Machine Learning Applications in Habitat Suitability Modeling

19

Sašo Džeroski

19.1 Introduction

Environmental sciences comprise the scientific disciplines, or parts of them, that consider the physical, chemical and biological aspects of the environment (Allaby 1996). Environmental sciences are possibly the largest grouping of sciences, drawing heavily on life sciences and earth sciences, both of which are relatively large groupings themselves. Life sciences deal with living organisms and include (among others) agriculture, biology, biophysics, biochemistry, cell biology, genetics, medicine, taxonomy and zoology. Earth sciences deal with the physical and chemical aspects of the solid Earth, its waters and the air that envelops it. Included are the geologic, hydrologic, and atmospheric sciences. The latter are concerned with the structure and dynamics of Earth's atmosphere and include meteorology and climatology.

The field of environmental science is very interdisciplinary. It exists most obviously as a body of knowledge on its own right when a team of specialists assembles to address a particular issue (Allaby 1996). For instance, a comprehensive study of a particular stretch of a river would involve determining the geological composition of the riverbed (geology), determining the chemical and physical properties of the water (chemistry, physics), as well as sampling and recording the species living in and near the water (biology). Environmental sciences are highly relevant to environmental management, which is concerned with directing human activities that affect the environment.

Sašo Džeroski (✉)
Jozef Stefan Institute, Department of Knowledge Technologies, Jamova 39, 1000 Ljubljana, Slovenia
email: Saso.Dzeroski@ijs.si

The most typical representative of environmental sciences is ecology, which studies the relationships among members of living communities and between those communities and their abiotic (non-living) environment. Ecology is frequently defined as the study of the distribution and abundance of plants and animals (e.g., Krebs 1972). The distribution can be considered along the spatial dimension(s) and/or the temporal dimension.

Within ecology, the topic of ecological modeling (Joergensen and Bendoricchio 2001) is rapidly gaining importance and attention. Ecological modeling is concerned with the development of models of the relationships among members of living communities and between those communities and their abiotic environment. These models can then be used to better understand the domain at hand or to predict the behavior of the studied communities and thus support decision making for environmental management. Typical modeling topics are population dynamics of several interacting species and habitat suitability for a given species (or higher taxonomic unit).

Machine learning is one of the essential and most active research areas in the field of artificial intelligence. In short, it studies computer programs that automatically improve with experience (Mitchell 1997). The most researched type of machine learning is inductive machine learning, where the experience is given in the form of learning examples. Supervised inductive machine learning, sometimes also called predictive modeling, assumes that each learning example includes some target property, and the goal is to learn a model that accurately predicts this property.

Machine learning (and in particular predictive modeling) is increasingly often used to automate the construction of ecological models (Džeroski 2001). Most

S. E. Haupt et al. (eds.), *Artificial Intelligence Methods in the Environmental Sciences*

frequently, models of habitat suitability and population dynamics are constructed from measured data by using machine learning techniques. The most popular machine learning techniques used for modeling habitat suitability include decision tree induction (Breiman et al. 1984, see also Chapter 4 of this volume by Dattatreya), rule induction (Clark and Boswell 1991), and neural networks (Lek and Guegan 1999, see also Chapter 2 of this volume by Marzban).

In this chapter, we will focus on applications of machine learning in ecological modeling, more specifically, applications in habitat suitability modeling. Habitat-suitability modeling studies the effect of the abiotic characteristics of the habitat on the presence, abundance or diversity of a given taxonomic group of organisms. For example, one might study the influence of soil characteristics, such as soil temperature, water content, and proportion of mineral soil on the abundance and species richness of springtails, the most abundant insects in soil. To build habitat-suitability models, machine learning techniques can be applied to measured data on the characteristics of the environment and the abundance of the taxonomic group(s) studied.

In the remainder of this chapter, we first discuss in more detail the task of habitat suitability modeling. We next briefly describe two approaches to machine learning that are often used in habitat suitability modeling: decision tree induction and rule induction. We then give examples of using machine learning to construct models of habitat suitability for several kinds of organisms. These include habitat models for bioindicator organisms in a river environment, springtails and other soil organisms in an agricultural setting, brown bears in a forest environment, and finally habitat suitability models for sea cucumbers in a sustainable fishing setting.

19.2 Habitat Suitability Modeling

If ecology is defined as the study of the distribution and abundance of plants and animals, habitat suitability modeling is concerned with the spatial aspects of the distribution and abundance. Habitat suitability models relate the spatially varying characteristics of the environment to the presence, abundance or diversity of a given (taxonomic) group of organisms. For example, one might study the influence of soil characteristics, such as soil temperature, water content, and proportion of mineral soil on the abundance and species richness of springtails, the most abundant insects in soil.

The input to a habitat model is thus a set of environmental characteristics for a given spatial unit of analysis. The output is a target property of the given (taxonomic) group of organisms. Note that the size of the spatial unit, as well as the type of environmental variables, can vary considerably, depending on the context, and so can the target property of the population (even though to a lesser extent).

The spatial unit considered may be of different size for different habitat models. For example, in the study of Collembola habitat, the soil samples taken were of size 7.8 cm diameter and 5 cm depth (Kampichler et al. 2000), in the study of sea cucumber habitat transects of 2 by 50 m of the sea bed were considered (Džeroski and Drumm 2003), and in ongoing studies of potential habitats for different tree species under varying climate change scenarios, 1 by 1 km squares are considered (Ogris and Jurc 2007). Habitat models can thus operate at very different spatial scales.

The input to a habitat model is a set of environmental variables, which may be of three different kinds. The first kind concerns abiotic properties of the environment, e.g., physical and chemical characteristic thereof. The second kind concerns some biological aspects of the environment, which may be considered as en external impact on the group of organisms under study. Finally, the third kind of variables are related to human activities and their impacts on the environment.

The environmental variables that describe the abiotic part of the environment can be of different nature, depending for example on whether we study a terrestrial or an aquatic group of organisms. Typical groups of variables concern properties of the terrain (calculated from a digital elevation model), such as elevation, slope and exposition; geological composition of the terrain or the riverbed/seabed; physical and chemical properties of the soil/water/air, such as moisture, pH, quantities of pollutants, and so on. An important group of variables concerns climate and encompasses temperature, precipitation, etc.

Biological aspects of the environment that are considered in habitat models are typically more specific

and more directly related to the target group of organisms as compared to the abiotic variables. They may be rather coarse and refer to the community, e.g., when modeling brown bear habitat one of the inputs may be the type of forest at a particular location. They may also refer to more specific types of organisms that are related to the target group, e.g., when modeling the habitat of wolves, information on important prey species such as hare and deer may be taken into account.

Some environmental variables may involve both abiotic and biotic aspects. Land cover is a typical example: possible values for this variable may be forest, grassland, water, etc. Finally, some environmental variables are related to human activity: examples are proximity to settlements, population density, and proximity to roads/railways.

The output of a habitat model is some property of the population of the target group of organisms at the spatial unit of analysis. There are two degrees of freedom here: one stems from the target property, the other from the group of organisms studied. In the simplest case, the output is just the presence/absence of a single species (or group). In this case, we simply talk about habitat models.

An example habitat model for brown bears in Slovenia (taken from Jerina et al. 2003) is given in Table 19.1. It has the form of an IF-THEN rule, which specifies the conditions that define suitable habitat for brown bears. The rule uses three environmental variables PREDOMINANT-LAND-COVER, FOREST-ABUNDANCE and PROXIMITY-TO-SETTLEMENTS: it was actually learned by applying machine learning techniques to observational data.

We can also be interested in the abundance or density of the population. If we take these as indicators of the suitability of the environment for the group of organisms studied, we talk about habitat suitability models: the output of these models can be interpreted as a degree of suitability. The abundance of the population can be measured in terms of the number of individuals or their total size (e.g., the dry biomass of a certain species of algae). If the (taxonomic) group is large enough, we can also consider the diversity of the group (Shannon index, species richness or such like, see Krebs 1989).

Table 19.1 A habitat model for the brown bear (*Ursus arctos*) in Slovenia

IF	PREDOMINANT-LAND-COVER = Forest
AND	FOREST-ABUNDANCE > 60%
AND	PROXIMITY-TO-SETTLEMENTS > 1.5 km
THEN	BrownBearHabitat = Suitable
ELSE	BrownBearHabitat = Unsuitable

In the most general case of habitat modeling, we are interested in the relation between the environmental variables and the structure of the population at the spatial unit of analysis (absolute and relative abundances of the organisms in the group studied). One approach to this is to build habitat models for each of the organisms (or lower taxonomic units) in the group, then aggregate the outputs of these models to determine the structure of the population (or the desired target property). An alternative approach is to build a model that simultaneously predicts the presence/abundance of all organisms in the group or directly the desired target property of the entire group. A comparison of the two approaches in the context of machine learning of habitat models is given by Demšar et al. (2006a).

We should note here that observing the presence or absence of a species/group (or its abundance/density) within a given spatial unit can be a nontrivial task. While most plants and certain animals (such as sea cucumbers) are relatively immobile, many animals (including brown bears) can move fast and cover wide spatial areas. In the latter cases, one might consider areals of activity (home ranges) and sample from these to obtain data for learning habitat suitability models: this is what was done in the study by Jerina et al. (2003).

Another issue that commonly occurs in habitat modeling, especially in the context of machine learning, is the fact that only presence data are often collected (i.e., no absence data are usually available). In such cases, additional care is necessary when preparing the data for the modeling task. Examples (spatial units) where the target group can be reasonably expected not to occur (based on domain knowledge) may be considered as absence data.

Finally, let us reiterate that habitat modeling focuses on the spatial aspects of the distribution and abundance of plants and animals. It studies the relationships between some environmental variables and the presence/abundance of plants and animals, under

the implicit assumption that both are observed at a single point in time for a given spatial unit. It mostly ignores the temporal aspects of the distribution/abundance, the latter being the focus of population dynamics modeling. Still, some temporal aspects may be taken into account, for example, averages of environmental variables over a period of time preceding the observation are sometimes included in habitat models (e.g., average winter air temperature).

19.3 Machine Learning for Habitat Modeling

19.3.1 The Machine Learning Task of Predictive Modeling

The input to a machine learning algorithm is most commonly a single flat table comprising a number of fields (columns) and records (rows). In general, each row represents an object and each column represents a property (of the object). In machine learning terminology, rows are called examples and columns are called attributes (or sometimes features). Attributes that have numeric (real) values are called continuous attributes. Attributes that have nominal values (are called discrete attributes.

The tasks of classification and regression are the two most commonly addressed tasks in machine learning. They are concerned with predicting the value of one field from the values of other fields. The target field is called the class (dependent variable in statistical terminology). The other fields are called attributes (independent variables in statistical terminology).

If the class is continuous, the task at hand is called regression. If the class is discrete (it has a finite set of nominal values), the task at hand is called classification. In both cases, a set of data (dataset) is taken as input, and a predictive model is generated. This model can then be used to predict values of the class for new data. The common term predictive modeling refers to both classification and regression.

Given a set of data (a table), only a part of it is typically used to generate (induce, learn) a predictive model. This part is referred to as the training set. The remaining part is reserved for evaluating the predictive performance of the learned model and is called the testing set. The testing set is used to estimate the performance of the model on unseen data (and sometimes also called validation set, see Chapter 2 of this volume by Marzbahn).

More reliable estimates of performance on unseen data are obtained by using cross-validation, which partitions the entire data available into N (with N typically set to 10) subsets of roughly equal size. Each of these subsets is in turn used as a testing set, with all of the remaining data used as a training set. The performance figures for each of the testing sets are averaged to obtain an overall estimate of the performance on unseen data.

19.3.2 A Machine Learning Formulation of Habitat Modeling

In the case of habitat modeling, examples correspond to spatial units of analysis. The attributes correspond to environmental variables describing the spatial units, as these are the inputs to a habitat model. The class is a target property of the given (taxonomic) group of organisms, such as presence, abundance or diversity.

The habitat model from Table 19.1 has been learned from a dataset which includes the discrete attribute PREDOMINANT-LAND-COVER (which can have the value forest, among others) and the continuous attributes FOREST-ABUNDANCE and PROXIMITY-TO-SETTLEMENTS. The class BrownBear-Habitat is discrete, with Suitable and Unsuitable as possible values. Hence, we are dealing with a classification task. An excerpt from the dataset is given in Table 19.2.

The machine learning task of habitat modeling is thus defined as follows. Given is a set of data with rows corresponding to spatial locations (units of analysis), attributes corresponding to environmental variables, and the class corresponding to a target property of the population studied. The goal is to learn a predictive model that predicts the target property from the environmental variables (from the given dataset). If we are only looking at presence/absence or suitable/unsuitable as values of the class (as is the case above), we have a classification problem. If we are looking at the degree of suitability (density/abundance), we have a regression problem.

Table 19.2 An excerpt from the dataset for modeling brown bear habitat in Slovenia. PLC stands for PREDOMINANT-LAND-COVER, PTS for PROXIMITY-TO-SETTLEMENTS, and BBH for BrownBearHabitat

Location	PLC	FOREST-ABUNDANCE	PTS	OtherEnvVariables	BBH
l1	Forest	80	21.4	–	Yes
l2	Forest	66	13.9	–	Yes
l3	Forest	55	50.0	–	No
l4	Forest	72	1.2	–	No
l5	Grassland	6	19.1	–	No
l6	Grassland	0	11.4	–	No
l7	Wetland	3	5.8	–	No
l8	Water	0	3.9	–	No

19.3.3 Decision Tree Induction

19.3.3.1 What Are Decision Trees?

Decision trees (Breiman et al. 1984, see also Chapter 4 of this volume by Dattatreya) are hierarchical structures, where each internal node contains a test on an attribute, each branch corresponds to an outcome of the test, and each leaf node gives a prediction for the value of the class variable. Depending on whether we are dealing with a classification or a regression problem, the decision tree is called a classification or a regression tree, respectively. An example classification tree modeling the habitat of sea cucumbers is given in Fig. 19.1. The tree has been derived from actual data by using machine learning (Džeroski and Drumm 2003).

Regression tree leaves contain constant values as predictions for the class value. They thus represent piece-wise constant functions. Model trees, where leaf nodes can contain linear models predicting the class value, represent piece-wise linear functions. An example model tree that predicts the total abundance of hemi- and eu-edaphic Collembola is given in Fig. 19.2 (Kampichler et al. 2000).

Note that decision trees represent total partitions of the data space, where each test corresponds to an axis-parallel split. Most algorithms for decision tree induction consider axis-parallel splits. However, there are a few algorithms that consider splits along lines that need not be axis-parallel or even consider splits along non-linear curves.

19.3.3.2 Top-Down Induction of Decision Trees

Finding the smallest decision tree that would fit a given data set is known to be computationally expensive (NP-hard). Heuristic search, typically greedy, is thus employed to build decision trees. The common way to induce decision trees is the so-called Top-Down Induction of Decision Trees (TDIDT, Quinlan 1986). Tree construction proceeds recursively starting with the entire set of training examples (entire

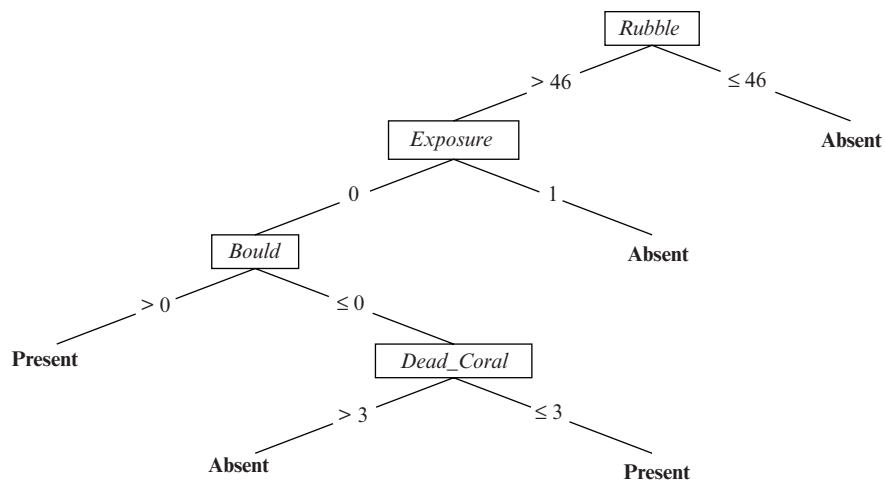

Fig. 19.1 A classification tree that predicts the suitability of habitat for the sea cucumber species *Holothuria Leucospilota* on Rarotonga, Cook Islands

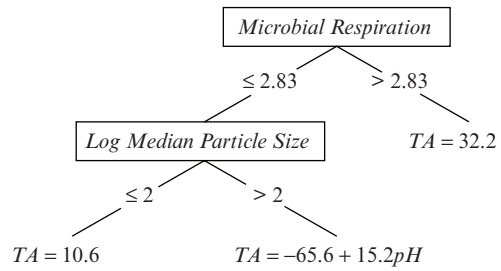

Fig. 19.2 A model tree that predicts the total abundance (TA) of hemi- and eu-edaphic Collembola on the FAM experimental farm at Scheyern (near Munich), Germany

table). At each step, an attribute is selected as the root of the (sub)tree and the current training set is split into subsets according to the values of the selected attribute.

For discrete attributes, a branch of the tree is typically created for each possible value of the attribute. For continuous attributes, a threshold is selected and two branches are created based on that threshold. For the subsets of training examples in each branch, the tree construction algorithm is called recursively. Tree construction stops when the examples in a node are sufficiently pure (i.e., all are of the same class) or if some other stopping criterion is satisfied (e.g., there is no good attribute to add at that point). Such nodes are called leaves and are labeled with the corresponding values of the class.

Different measures can be used to select an attribute in the attribute selection step. These also depend on whether we are inducing classification or regression trees (Breiman et al. 1984). For classification, Quinlan (1986) uses information gain, which is the expected reduction in entropy of the class value caused by knowing the value of the given attribute. Other attribute selection measures, however, such as the Gini index (Breiman et al. 1984) or the accuracy of the majority class, can and have been used in classification tree induction. In regression tree induction, the expected reduction in variance of the class value can be used.

An important mechanism used to prevent trees from over-fitting data is tree pruning. Pruning can be employed during tree construction (pre-pruning) or after the tree has been constructed (post-pruning). Typically, a minimum number of examples in branches can be prescribed for pre-pruning and a confidence level in accuracy estimates for leaves for post-pruning.

19.3.4 Rule Induction

19.3.4.1 What Are Predictive Rules?

We will use the word rule here to denote patterns of the form "IF Conjunction of conditions THEN Conclusion." The individual conditions in the conjunction will be tests concerning the values of individual attributes, such as "PROXIMITY-TO-SETTLEMENTS > 1.5 km" or "PREDOMINANT-LAND-COVER=Forest". For predictive rules, the conclusion gives a prediction for the value of the target (class) variable.

If we are dealing with a classification problem, the conclusion assigns one of the possible discrete values to the class, e.g., "BrownBearHabitat=Unsuitable". A rule applies to an example if the conjunction of conditions on the attributes is satisfied by the particular values of the attributes in the given example. Each rule corresponds to a hyper-rectangle in the data space.

Predictive rules can be ordered or unordered. Unordered rules are considered independently and several of them may apply to a new example that we need to classify. A conflict resolution mechanism is needed if two rules which recommend different classes apply to the same example. A default rule typically exists, whose recommendation is taken if no other rule applies.

Ordered rules form a so-called decision list. Rules in the list are considered from the top to the bottom of the list. The first rule that applies to a given example is used to predict its class value. Again, a default rule with an empty precondition is typically found as the last rule in the decision list and is applied to an example when no other rule applies.

An ordered list of rules describing brown bear habitat is given in Table 19.1: the second rule in this list is the default rule which always applies. An unordered list of rules that predicts the suitability of habitat for sea cucumbers is given in Table 19.3. Note that classification trees can be transcribed into sets of classification rules, since each of the leaves of a classification tree corresponds to a classification rule. Although less

Table 19.3 A set of unordered rules that predicts the suitability of habitat for the sea cucumber species *Holothuria Leucospilota* on Rarotonga, Cook Islands. The default rule, which predicts the class Absent is not listed

IF Sand < 7.5 AND Rubble > 62.0 AND Rock_Pave < 15.0 AND Dead_Coral < 13.5 THEN Presence = Present [3 absent, 15 present]
IF Rubble < 54.0 AND 7.5 < Consol_Rubble < 77.5 AND Bould < 25.0 AND Rock_Pave < 30.0 AND Dead_Coral < 45.0 THEN Presence = Present [1 absent, 6 present]
IF Rubble < 9.5 AND Live_Coral < 27.5 THEN Presence = Absent [65 absent]
IF Sand > 8.5 AND Consol_Rubble < 5.0 THEN Presence = Absent [64 absent]
IF Bould > 2.5 AND Rock_Pave > 30.0 THEN Presence = Absent [10 absent]

common in practice, regression rules also exist, and can be derived, e.g., by transcribing regression trees into rules.

19.3.4.2 The Covering Algorithm for Rule Induction

In the simplest case of binary classification, one of the classes is referred to as positive and the other as negative. For a classification problem with several class values, a set of rules is constructed for each class. When rules for class c_i are constructed, examples of this class are referred to as positive, and examples from all the other classes as negative.

The covering algorithm works as follows. We first construct a rule that correctly classifies some examples. We then remove the examples covered by the rule from the training set and repeat the process until no more examples remain. When learning ordered rules we remove all examples covered and when learning unordered rules only the positive examples covered by the rule.

Within this outer loop, different approaches can be taken to find individual rules. One approach is to heuristically search the space of possible rules top-down, i.e., from general to specific (in terms of examples covered this means from rules covering many to rules covering fewer examples) (Clark and Boswell 1991). To construct a single rule that classifies examples into class c_i, we start with a rule with an empty antecedent (IF part) and the selected class c_i as a consequent (THEN part). The antecedent of this rule is satisfied by all examples in the training set, and not only those of the selected class. We then progressively refine the antecedent by adding conditions to it, until only examples of class c_i satisfy the antecedent. To allow for handling imperfect data, we may construct a set of rules which is imprecise, i.e., does not classify all examples in the training set correctly.

19.4 Case Studies of Habitat Modeling with Machine Learning

In this section, we exemplify the machine learning approach to habitat modeling through four case studies. For each case study, we briefly describe the data available, the machine learning approach used, and the results obtained. We also give examples of habitat models learned in the process.

19.4.1 Bioindicator Organisms in Slovenian Rivers

In this study (Džeroski et al. 1997), we learned habitat models for 17 organisms that can be found in Slovenian rivers and are used as indicator organisms when determining the biological quality of river waters. The habitat models explicate the influence of physical and chemical parameters of river water on 10 plant taxa and seven animal taxa. On the plant side, eight kinds of diatoms (BACILLARIOPHYTA) and two kinds of green algae (CHLOROPHYTA) were studied. The animal taxa chosen for study include worms (OLIGOCHAETA), crustaceans (AMPHIPODA) and five kinds of insects.

The plant taxa studied were: *Coconeis placentula*, *Cymbella sp.*, *Cymbella ventricosa*, *Diatoma vulgare*, *Navicula cryptocephala*, *Navicula gracilis*, *Nitzschia palea*, *Synedra ulna*, *Cladophora sp.*, and *Oedogonium sp.* The animal taxa studied were *Tubifex sp.*,

Table 19.4 Example rules from the habitat models for bioindicator organisms in Slovenian rivers (*Nitzschia palea*, *Elmis sp.*, and *Plecoptera leuctra sp.*)

IF Hardness > 11.85 AND NO_2 > 0.095 AND NH_4 > 0.09 THEN Nitzschia = Present	IF NO_2 < 0.005 AND NO_3 < 7.1 AND PO_4 < 0.125 AND Detergents < 0.055 AND BOD < 2 THEN Nitzschia = Absent
IF Temperature > 12.75 AND BOD < 0.65 THEN Elmis = Present	IF PH > 7.05 AND BOD > 12.15 THEN Elmis = Absent
IF Temperature < 23 AND 120 < Saturation < 150 AND COD > 10.9 AND BOD < 3.75 THEN Leuctra = Present	IF Temperature < 22.25 AND Total Hardness < 18.55 AND BOD > 6.9 THEN Leuctra = Absent

Gammarus fossarum, *Baetis sp.*, *Leuctra sp.*, *Chironomidae (green)*, *Simulium sp.*, *Elmis sp.*

The data used in the study came from the Hydrometeorological Institute of Slovenia (now Environment Agency of Slovenia) that performs regular water quality monitoring for most Slovenian rivers and maintains a database of water quality samples. The data used cover a 4 year period, from 1990 to 1993. In total, 698 water samples were available on which both physical/chemical and biological analyses were performed: the former provided the environmental variables for the habitat models, while the latter provided information on the presence/absence of the studied organisms.

Plants are more or less influenced by the following physical and chemical parameters (water properties): total hardness, nitrogen compounds (NO_2 , NO_3, NH_4), phosphorus compounds (PO_4), silica (SiO_2), iron (Fe), surfactants (detergents), chemical oxygen demand (COD), and biochemical oxygen demand (BOD). The last two parameters indicate the degree of organic pollution: the first reflects the total amount of degradable organic matter, while the second reflects the amount of biologically degradable matter. Animals are mostly influenced by a different set of parameters: water temperature, acidity or alkalinity (pH), dissolved oxygen (O_2, saturation of O_2), total hardness, chemical (COD), and biochemical oxygen demand (BOD).

The habitat models for the plant/animal taxa used the following environmental variables: Hardness, NO_2, NO_3, NH_4, PO_4, SiO_2, Fe, Detergents, COD, BOD for plants and Temperature, PH, O_2, Saturation, COD, BOD for animals. The class is the presence of the selected taxon (with values Present and Absent). Seventeen machine learning problems were thus defined, one for each taxon. Each of the datasets contained 698 examples.

Rule induction, and in particular the CN2 system (Clark and Boswell 1991), was used to construct the habitat models. The rules induced on the complete data were given to a domain expert (river ecologist) for inspection. Their accuracy on unseen data was also estimated by dividing the data into a training set (70%) and a testing set (30%), repeating this 10 times and averaging the results (accuracy on the test set).

The accuracy of the 17 models on the whole (training) dataset ranges between 66% and 85%, while the default accuracy, i.e., the majority class frequency ranges from 50% to 70%. The estimated accuracy on unseen cases ranges from 53% to 71%. In nine of the 17 cases, the models substantially improve upon the default accuracy and provide interesting knowledge about the taxa studied.

In several cases, the induced rules are consistent with and confirm the expert knowledge about the organism studied. The diatom *Nitzschia palea*, the most common species in Slovenian rivers, is very tolerant to pollution. The rules confirm that a larger degree of pollution is beneficial to this species: they indicate that *Nitzschia palea* needs nitrogen compounds, phosphates, silica, and larger amounts of degradable matter (COD and BOD). *Elmis sp.* is known to inhabit clean waters: the rules demand a low quantity of biodegradable matter (pollution) in order for the taxon to be present, and predict that the taxon will be absent if the

water is overly polluted (has high values of BOD, COD and pH).

Not all of the induced rules agree with existing expert knowledge. For example, the rules that predict the presence of the taxon *Plecoptera leuctra sp.*, which is used as an indicator of clean waters, confirm that it is indeed found mainly in clean waters. However, they also state that it can be found in quite polluted water, provided there is enough oxygen. Thus, they enhance current knowledge on the bioindicator role of this taxon.

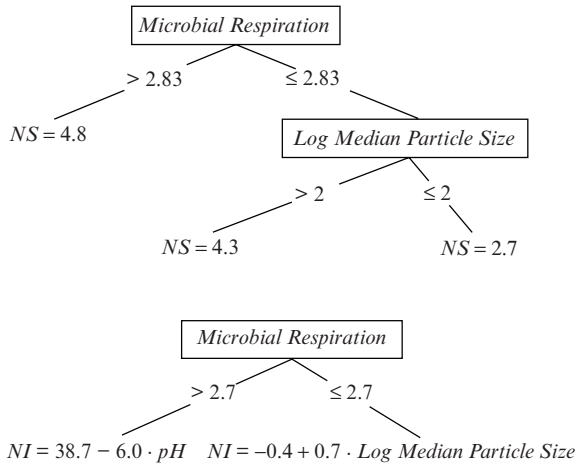

Fig. 19.3 A regression (top) and a model tree (bottom) that predict the number of species (NS) of hemi- and eu-edaphic Collembola and the number of individuals (NI) of the collembolan species *Folsomia quadrioculata*, respectively, on the FAM experimental farm at Scheyern (near Munich), Germany

19.4.2 Soil Insects on an Experimental Farm in Germany

Kampichler et al. (2000) used machine learning techniques to build habitat models for Collembola (springtails), the most abundant insects in soil, in an agricultural soil environment. They study both the taxonomic group of Collembola, as well as the dominant species in the study area, (*Folsomia quadrioculata*). The habitat models constructed relate the total abundance and species number of Collembola, as well as the abundance of the dominant species, to habitat characteristics, i.e., properties of the soil.

The data used in the study come from an experimental farm at Scheyern (near Munich), Germany, run by the FAM Research Network on Agroecosystems. The farm was of size approximately 153 ha, located at an elevation of 450–490 m above sea level, with mean annual temperature and mean annual precipitation of 7.58°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 diameter, 5 cm depth) and yielded a total of 396 cores. The majority of these points were situated in arable fields, the remainder in pastures, meadows and arable fields on former hop fields. Microarthropods were counted and Collembola identified by species. Only data of euedaphic (soil-dwelling) Collembola and hemiedaphic Collembola (which live near the soil surface) were included in the analysis.

To measure environmental factors, cores were taken from the same sampling points, at a distance of approximately 25 cm from the first cores. The following environmental variables were measured: microbial biomass, microbial respiration, soil moisture, soil acidity, carbon content (Ct) and nitrogen content (Nt). Soil texture at the sampling points was also determined and expressed by the (base 10) logarithm of the median particle size (diameter). 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 samples.

To build habitat models, we used regression trees. More specifically, the system M5 (Quinlan 1992) for model tree induction was used. Trees were built separately for each of the three target variables: the abundance and diversity (species number) of Collembola, and the abundance of the dominant species *Folsomia quadrioculata*. Example trees for the last two are given in Fig. 19.3, while an example tree for the first is given in Fig. 19.2. Linear regression models, as well as neural networks with one hidden layer, were also constructed for each of the target variables.

In terms of predictive power, model trees fared better than linear regression and worse than neural networks. All of them, however, had quite low predictive power (for unseen cases, the correlation coefficients were estimated by 10-fold cross-validation at approx. 0.3 for linear regression, 0.4 for model trees and 0.5 for neural networks). The most probable reason for the low performance is that the aggregated spatial

distribution of collembolans sets limit to the possibility of predicting the actual number of collembolans. In this context, the quality of trees of being transparent and providing explicit information about the quantitative relationships between the variables proved very appealing to the domain experts.

The trees clearly identify microbial respiration as the most important factor influencing the collembolan community, followed by soil texture and soil acidity. The same environmental variables seem to be important for all three target variables and the structure of the individual trees is very similar. In this case, simultaneous prediction of all target variables seems reasonable: this can be done by applying predictive clustering trees (Blockeel et al. 1998), a generalized version of decision trees. This methodology, also called multi-objective classification/prediction, has been applied to habitat modeling for river communities (Džeroski et al. 2001) and for soil insects, including mites and springtails (Demšar et al. 2006b).

19.4.3 Brown Bears in Slovenia

The brown bear (*Ursus arctos*) occurs today in only a small part of its historical range: Slovenia is among the few European countries with a preserved viable indigenous brown bear population, as well as populations of other large predator species, such as wolf and lynx. The Slovenian bear population is a part of the continuous Alps-Dinaric-Pindos population: its core habitat (the forests of Kočevska and Snežnik in South-Western Slovenia) is connected with Gorski Kotar in Croatia in a unified block of bear habitat. This bear population is important also because it represents the source for natural re-colonization or reintroduction of the bear into Slovenia's neighboring countries Austria and Italy.

In their study, Jerina et al. (2003) address three aspects of the brown bear population in Slovenia: its size (and its evolution over time), its spatial expansion out of the core area, and its potential habitat based on natural habitat suitability. The results of the study include estimates of population size, a picture of the spatial expansion of the population and maps of its optimal and maximal potential habitat (based on natural suitability). All of these are relevant to the management of the Slovenian brown bear population. In this section, we summarize the habitat modeling aspect of the study.

The habitat models built were based on bear sightings data acquired in the last decade of the 20th century by the Hunters association of Slovenia, as well as data from a previous radio-tracking project. Since we were interested in the optimal habitat, best represented by females with cubs, we selected only such sightings. Instead of using a cloud of sighting location points as the basis for the models, we used an estimation of the inhabited area (IA) constructed by a kernel method: this method gives as output the frequency/probability with which individual points in space are occupied by brown bears.

The spatial unit of analysis was a pixel of size 500×500 m. Positive examples were sampled from the inhabited area. Examples for the "optimal" habitat model were sampled from areas that exceeded a high threshold of the probability of bear occupancy: This threshold was lower when sampling positive examples for the "maximal" potential habitat model. Negative examples were randomly sampled from the rest of the study area (i.e., not the IA), which presumably is less (or not at all) suitable for bear habitat.

The explanatory environmental variables were derived from several GIS (Geographical Information Systems) layers. These included land cover data, forest inventory data, settlements map, road map, and a digital elevation model. Example variables include forest abundance and proximity to settlements. A value of each of these variables was associated with each 500×500 m pixel. The method of decision tree induction, and in particular the See5 commercial product, based on the C4.5 (Quinlan 1993) algorithm, was used to build the "optimal" and "maximal" habitat models.

The decision tree for optimal habitat (Fig. 19.4a) takes into account the surrounding forest matrix size, forest abundance in each pixel, predominant land cover type, sub-regional density of human population, and the predominant forest association within each forest pixel. The decision tree for maximal habitat (Fig. 19.4b) is much simpler and only takes into account the predominant land cover type, forest abundance, and proximity to settlement. Note that the classification rule for predicting "maximal bear habitat", given in Table 19.1, is obtained by rewriting the tree in Fig. 19.4b.

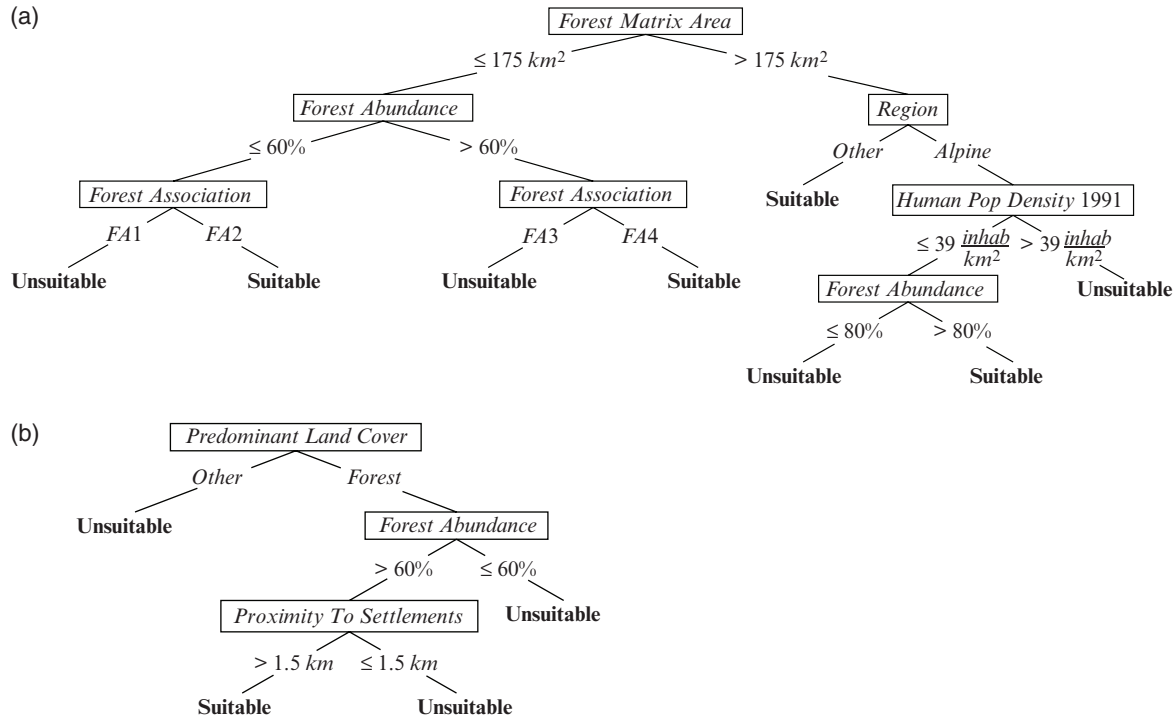

Fig. 19.4 Two decision trees predicting the (a) optimal and (b) maximal habitat of the brown bear (*Ursus arctos*) in Slovenia. FA1, FA2, FA3, and FA4 denote four different groups of forest associations, where FA1 and FA3 contain oak and FA2 and FA4 contain beech

The learned trees were used to produce the respective habitat maps. The thematic accuracy for the first map was estimated by 10-fold cross-validation as 89%, and 84% for the latter. The optimal habitat covers 12.3% of Slovenias territory, mostly in the southern part, bordering to Croatia. The possible maximal habitat extent includes additional 26.4% of the territory, mostly in the alpine region in the northern and western part of Slovenia, thus totaling 38.7% of the country.

It can be gleaned both from the decision trees as well as from the final habitat maps, that the bear habitat suitability in Slovenia largely depends on the presence of a dense forest cover, while it depends less upon food availability. Considering the increasing trend of forest cover in Slovenia, and assuming a continuation of high reproduction rates, we could even expect a further expansion of bear-inhabited areas in the future. It is furthermore obvious that the six-lane Ljubljana Trieste highway cuts through the optimal habitat at two vulnerable bottlenecks, disrupting the dispersion corridors towards the Alps: This can be seen from a large number of bear related traffic accidents on the highway. The habitat maps we constructed were used to recommend suitable locations for eco-ducts (wildlife bridges) across this highway to the Highway authority of Slovenia.

19.4.4 Sea Cucumbers on Rarotonga, Cook Islands

In the Pacific Islands, invertebrates including sea cucumbers are among the most valuable and vulnerable inshore fisheries resources. The sea cucumber (*Holothuria leucospilota*) forms an important part of the traditional subsistence fishery on Rarotonga, Cook Islands, yet little is known of this species present spatial distribution and abundance around the island. To contribute to the knowledge about this species, Džeroski and Drumm (2003) apply machine learning to measured data and build a habitat model that predicts the number of sea cucumber

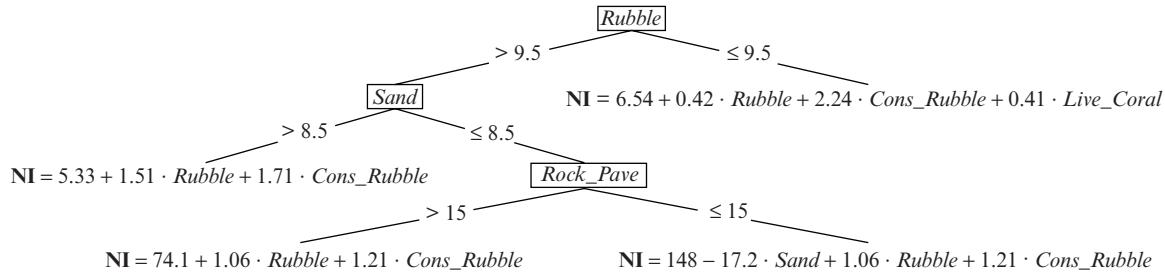

Fig. 19.5 A model tree predicting the number of individuals (NI) of the sea cucumber species *H. Leucospilota* in a 2 × 50 m transect of the sea bed near Rarotonga, Cook Islands

individuals from environmental characteristics of a location.

The spatial unit of analysis was a 2 × 50 m (100 m^2) strip transect: This size was selected to account for the patchy distribution of the animals. A total of 128 sites were sampled for environmental and biological variables. The number of *H. leucospilota* individuals encountered along each transect was recorded. In addition to the species abundance, 10 environmental variables that were expected to have an influence on the habitat preference of the sea cucumber were recorded. These included the exposure of the site (windward or leeward side of the island), and the following microhabitat variables, estimated as a percentage (with possible values from 0% to 100%) of the total 100 m^2 area sampled: Sand, Rubble, Cons_Rubble (consolidated rubble), Boulder, reef rock/pavement (Rock_Pave), live coral (Live_Coral), dead coral (Dead_Coral), mud/silt (Mud_Silt), and Gravel.

The number of *H. leucospilota* individuals was the class variable, while the 10 environmental variables were the attributes. Model tree induction was used to build the habitat model. More specifically, M5', a re-implementation of the system M5 (Quinlan 1992) within the software package WEKA (Witten and Frank 1999) was used. The model tree constructed is given in Fig. 19.5. The correlation coefficient for predictions on unseen cases was estimated to be 0.5 (by using 10-fold cross-validation).

The tree identifies the most important influences of the site characteristics on habitat suitability (rubble and sand, followed by rock pavement, consolidated rubble, and live coral). It identifies four types of sites (one leaf for each) and constructs different linear models to predict the number of sea cucumbers at each.

Two of the site types are essentially not very suitable as sea cucumber habitat: the first (LM1) does not have enough rubble, while the second (LM4) does have enough rubble, but also has too much sand. The average numbers of individuals recorded at the two types of sites are 15 and 35, respectively. One site type (LM2) is very suitable as sea cucumber habitat, as evidenced by the average of 236 animals found per site. This type of site is characterized by enough rubble, little sand and little rock pavement. The last type of site (LM3) represents a moderately suitable habitat for sea cucumbers: it has the same characteristics as the most suitable habitat, except for too much rock pavement. The sea cucumbers prefer larger percentages of rubble and consolidated rubble in all four types of sites (positive coefficients for rubble/consolidated rubble in each of the four linear models).

19.5 Summary and Discussion

In this chapter, we have introduced the task of habitat suitability modeling and formulated it as a machine learning problem. Habitat-suitability modeling studies the effect of the abiotic characteristics of the habitat on the presence, abundance or diversity of a given taxonomic group of organisms. We have briefly described two approaches to machine learning that are often used in habitat suitability modeling: decision tree induction and rule induction.

Applications of machine learning to habitat suitability modeling can be grouped along two dimensions. One dimension is the type of environment where

the studied group of organisms lives, e.g., aquatic (river or sea) or terrestrial (forest or agricultural fields). Another dimension is the type of machine learning approach used, e.g., symbolic (decision trees or classification rules) or statistical (logistic regression or neural networks).

In this chapter, we have given examples of using symbolic machine learning approaches to construct models of habitat suitability for several kinds of organisms in the abovementioned environments. These include habitat models for springtails and other soil organisms in an agricultural setting, brown bears in a forest environment, bioindicator organisms in a river environment, and finally sea cucumbers in a sustainable fishing setting. Many more examples of using machine learning for habitat modeling exist, some of which we point to below. A collection of papers, devoted specifically to the topic of habitat modeling, has been edited by Raven et al. (2002) and describes several applications of machine learning methods.

The author has been involved in quite a few other habitat modeling applications of machine learning, besides those summarized above. These include another, more realistic application in modeling the effects of agricultural actions on soil insects, including mites and collembolans (Demšar et al. 2006). This has been also studied in the context of farming with genetically modified crops and their effects on soil fauna, including earthworms (Debeljak et al. 2005). We have also studied habitat suitability for red deer in Slovenian forests using GIS data, such as elevation, slope, and forest composition (Debeljak et al. 2001).

Neural networks are often used for habitat modeling: several applications are described in (Lek and Guegan 1999). For example, (Lek-Ang et al. 1999) use them to study the influence of soil characteristics, such as soil temperature, water content, and proportion of mineral soil on the abundance and species richness of *Collembola* (springtails). Another study of habitat suitability modeling by neural networks is given by Ozesmi and Ozesmi (1999).

Several habitat-suitability modeling applications of other data mining methods are surveyed by Fielding (1999b). Fielding (1999a) applies a number of methods, including discriminant analysis, logistic regression, neural networks and genetic algorithms, to predict nesting sites for golden eagles. Bell (1999) uses decision trees to describe the winter habitat of pronghorn antelope. Jeffers (1999) uses a genetic algorithm to discover rules that describe habitat preferences for aquatic species in British rivers.

As compared to traditional statistical methods, such as linear and logistic regression, the use of machine learning offers several advantages. On one hand, machine learning methods are capable of approximating nonlinear relationships (typical for the interactions between living organisms and the environment) better than traditional linear approaches. On the other hand, symbolic learning approaches, such as decision trees and classification rules, provide understandable models that can be inspected to give insight into the domain studied.

Let us conclude by mentioning several recent research topics related to the use of machine learning for habitat suitability modeling. These include machine learning methods for simultaneous prediction of several target variables, machine learning methods that are spatially aware and finally the use of habitat suitability modeling in the context of predicting the effects of climate change. We discuss each of these briefly below.

When modeling the habitat of a group of organisms, we are interested in the relation between the environmental variables and the structure of the population at the spatial unit of analysis (absolute and relative abundances of the organisms in the group studied). While one approach to this is to build habitat models for each of the organisms, the alternative approach of building a model that simultaneously predicts the presence/abundance of all organisms in the group is more natural. For this purpose, we can use a neural network with several output nodes that share a common hidden layer. Recently, however, symbolic machine learning approaches have been developed that address this problem, namely predictive clustering trees (Blockeel et al. 1998) and predictive clustering rules (Ženko et al. 2006) for multi-target prediction.

When using machine learning to build habitat models, individual spatial points are treated as training examples. These are assumed to be completely independent and their relative spatial position (proximity) is ignored. This can result in unrealistic predictions of very small patches of habitat: this was, e.g., the case in the brown bear habitat modeling study described

earlier in the chapter. This problem is usually dealt in a post-processing phase, where the prediction of the habitat model for each spatial unit are corrected by taking into account (the predictions for) the neighborhood of that unit. However, spatially aware machine learning methods have recently started to emerge (Lee et al. 2005, Andrienko et al. 2005), although applications of such methods in habitat modeling are still rare.

Finally, let us mention climate change, which is already causing significant changes in the distribution of animals and vegetation across the globe. Predicting future effects along these lines is an emerging area where the use of machine learning for habitat modeling is likely to increase drastically. The idea in this context is to build habitat models for the target groups of organisms, which include climate-related variables, such as mean annual temperature and precipitation. By applying the habitat models to the predictions produced by climate models, one can predict the changes of the distribution of the target group of organisms. For example, Ogris and Jurc (2007) study the change of potential habitats for different tree species under varying climate change scenarios.

Harrison et al. (2006) conduct a more global study where the changes of habitat are investigated for a much larger and more diverse group of organisms. In their study, the availability of suitable climate space across Europe for the distributions of 47 species was modelled. These were chosen to encompass a range of taxa (including plants, insects, birds and mammals) and to reflect dominant and threatened species from 10 habitats. Habitat availability was modelled for the current climate and three climate change scenarios using a neural network model, showing that the distribution of many species in Europe may be affected by climate change, but that the effects are likely to differ between species.

In sum, machine learning methods have been successfully used and are increasingly more often used for habitat modeling, establishing the relations between abiotic characteristics of the environment and the properties of a target population of organisms (such as presence, abundance or diversity). The learned models can be used as tools for the management of the population studied. Perhaps even more importantly, the learned model can enhance our knowledge of the studied population.

References

Allaby, M. (1996). *Basics of environmental science*. London: Routledge.

Andrienko, G., Malerba, D., May, M., & Teisseire, M. (Eds.) (2005). *Proceedings of the ECML/PKDD 2005 Workshop on Mining Spatio-Temporal Data*. Portugal: University of Porto.

Bell, J. F. (1999). Tree based methods. In A. H. Fielding, (Ed.) *Machine learning methods for ecological applications* (pp. 89–105). Dordrecht: Kluwer.

Blockeel, H., De Raedt, L., & Ramon, J. (1998). Top-down induction of clustering trees. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 55–63). San Francisco: Morgan Kaufmann.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Clark, P., & Boswell, R. (1991). Rule induction with CN2: Some recent improvements. *Proceedings of the Fifth European Working Session on Learning* (pp. 151–163). Berlin: Springer.

Debeljak, M., Džeroski, S., Jerina, K., Kobler, A., & Adamič, M. (2001). Habitat suitability modelling of red deer (Cervus elaphus, L.) in South-Central Slovenia. *Ecological Modelling*, *138*, 321–330.

Debeljak, M., Cortet, J., Demšar, D., & Džeroski, S. (2005). Using data mining to assess the effects of Bt maize on soil microarthropods. *Proceedings of the Nineteenth International Conference Informatics for Environmental Protection* (pp. 615-620). Czech Republic: Brno.

Demšar, D., Džeroski, S., Debeljak, M., & Henning Krogh, P. (2006a). Predicting aggregate properties of soil communities vs. community structure in an agricultural setting. *Proceedings of the Twentieth International Conference on Informatics for Environmental Protection* (pp. 295–302). Aachen: Shaker Verlag.

Demšar, D., Džeroski, S., Larsen, T., Struyf, J., Axelsen, J., Bruns-Pedersen, M., & Henning Krogh, P. (2006b). Using multi-objective classification to model communities of soil microarthropods. *Ecological Modelling*, *194*, 131–143.

Džeroski, S. (2001). Applications of symbolic machine learning to ecological modelling. *Ecological Modelling*, *146*, 263-273, 2001.

Džeroski, S., & Drumm, D. (2003). Using regression trees to identify the habitat preference of the sea cucumber (Holothuria leucospilota) on Rarotonga, Cook Island. *Ecological Modelling, 170*, 219–226.

Džeroski, S., Grbović, J., & Blockeel, H. (2001). Predicting river water communities with logical decision trees. *Book of Abstracts, Third European Ecological Modelling Conference*, Croatia: Dubrovnik.

Džeroski, S., Grbović, J., Walley, W. J., & Kompare, B. (1997). Using machine learning techniques in the construction of models. Part II: Rule induction. *Ecological Modelling*, *95*, 95–111.

Fielding, A. H. (1999a). An introduction to machine learning methods. In A. H. Fielding, (Ed.) *Machine learning methods for ecological applications* (pp. 1–35). Dordrecht: Kluwer.

Fielding, A. H. (Ed.) (1999b). *Machine learning methods for ecological applications*. Dordrecht: Kluwer.

Harrison, P. A., Berry, P. M., Butt, N., & New, M. (2006). Modelling climate change impacts on species distributions at the European scale: implications for conservation policy. *Environmental Science and Policy*, *9*(2). 116–128.

Jeffers, J. N. R. (1999). Genetic algorithms I. In A. H. Fielding (Ed.) , *Machine learning methods for ecological applications* (pp. 107–121). Dordrecht: Kluwer.

Jerina, K., Debeljak, M., Kobler, A., & Adamič, M. (2003). Modeling the brown bear population in Slovenia: A tool in the conservation management of a threatened species. *Ecological Modelling*, *170*, 453-469.

Joergensen, S. E., & Bendoricchio, G. (2001). *Fundamentals of ecological modelling*. Amsterdam: Elsevier.

Kampichler, C., Džeroski, S., & Wieland, R. (2000). The application of machine learning techniques to the analysis of soil ecological data bases: Relationships between habitat features and *Collembola* community characteristics. *Soil Biology and Biochemistry 32*, 197–209.

Krebs, C. J. (1972). *Ecology*. New York: Harper and Row.

Krebs, C. J. (1989). *Ecological methodology*. HarperCollins, New York, NY.

Lee, C.-H., Greiner, R., & Schmidt, M. (2005). Support vector random fields for spatial classification. *Proceedings of the Ninth European Conference on Principles and Practice of Knowledge Discovery in Databases* (pp. 121–132). Berlin: Springer.

Lek-Ang, S., Deharveng, L., & Lek, S. (1999). Predictive models of collembolan diversity and abundance in a riparian habitat. *Ecological Modelling*, *120*(2-3), 247–260.

Lek, S., & Guegan, J. F. Guest (Eds.) (1999). Application of artificial neural networks in ecological modelling. Special issue of *Ecological Modelling 120*(2-3).

Mitchell, T. (1997). *Machine learning*. New York: McGraw-Hill.

Ogris, N., & Jurc, M. (2007). Potential changes in the distribution of maple species (*Acer pseudoplatanus, A. campestre, A. platanoides, A. obtusatum*) due to climate change in Slovenia. *Proceedings of the Symposium on Climate Change Influences on Forests and Forestry*. Slovenia: University of Ljubljana.

Ozesmi, S. L., & Ozesmi, U. (1999). An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling*, *116*(1), 15–31.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106.

Quinlan J. R. (1992). Learning with continuous classes. *Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence* (pp. 343–348). Singapore: World Scientific.

Quinlan, J. R. (1993). *Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Raven, P. H., Scott, J. M., Heglund, P., & Morrison, M. (2002). *Predicting species occurrences: Issues of accuracy and scale*. Washington, DC: Island Press.

Witten, I. H., & Frank, E. (1999). *Data mining: Practical machine learning tools and techniques with Java implementations*. San Francisco: Morgan Kaufmann.

Ženko, B., Džeroski, S., & Struyf, J. (2006). Learning predictive clustering rules. *Proceedings of the Fourth International Workshop on Knowledge Discovery in Inductive Databases* (pp. 234–250). Berlin: Springer.