. exclusively on acid soils |
| | max. exclusively on calcareous soils |
| nutrient | min. on sites with lowest N concentrations |
| | max. on sites with highest N concentrations |

Each plant species has an indicator value for a number of factors like light, temperature, etc. ranging from 0 [very low] to 9 [very high], for example “light”: 0—exclusively in deep shadow, 9—exclusively in open habitats. Mean values calculated by presence/absence of species without weighting by abundance.

Table 3
Cover (estimated in the field as a vertical projection of the plants onto the soil surface, thus the sum is >100%)

| | |
|---|---|
| herb_layer | Cover of all plants in the herb layer |
| soil_layer | Open soil |
| litter_layer | Cover of plant litter |

Table 4
Cover (estimated in the field) in different heights (sum >100%)

| | |
|---|---|
| small_cover (cm) | <25 |
| med_cover (cm) | 26–50 |
| high_cover (cm) | 51–100 |
| vhigh_cover (cm) | >100 |

A short explanation[1] of the variables in the spiders data set is given in Tables 1–4.

## 3. Interaction analysis

A fundamental goal of data analysis is the identification of connections between variables. Connections in-

[1] Provided by Gabriele Andersik-Wesinger.

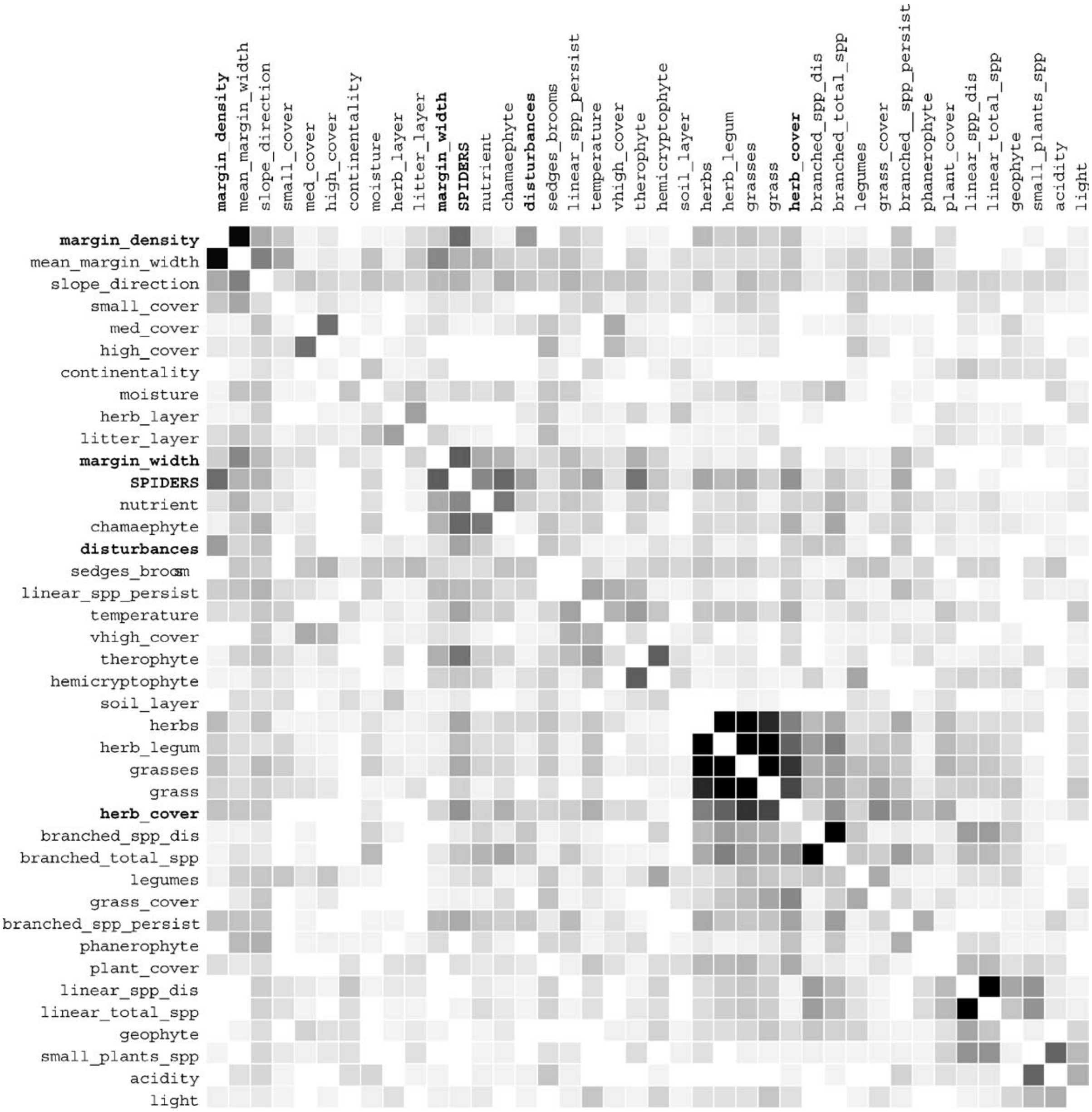


Fig. 1. The variable proximity matrix illustrates the strongest connections between individual variables: the darker the corresponding box the higher the mutual information between the variables.

dicate the existence of a pattern which can be identified by examining the connected variables simultaneously. If there is no such pattern, there is no need for connecting the variables, and we can safely assume them to be independent. Such a pattern-based view of connections subsumes the notions of a correlation as one of possible patterns, an interaction as the cause of a pattern, or an association as the psychological reaction to the discovery of a pattern. But how to define the existence of a 'pattern' mathematically? Interaction analysis (Jakulin and Bratko, 2003) is one approach to this task.

### 3.1. Methodology

Let us examine interaction analysis on a simple example of two variables, $A$ and $B$. There are two possibilities: we can assume that there is a pattern, and the resulting probabilistic model will take the form of $P(A, B)$. Alternatively, we can assume the two variables to be independent, and the model will be $P(A)P(B)$. Let us examine such a probabilistic model on an example, by assuming that $A$ and $B$ are discrete variables, and $a$ and $b$ are individual values that they can take. If $A$ denotes diversity, and $B$ the margin density, the resulting distribution of the 97 examples can be shown in a contingency table:

| B | A | | |
|---|---|---|---|
| | Low diversity (%) | High diversity (%) | Total margin density |
| High margin density | 46 | 25 | 71 |
| Low margin density | 3 | 26 | 29 |
| Total diversity | 49 | 51 | 100 |

In the center, the percentages describe the joint probability mass function $P(A, B)$, on the right $P(B)$, and in the bottom $P(A)$.

If we employ a reliable approach for obtaining all the probabilistic models, the less restricted model $P(A, B)$ can be used as a reference to which we compare the loss of the restricted model $P(A)P(B)$. In the above example, we can see a distinct deviation caused by the unexpectedly low percentage of low diversity of spider species in areas with low margin density, where we compare $P(\text{low\_diversity}, \text{low\_margin\_density}) = 0.03$ with $P(\text{low\_diversity})P(\text{low\_margin\_density}) = 0.14$. The measure of the pattern in interaction analysis is the loss caused by the assumption of variable independence in the model $P(A)P(B)$, relative to the dependence-assuming model $P(A, B)$. Employing the device of Kullback–Leibler divergence we can compute the 'distance' between both models:

$$D(P(A, B)||P(A)P(B)) = \sum_{a,b} P(a, b) \log_2 \frac{P(a, b)}{P(a)P(b)} \quad (1)$$

We can express the same in terms of the Shannon entropy, for two variables defined as $H(A, B) = -\sum_{a,b} P(a, b) \log_2 P(a, b)$. Entropy measures the lack of structure in $P(A, B)$, and is similar in meaning to variance, uncertainty or disorder. The lower the entropy $H(A, B)$ the stronger the pattern in $P(A, B)$ and the more telling the model. However, entropy does not provide the information about the reducibility of the model through the assumption of independence. For that purpose, we can restate Eq. (1) in terms of entropy, matching the definition of mutual information $I(A; B)$:

$$D(P(A, B)||P(A)P(B)) = H(A) + H(B) - H(A, B) = I(A; B) \quad (2)$$

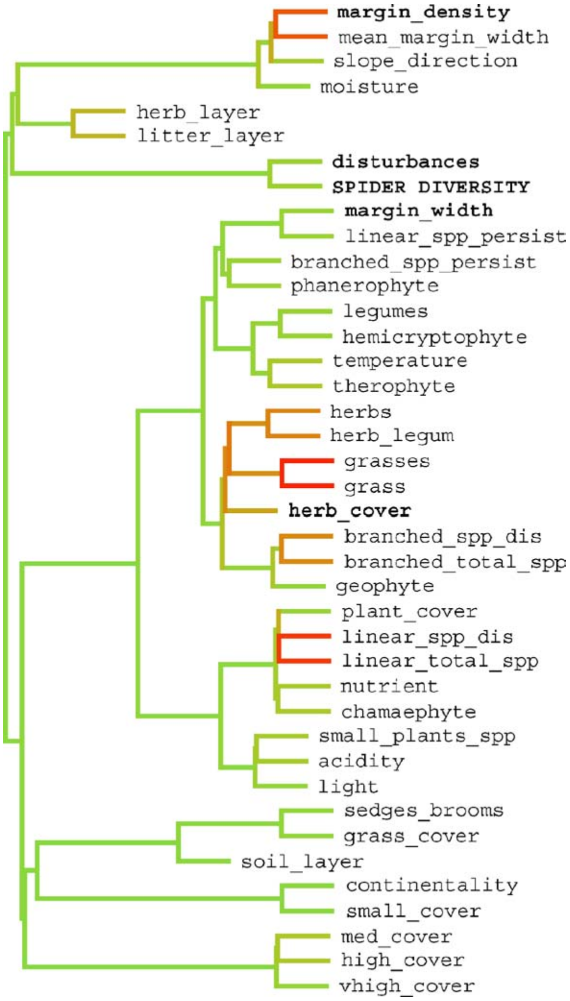


Fig. 2. The two-way interaction dendrogram summarizes the mutual information between individual variables. The color hue indicates the power of the interaction. Strong ones are red (darker) and the weak ones are green (lighter). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

If $I(A;B)$ is sufficiently high, we say that $A$ and $B$ interact. Mutual information can be seen as a measure of a two-way interaction between two variables. The role of mutual information is analogous to that of the non-parametric measure of correlation or association.

## 3.2. Results

### 3.2.1. Analysis of associations

In association analysis, we are primarily interested in mutual information between pairs of variables, without regard for any particular dependent variable. The dependent variable is considered to be equivalent to other independent variables. The task of association analysis helps to understand the general structure of variables in the data, which then helps separate the variables into groups. In the context of interaction analysis, the association between two variables is quantified with mutual information.

For the purpose of this analysis, each numerical variable was converted into a three-level discrete one. Each discrete value corresponds to a tercile in the distribution of the numerical values. Three-valued discrete variables allow relatively robust maximum likelihood estimation of probabilities in the resulting 9 groups and the available 97 instances. Fewer levels would cause a loss in pattern detection, while more levels would cause unreliable probability estimates. In this analysis, the outcome was handled in the same way as all other variables.

It is a well-known result that the joint entropy $H(A, B)$ is the upper bound for the mutual information $I(A;B)$. Therefore, we can express mutual information as a percentage of joint entropy. This percentage is a measure of proximity between two variables across all the instances. Without such a normalization, the number of variable values would influence the mutual information, and the complex variables would therefore appear to be more connected than simple variables. Such a normalization was originally proposed by Rajski (1961), who has also shown that the resulting normed mutual information obeys the triangle inequality and is therefore a metric.

Fig. 1 illustrates the resulting variable proximity matrix, but in itself it is not very clear. A popular approach to summarizing proximity matrices is hierarchical clustering (Struyf et al., 1997), and the result is illustrated in Fig. 2. The dendrogram is effectively an empirical taxonomy of the variables, created purely from the data. We can see three clusters of variables. On the top, there are mainly the variables that reflect human intervention. The variables that describe the density of margin vegetation are at the bottom. In between there is a large group of diverse variables that tend to characterize the composition of the plant community. We can clearly see that the spider diversity is primarily associated with the human intervention variables. The clustering is also meaningful since the interactions are to some extent transitive.

### 3.2.2. Analysis of predictors

Sometimes the objective of data analysis is to predict a particular outcome, in our example it is the number of spider species in a particular area. The outcome plays the role of the dependent variable, while other variables are considered to be independent. For example, we are not interested in the mutual information between the margin density and the orientation of the field, as this is not within the context of the outcome. Instead, we are only interested in the mutual information between independent variables and the dependent variable. It is easy to see that the mutual information also quantifies the reduction in the uncertainty of the outcome (spider diversity) as allowed by the information about the margin density: $I$(margin_density; diversity) = $D$($P$(diversity | margin_density) || $P$ (diversity)).

When we are predicting the outcome with two independent variables on real data, the information provided by these two variables might not be truly independent: variable independence is a modelling assumption, but rarely an intrinsic property of the data. For example, the second variable might provide some information that the first variable already provided about the outcome (consider the relevance of the variables temperature and altitude). Or, the second variable might affect the effect of the first variable on the outcome. It may turn out, for example, that the influence of grass coverage on the spider diversity is not independent of mean margin width.

One way of quantifying these deviations from independence is based on three-way interaction information (Jakulin and Bratko, 2003; McGill, 1954):

$$I(A;B;C) = I(A, B;C) - I(A;C) - I(B;C)$$

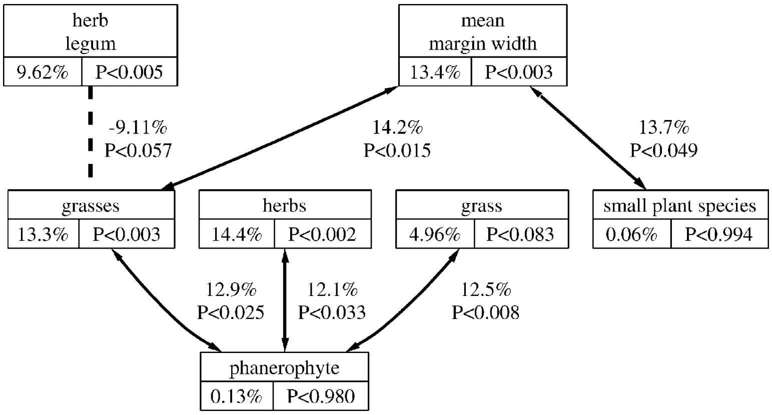


Fig. 3. The three-way interaction graph shows which pairs of variables interact with the outcome. The nodes identify individual variables, the undirected dashed edge indicates a redundancy, and the directed edges correspond to synergies between two variables and the outcome. Nodes are labeled with $I(A;C)$, the mutual information between the variable and the outcome, while interactions are labeled with $I(A,B;C)$, both expressed as a percentage of the outcome entropy $H(C)$. The $P$-values of the interactions are also shown.

Here, $I(A,B;C)$ is simply the mutual information between $A$ and $B$ together on one side, and $C$ on the other side: $I(A,B;C) = H(A,B) + H(C) - H(A,B,C)$. This way, interaction information compares the joint mutual information with the sum of individual mutual informations. We can also interpret $I(A;B;C)$ through the following formula: $I(A;B;C) = I(A;B|C) - I(A;B)$. Hence, $I(A;B;C)$ computes the change in mutual information between $A$ and $B$ if we control for $C$, and thereby evaluates the amount of $C$'s influence on the relationship between $A$ and $B$.

The interaction information may be either positive or negative. If it is distinctly positive, the pair of variables are opening a pattern that would otherwise not be detected by only considering individual variables' information about the outcome. For that reason, we say that the two variables are in a synergy with respect to the outcome. On the other hand, two variables may contribute partly the same information, and this situation of redundancy could result in duplication of evidence. For prediction, we can improve the performance by accounting for both redundancies and synergies. This is achieved by allowing for the dependence between the variables. Exercising the assumption of variable independence, which is made in many learning procedures, may result in underfitting for synergies and in overfitting for redundancies.

The usual practice of three-way interaction analysis is identifying the pairs of variables that are involved in the most distinct interactions with the dependent variable. To prepare our data for analysis, we employed the Fayyad-Irani (Fayyad and Irani, 1993) discretization with at least two variable values. The outcome was also discretized into two values with a split at the median value. Because we are estimating the three-way interactions, the number of discrete values per variable must be lower than in the earlier association analysis, not to risk unreliable probability estimates.

The resulting analysis in Fig. 3 shows a single negative interaction between the herb and legume coverage and the coverage of grasses. Usually, the two coverages sum up to approximately 100%, so only one of them is truly needed: once we know the coverage of grasses, the information about the herb and legume coverage eliminates merely $9.62 - 9.11 = 0.51\%$ of outcome (number of spider species) entropy. We can say that the latter provides negligible evidence once the coverage of grasses is known. Recently, a significance test has been proposed to measure the confidence in the reliability of the interaction (Jakulin and Bratko, 2004), and the results are included in the graph.

The proportion of phanerophytes (taller woody plants) appears to moderate a number of two-way interactions with the outcome. For example, the proportion of phanerophytes is not informative on its own (highly insignificant as a predictor), but controlling for the frequency of grass, it turns to be a significant predictor of spider diversity. Specifically, it turns out that there is much lower spider diversity when there are many phanerophytes with high grass coverage, in the absence of undergrowth. These two variables can be seen as predictors of the undergrowth. One dependence-assuming approach is to describe the undergrowth with a separate variable. On the other hand, the covering of grass has a positive interaction with mean margin width. These two variables together explain 13.4 + 14.2 + 13.3% of the outcome (rather than 13.4 + 13.3%). Another prominent positive interaction that was discovered is illustrated in more detail in Fig. 4. The number of small plant species appears to be uninformative on its own, but informative in combination with the mean margin width.

Interaction graphs become cluttered when the variables are many. For that reason, it is again possible to define proximity between variables, but this time in the context of the dependent variable. The resulting proximity measure is normed interaction information, the absolute value of interaction information expressed as a percentage of joint entropy: $d(A, B) = |I(A; B; C)|/H(A, B, C)$.

As before, we summarize the interactions using hierarchical clustering in a dendrogram that is shown in Fig. 5. Looking at the asterisks, it can be seen which variables are the best predictors (the top part of the dendrogram). For example, we can see that certain variables refer to the same aspects of the data: the influence of disturbances is also reflected in the presence or absence

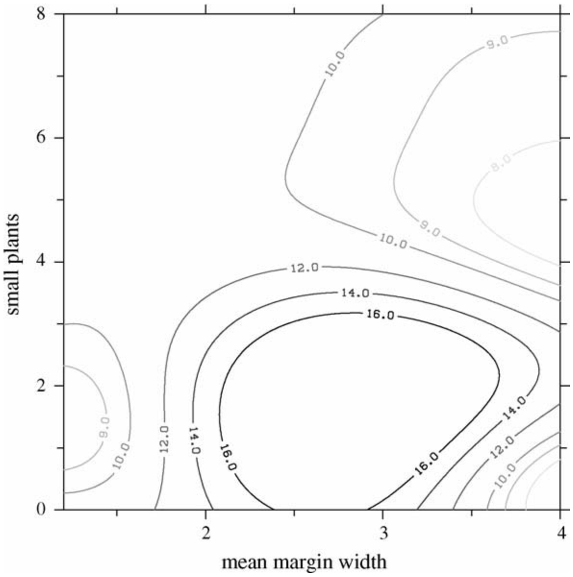


Fig. 4. The contour plot depicts the outcome variable (spider diversity) in dependence of two independent ones (number of small plant species and mean margin width). The influence of the two independent variables on the dependent one is highly nonlinear. The number of small plant species appears to be uninformative on its own, but informative in combination with the mean margin width. To generate the nonlinear regression model underlying the contours, we employed support vector regression with RBF kernels (Chang and Lin, 2001).

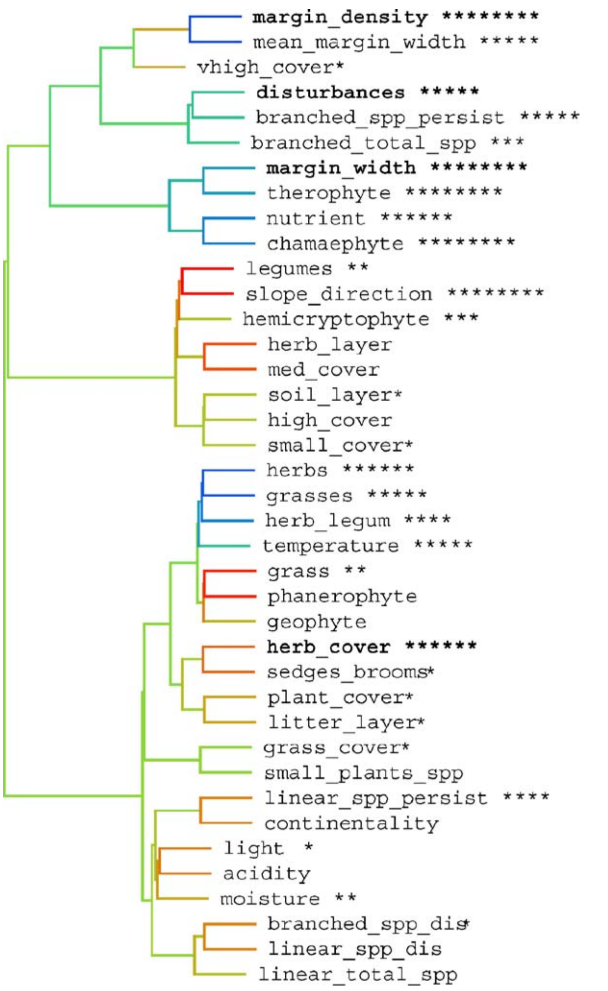


Fig. 5. The three-way interaction dendrogram shows both the two-way interactions between an independent and the dependent variable (denoted by asterisks), and the three-way interactions between two independent variables and the dependent one (denoted by the proximity in the dendrogram). The color hue indicates the type of the interaction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

of branched plants. Therefore, it is unclear whether it is the frequency of disturbances that affects the spiders directly or it is the indirect influence of disturbances through the number of branched plants.

## 4. Constructive induction

Using the tool HINT (Zupan et al., 1999), we performed data-driven constructive induction, construction of new concepts from the given variables in the spiders data set. HINT achieves this by function decomposition, a method that decomposes a complex function into a hierarchy of simpler ones.

### 4.1. Function decomposition and HINT

We will make a short explanation of function decomposition on a simple artificial example. Let us say, that we are interested in the rate of pollination of apple-trees and we decide to measure the pollination rate ($P$) along with some other factors that might be important in the pollination process. In this case let these factors be the temperature ($T$), the rainfall ($R$) and the distance to the nearest beehive ($B$). The rainfall and temperature can take values low and high, whereas the other two variables can take values low, medium and high.

The goal of our analysis would be to find out how the measured factors influence the target concept. In this respect, we can view the measurements as variables and the goal of the analysis as finding the function $F$ so that $P = F(T, R, B)$.

Suppose we have some measurements that are given in Table 5. Notice that these measurements do not totally specify our function.

Table 5
Function $F$

| T | R | B | P |
|---|---|---|---|
| low | low | low | medium |
| low | low | medium | medium |
| low | high | low | low |
| low | high | high | high |
| low | high | medium | medium |
| high | low | low | high |
| high | low | high | high |
| high | high | low | medium |
| high | high | medium | medium |

The function decomposition tries to build new concepts based on possible partitions of the variables. In this case, there are three non-trivial partitions and three corresponding decompositions of the variables: $P = F(T, H(R, B))$, $P = F(H(T, R), B)$ and $P = F(R, H(T, B))$. The concepts are built in a way that one value of a concept stands for every set of combinations of the values of its variables where such set produces the same values of the target concept given also the values of variables outside the concept. This process is explained in much more detail in the literature (Zupan et al., 1998).

The decompositions obtained this way are in Fig. 6. We can see that the new concept in the second decomposition has the lowest number of values. Feature decomposition methods usually select the decomposition that has concepts with the smallest set of values or the lowest number of examples in the definition of the functions. In our case, the second decomposition in Fig. 6 is the best one regarding both criteria.

If we had more variables in our dataset, the function decomposition would proceed further and build a more complex hierarchy, such as the one in Figs. 7–9. Such a hierarchy of concepts provides an insight into the relations among the measured variables and can be used as a first approximation of a rule-based prediction model. There have been some successful tests of this approach in the past (e.g., Zupan et al., 1999, 1998).

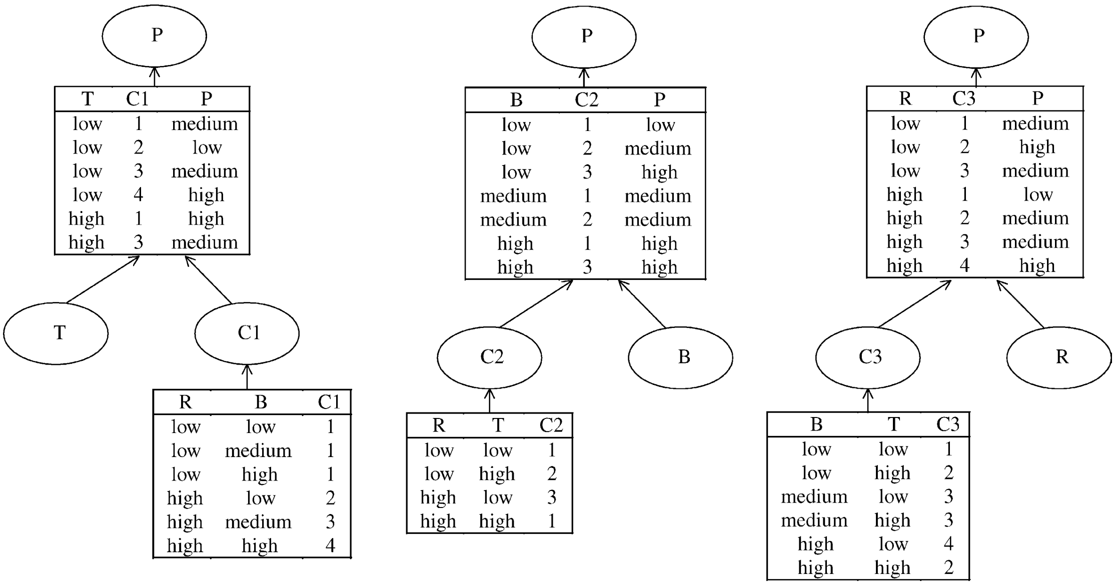
### 4.2. Results

We have used HINT for construction of concept hierarchies and models from the spiders data set with various settings. The hierarchies built from data set are shown and a model obtained from four variables is compared with the model previously built by the experts (Kampichler et al., 2000).

#### 4.2.1. Full hierarchy

A hierarchy construction from the whole data set was made first. The data had to be preprocessed to be used by HINT's methods, the variables with many missing values ('legumes', 'sedges and brooms') were discarded and the instances with missing values were removed, since missing values obstruct the constructive induction with HINT.

A hierarchical structure of the variables in the preprocessed spiders data set was built using the func-



| T | C1 | P |
|---|---|---|
| low | 1 | medium |
| low | 2 | low |
| low | 3 | medium |
| low | 4 | high |
| high | 1 | high |
| high | 3 | medium |

| R | B | C1 |
|---|---|---|
| low | low | 1 |
| low | medium | 1 |
| low | high | 1 |
| high | low | 2 |
| high | medium | 3 |
| high | high | 4 |

| B | C2 | P |
|---|---|---|
| low | 1 | low |
| low | 2 | medium |
| low | 3 | high |
| medium | 1 | medium |
| medium | 2 | medium |
| high | 1 | high |
| high | 3 | high |

| R | T | C2 |
|---|---|---|
| low | low | 1 |
| low | high | 2 |
| high | low | 3 |
| high | high | 1 |

| R | C3 | P |
|---|---|---|
| low | 1 | medium |
| low | 2 | high |
| low | 3 | medium |
| high | 1 | low |
| high | 2 | medium |
| high | 3 | medium |
| high | 4 | high |

| B | T | C3 |
|---|---|---|
| low | low | 1 |
| low | high | 2 |
| medium | low | 3 |
| medium | high | 3 |
| high | low | 4 |
| high | high | 2 |

Fig. 6. The three possible decompositions of the function defined with examples from Table 5.

tion decomposition method described by Zupan et al. (1998). The variables in the spiders data set are continuous, but the method works only with categorical data, so each variable's values were categorized (discretized) into three categories using equal frequency discretization. The resulting hierarchy is shown in Figs. 7–9.

A comparison with the model built by the experts (Kampichler et al., 2000) can be made only using the four variables they used in their model ('disturbance',

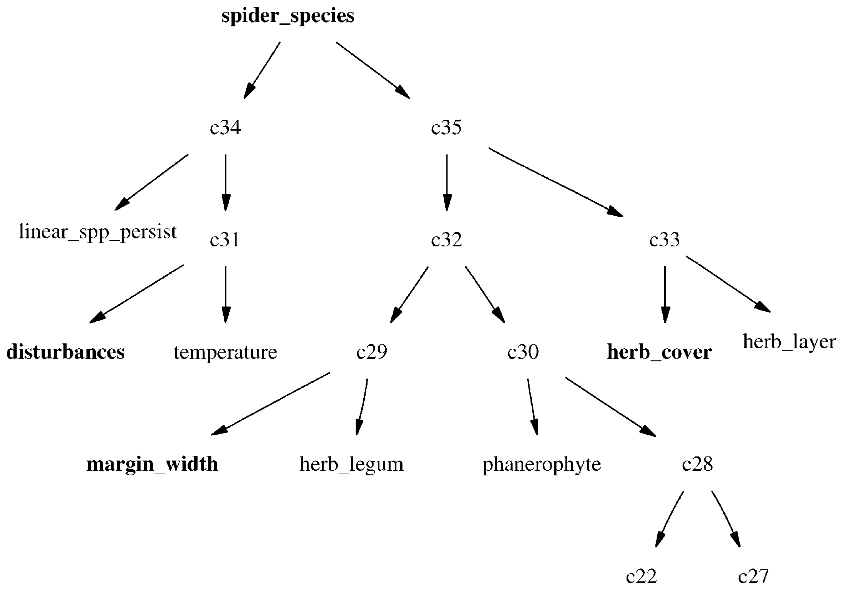


Fig. 7. The top part of the hierarchy obtained with HINT on the preprocessed spiders data set. The nodes marked with 'c' and a number, are artificial concepts created by the function decomposition method.

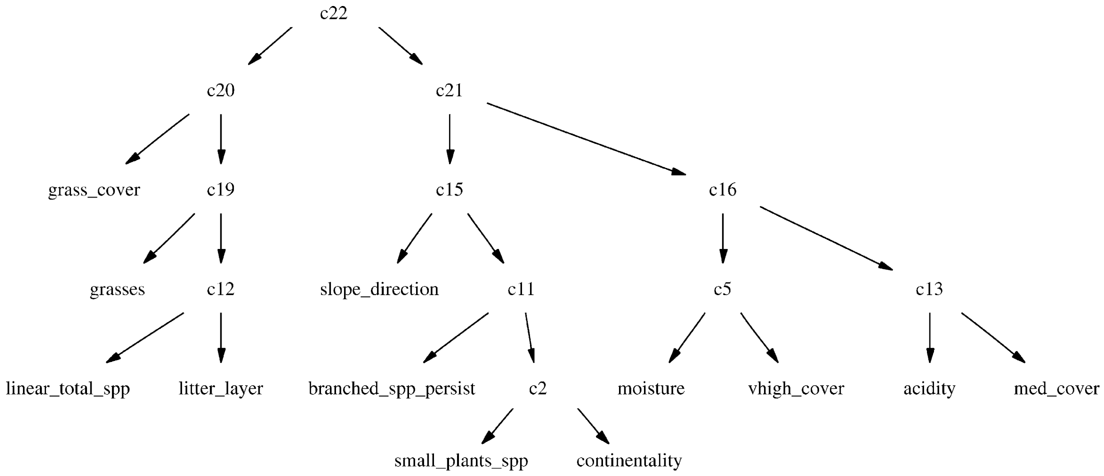


Fig. 8. The left part of the hierarchy obtained with HINT on preprocessed spiders data set. The nodes marked with 'c' and a number, are artificial concepts created by the function decomposition method.

'margin width', 'margin density' and 'herb cover'). The relations of these four variables in the hierarchy obtained with HINT do not match the relations in the experts' model. The methodologies employed are very different, so we cannot compare the fuzzy model, the interaction dendrogram and the hierarchy from HINT. Therefore, we cannot explain why there are differences in their structure.

Model consisting of so many attributes also proved to be too complex to be evaluated or explained by the experts. The interpretation of constructed concepts can not be made without a multitude of speculations that are making it worthless.

### 4.2.2. Model comparison

Because the structure obtained from the whole data set did not match the structure presented by Kampichler et al. (2000) and was too complex for the experts to evaluate, we also applied HINT to the spiders data set with only four variables present, the ones that were

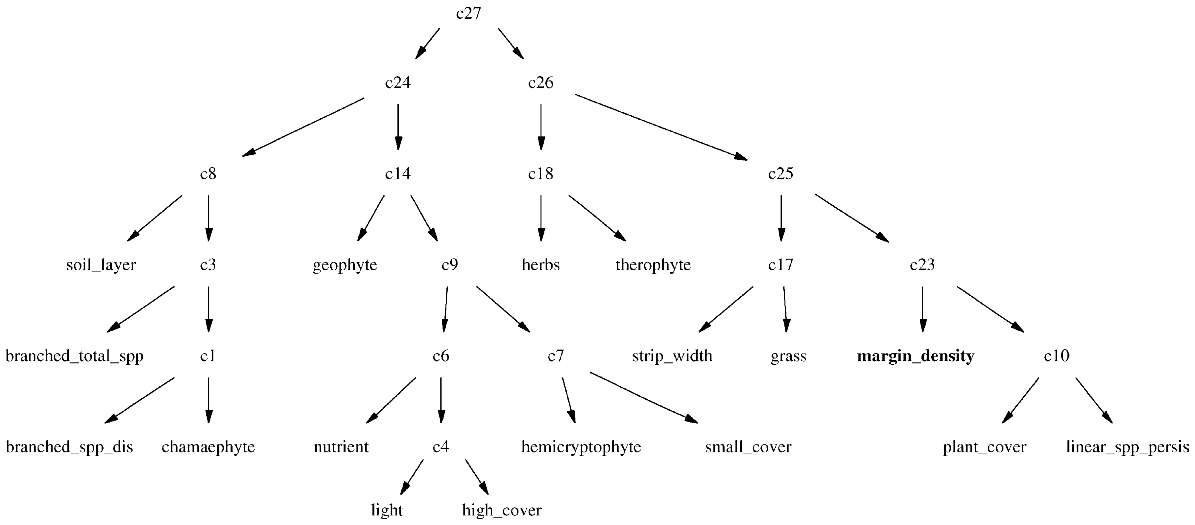


Fig. 9. The right part of the hierarchy obtained with HINT on preprocessed spiders data set. The nodes marked with 'c' and a number, are artificial concepts created by the function decomposition method.

used in experts' model (Kampichler et al., 2000). The discretization intervals of the values of these variables were identical to the ones used in the expert's model. The only difference was that the interval membership functions of values were not fuzzy, but crisp. In such a setting, besides the hierarchy, the functions in the model and the predictive performance of the model are interesting to observe and compare.

The hierarchy obtained in this setting is shown in Fig. 10. The four presented variables are in a different relation to the one in the hierarchy obtained from the data set with all the variables present. This can be due to the changes in data set, different categorization than in the previous experiment and a much smaller number of variables. The functions for the Species Number and both of the artificial concepts are given in Table 6. Some combinations of variable values never occur in the rules, because these combinations never occur in the data set.

There is no straightforward interpretation of the artificial concepts in the hierarchy available. However, we can recognize that concept 1 includes information on spatial and temporal scales of margin existence: margin density characterizes margin area available per unit area of landscape, thus higher density means higher margin accessibility for colonizers; less disturbance means larger life-time of a given margin. High margin accessibility and large margin persistence should promote accumulation of species and increase species number. Concept 2 includes only local habitat information: local size of margins and local supply with structural complexity (herbs supply more structure than grasses). We could roughly translate these observations to 'metacommunity concept' and 'habitat quality concept'. These concepts parallel recent attempts in explaining spatial distributions of animals by a hierarchical framework (Mackey and Lindenmayer, 2001), sometimes called 'filters', that must be passed by species from the regional pool to have access to local communities (Poff, 1997; Schröder and Reineking, 2004). For example, the artificial 'habitat quality' and 'metacommunity' concepts correspond to filter I (local resources and conditions) and filter II (spatial relationships—isolation, area size, dispersal capacity) in the filter cascade proposed by Schröder and Reineking (2004).

In the paper by Kampichler et al. (2000), the predictive performance of the models was measured with

Table 6
Table functions for Species Number, concept1 and concept2. All other names represent labels for artificial concepts' values

| MarDen | Distur | c1 |
|---|---|---|
| medium | low | ml |
| medium | high | mh |
| low | low | ll |
| low | high | hhlh |
| high | low | hl |
| high | high | hhlh |

| MarWid | HerCov | c2 |
|---|---|---|
| low | low | ll |
| low | medium | lm |
| low | high | mhlh |
| medium | medium | mm |
| high | medium | hm |
| medium | low | ml |
| high | high | hh |
| medium | high | mhlh |
| high | low | hl |

| c1 | c2 | SpecNum |
|---|---|---|
| ml | ll | medium |
| ml | lm | fairly high |
| ml | mm | fairly high |
| ml | hm | fairly high |
| ml | ml | medium |
| ml | hh | fairly high |
| ml | mhlh | medium |
| ml | hl | medium |
| mh | ll | low |
| mh | lm | low |
| mh | mm | fairly high |
| mh | ml | low |
| mh | hh | fairly high |
| mh | mhlh | fairly high |
| mh | hl | medium |
| ll | ll | low |
| ll | lm | low |
| ll | mm | high |
| ll | hm | medium |
| ll | hh | medium |
| ll | hl | fairly high |
| hhlh | ll | fairly high |
| hhlh | lm | very low |
| hhlh | mm | low |
| hhlh | ml | low |
| hhlh | hl | low |
| hl | lm | high |
| hl | hm | high |
| hl | mhlh | very high |

SpecNum = Species Number, $c1$ = concept1, $c2$ = concept2, MarDen = Margin Density, Distur = Disturbance, MarWid = Margin Width, HerCov = Herb Cover.

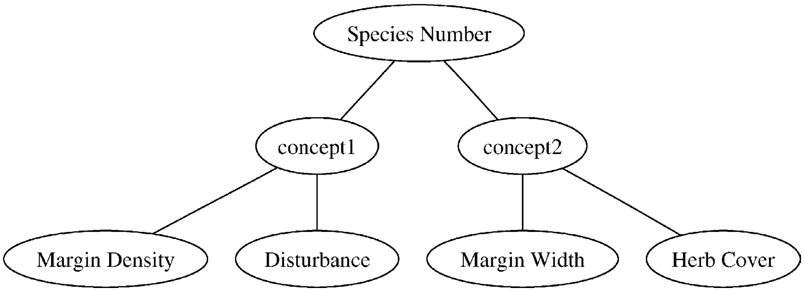


Fig. 10. The hierarchy obtained with HINT on spiders data set of four variables.

mean absolute error (MAE). This measure was evaluated on test data, which was not used for the tuning of fuzzy sets in the model. For a model based on multiple regression, the reported MAE was 3.17 species and 1.38 species for the fuzzy rule-based model.

We tested our crisp rule-based model on the discretized spiders data. The same data split into train and test set was used as in Kampichler et al. (2000) (87 study plots, 9 test plots) and the results are an average of 100 test runs. At each run, HINT constructed a model from the training data and the model was tested on the test data set. Because the goal variable is categorical in this data set, the result had to be 'decategorized', similarly as the result was 'defuzzified' when the fuzzy model was used (Kampichler et al., 2000). The procedure of 'decategorization' is very simple, each interval is represented by its mean value. As this approach is very rough, the results are expected to be somewhat worse. They also cannot be directly compared to the results in the paper by Kampichler et al., since the model was not fuzzy.

HINT reached a MAE of 2.56, a good result that indicates that our model certainly is valid and is not trivial. However, it would be interesting to know whether the fuzzy model is so much better because of the different structure and functions, or because of the fuzzy approach. To answer this question, a crisp model was built, based on the structure and functions (rules) from the original fuzzy model from the paper by Kampichler et al. (2000). This model was compared to the HINT's model made from all the data items (study + test), so neither model was tuned or fuzzy and the results could be directly compared. The resulting MAE of HINT's model (hierarchy in Fig. 10, functions in Table 6) was 2.49. The result of the crisp expert's model was 3.28, worse than the result of the crisp model with hierarchy and functions from HINT.

To further improve the automatically constructed model, we employed another machine learning technique. The HINT's model from the whole data set was revised with a data-driven model revision technique that tries to make changes in the model that would solve the model's misclassifications of data from a given data set. The method tries to make smallest possible changes and is limited only to changes in the rules of functions in hierarchy. A description of the data-driven revision is described in more detail in the literature (Žnidaršič and Bohanec, 2004). The method proposed a single change in the original model, the change of rule in the function of SpecNum from $< ll, ll, \text{low} >$ to $< ll, ll, \text{medium} >$. This change indeed is beneficial, but the effect is insignificant, the resulting MAE of the revised model is 2.47. This is currently the limit we reached on this data set with completely automatic methods of rule-based model construction.

It seems that the fuzzy approach is the main advantage of the expert's fuzzy model. Also, fine tuning of the shapes of fuzzy sets had a major impact on the result of fuzzy expert's model (Kampichler et al., 2000). The authors claim that an overfitted fuzzy model even had a MAE of over 5. The crisp version of the model does not have the advantage of tuning, that is a reason why the results of the crisp models are somewhat worse. This also puts the result of HINT's model into a different perspective. Its hierarchy and functions might be even more appropriate than the ones obtained by the experts, since the crisp HINT's model performed better than the crisp experts' model. The experimental performance measurements in terms of MAE are summarized in Table 7.

Table 7
Spider species number prediction MAE of the models in experiments

| Model | MAE |
|---|---|
| Multiple regression | 3.17 |
| Experts' fuzzy | 1.38 |
| HINT's | 2.56 |
| HINT's final | 2.49 |
| HINT's final + revision | 2.47 |
| Experts' final crisp | 3.28 |

The last three results correspond to the final models, they were learned and tested on all available examples.

## 5. Discussion

Two machine learning methods that discover interconnections and hierarchical relations were presented in the paper, interaction analysis and constructive induction. Both were applied to a data set from an ecological domain, the data set that describes features of field margins that could influence the diversity of spider species. The interaction analysis provides many interesting hypotheses and patterns, whereas the constructive induction is able to build a complete rule-based model of the target concept, ready to be used for prediction.

We have demonstrated several aspects of interaction analysis: the analysis of associations between variables, and the analysis of the interactions between independent variables for the purpose of predicting the outcome. In interaction analysis, the deviation from independence quantifies, as assessed through information-theoretic measures, both the proximity of variables and the existence of patterns. Interaction dendrograms summarize the variable proximity matrices, and interaction graphs pinpoint the most distinct interactions which can be examined through low-dimensional nonlinear regression models or scatter plots.

The variable structure obtained through the interaction dendrogram is very similar to the one designed by the experts. All the variables chosen by the experts are highly informative in the dendrogram, and dominate their respective sub-clusters. Namely, a common feature selection heuristic is to pick the one best predictor to represent the whole group. The difference is that interaction analysis suggests first merging margin density and disturbances, rather than margin width and disturbances. The second difference is that two other attributes could be considered for inclusion: slope direction and proportion of herbs in the total biomass.

Interaction analysis is a useful tool for exploratory data analysis. Its primary purpose is to help structure the variables in the domain and to guide the examination of interactions. However, care must be taken when acquiring probabilistic models from data. It is very easy to overfit probabilistic models when the data is scarce, and currently the automated tools for robust estimation are either inefficient (posterior sampling with MCMC), or may induce bias (Bayesian priors). This is one of the areas of our future work.

Using constructive induction with HINT, we have constructed two complete hierarchical rule-based models for prediction of the number of spider species. The hierarchies proposed by HINT were not similar to the ones in the model made by the experts, but domain experts were able to find a possible explanation for the smaller one. However, the large hierarchy of all the attributes proved to be too big and complex to be properly interpreted.

The model based on the variables that were used in the model in the referenced literature was evaluated for predictive power and compared to the models in the literature. Its performance was slightly worse, however, the other models had the advantage of using continuous values or pre-tuned fuzzy sets, so a direct comparison is not fair.

To examine only the hierarchies and crisp rule-based functions of the models, we adapted the model of the experts and rerun the tests. In this setting a direct comparison of results could be made. The HINT's hierarchies and functions gained a better result in this experiment, indicating that they might be even more appropriate than the ones in the model of the experts. This is another confirmation of the claim by Kampichler et al. (2000), that fuzzy models have a good predictive power. But it also indicates, that HINT can be a useful tool for environmental modelling, since the hierarchical rule-based models it constructs from data, are valid and even have a better predictive power then comparable crisp models with manually made hierarchy and rules.

There are some interesting areas for further work that appeared during our research. The most appealing, according to our results, would be to develop methods for fuzzy or probabilistic rule learning for the given structure of variables. It would also be interesting to see the results of using some variable selection methods prior to constructive induction. This way we might eliminate some noise and improve performance, while at the same time obtain a simpler full hierarchy that would probably be easier to interpret.

## Acknowledgements

We dedicate this paper to late Jutta Barthel who had collected the data base on spiders and field margin properties during her Ph.D. thesis with the FAM Munich Research Network on Agroecosystems (Forschungsverbund Agrarökosysteme München). We would like to thank her former collaborator, Gabriele Anderlik-Wesinger, for providing valuable explanations about Jutta's original data-sheets.

## References

Anderlik-Wesinger, G., Barthel, J., Pfadenhauer, J., Plachter, H., 1996. Einflußstruktureller und floristischer Ausprägungen von Rainen in der Agrarlandschaft auf Spinnen (Araneae) der Krautschicht. Verh. Ges. Ökol. 26, 711–720.

Baines, M., Hambler, C., Johnson, P.J., Macdonald, D.W., Smith, H., 1998. The effects of abale field management on the abundance and species richness of Araneae (spiders). Ecography 21, 74–86.

Barthel, J., Placher, H., 1996. Significance of field margins for foliage-dwelling spiders (Arachnida, Araneae) in an agricultural landscape of Germany. Rev. Suisse Zool. 45–59, vol. hors serie.

Chang, C.-C., Lin, C.-J., 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Demšar, J., Zupan, B., 2004. Orange: From Experimental Machine Learning to Interactive Data Mining. White Paper. Faculty of Computer and Information Science, University of Ljubljana, Slovenia. http://www.ailab.si/orange.

Ellenberg, H., Weber, H.E., Düll, R., Wirth, V., Werner, W., Paulißen, D., 1992. Zeigerwerte von Pflanzen in Mitteleuropa. Scripta Geobot. 18, 1–258.

Fayyad, U.M., Irani, K.B., 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In: Bajcsy, R. (Ed.), Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI 1993), pp. 1022–1029.

Gibson, C.W., Hambler, C., Brown, V.K., 1992. Changes in spider (Araneae) assemblages in relation to succession and grazing management. J. Appl. Ecol. 29, 132–142.

Jakulin, A., Bratko, I., 2003. Analyzing attribute dependencies. In: Lavrač, N., Gamberger, D., Blockeel, H., Todorovski, L. (Eds.), Proceedings of the Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2003), pp. 229–240.

Jakulin, A., Bratko, I., 2004. Testing the significance of attribute interactions. In: Greiner, R., Schuurmans, D. (Eds.), Proceedings of the 21st International Conference on Machine Learning (ICML-2004), pp. 409–416.

Kampichler, C., Barthel, J., Wieland, R., 2000. Species density of foliage-dwelling spiders in field margins: a simple fuzzy rule-based model. Ecol. Model. 129, 87–99.

Mackey, B.G., Lindenmayer, D.B., 2001. Towards a hierarchical framework for modelling the spatial distributions of animals. J. Biogeogr. 28, 1147–1166.

McGill, W.J., 1954. Multivariate information transmission. Psychometrika 19 (2), 97–116.

Poff, N.L., 1997. Landscape filters and species traits: towards mechanistic understanding and prediction in stream ecology. J. North Am. Benthol. Soc. 16, 391–409.

Rajski, C., 1961. A metric space of discrete probability distributions. Informat. Contr. 4, 373–377.

Schröder, B., Reineking, B., 2004. Modellierung der Art-Habitat-Beziehung—ein Überblick über die Verfahren der Habitatmodellierung. In: Dormann, C.F., Blaschke, T., Lausch, A., Schröder, B., Söndgerath, D. (Eds.), Habitatmodelle—Methodik, Anwendung, Nutzen, UFZ-Berichte, 9/2004, pp. 5–26.

Struyf, A., Hubert, M., Rousseeuw, P.J., 1997. Integrating robust clustering techniques in S-PLUS. Comput. Statist. Data Anal. 26, 17–37.

Thomas, M.B., Wratten, S.D., Sotherton, N.W., 1992. Creation of 'island' habitats in farmland to manipulate populations of beneficial arthropods: predator densities and species composition. J. Appl. Ecol. 29, 524–531.

Zupan, B., Bohanec, M., Demšar, J., Bratko, I., 1998. Feature transformation by function decomposition. IEEE Intell. Syst. Appl. 13, 38–43.

Zupan, B., Bohanec, M., Demšar, J., Bratko, I., 1999. Learning by discovering concept hierarchies. Artif. Intell. 109, 211–242.

Žnidaršič, M., Bohanec, M., 2004. Revision of qualitative multi-attribute decision models. In: Meredith, R., Shanks, G., Arnott, D., Carlsson, S. (Eds.), Proceedings of the 2004 IFIP International Conference on Decision Support Systems (DSS2004), pp. 881–888.