

Habitat modelling with single- and multi-target trees and ensembles

Name Surname*, Name2 Surname2

*Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000
Ljubljana, Slovenia*

Abstract

This is not an abstract.

Keywords: Habitat modelling, Predictive clustering trees, Ensemble models, Multi-target prediction

1. Introduction

Ecology is frequently defined as the study of the distributions and abundances of organisms across space and time and their interactions with the environment [1]. The distribution can be considered along the spatial dimension(s) and/or the temporal dimension. Within ecology, the topic of ecological modeling [2] 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 (temporal dimension) and habitat suitability for a given species (spatial dimension). We further focus on the latter.

Habitat suitability/modelling 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

*Corresponding author (telephone: +386 1 XXXXXXXX)

Email addresses: mail@ijs.si (Name Surname), mail2@ijs.si (Name2 Surname2)

and animals. This is typically done under the implicit assumption that both are observed at a single point in time for a given spatial unit.

The input to a habitat model [3, 4] is a set of environmental characteristics for a given spatial unit of analysis. These environmental characteristics (i.e. environmental variables) may be of three different types. The first type concerns abiotic properties of the environment, e.g., physical and chemical characteristic thereof. The second type concerns some biological aspects of the environment, which may be considered as an external impact on the group of organisms under study. Finally, the variables of the third type are related to human activities and their impacts on the environment. The output of a habitat model is a target property of the given (taxonomic) group of organisms. Note that the type of environmental variables, as well as the size of the spatial unit, can vary considerably, depending on the context, and so can the target property of the population (even though to a lesser extent). If we take the abundance or density of the population 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 (e.g., Shannon index, species richness).

Machine learning (and in particular predictive modeling) is increasingly often used to automate the construction of ecological models [4]. 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 [5]. 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. The most popular machine learning techniques used for modeling habitat suitability include decision tree induction [6], rule induction [7], and neural networks .

In the most general case of habitat modelling, 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. An alternative approach is to build a model that simultaneously predicts the presence/abundance of all organisms in the group.

In this work, we investigate in more detail the later approach: predicting the presence of all organisms from a group, i.e., community modelling. To this end, we propose to use predictive clustering trees [8], and in particular their instantiation for predicting multiple targets [9] (called multi-target decision trees). This approach has several advantages over constructing a model for each species separately and then aggregate/combine their outputs. To begin with, they exploit the relations and connections that may exist between the species from the group. Next, they are smaller and faster to learn. Finally, it is easier to interpret a single multi-target decision tree than set of single-target decision trees.

To further increase the predictive performance of the predictive clustering trees we construct ensembles. Ensembles are set of predictive models, called base predictive models. The predictions of the base predictive models are combined by some combination scheme to obtain the overall prediction. There are many theoretical and empirical studies that show that ensembles lift the predictive performance of the base predictive models and offer high predictive performance [10, 11, 12]

We explore the potential of predictive clustering trees and ensembles thereof for community structure modelling on three case studies. First, we present community structure modelling in Slovenian rivers [13, 14]. Next, we present community modelling in soil samples from experimental farming systems in Denmark [15]. Finally, we show a case study for diatom community modelling in lake Prespa, Macedonia [16].

The remainder of this paper is organized as follows. In Section 2, we provide the relevant background for habitat modelling and machine learning. We present the methodology for community structure modelling in Section 3. In Section 4, we show three case studies that concern river, lake and soil communities of organisms. Finally, we state the conclusions in Section 5.

2. Background

In this section, we give the background for the work presented here. We first present the habitat modelling within the framework of ecology. Namely, we present the types of environmental variables that are typically encountered

in the habitat modelling applications and discuss several issues that need to be considered during the construction of habitat models. Then, we define the machine learning tasks of classification, regression in the context of habitat modelling.

2.1. Habitat modelling

If ecology is defined as the study of the distribution and abundance of plants and animals, habitat suitability modelling is concerned with the spatial aspects of the distribution and abundance. Habitat suitability models relate the spatially varying characteristics of the environment on the presence, abundance or diversity of a 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 soil microarthropods habitat (Section 4.3 and [15]), the soil samples taken were of size 6cm diameter and 5.5cm depth, in a study of sea cucumber [17] habitat transects of 2m by 50m of the sea bed were considered, and in some studies of potential habitats for different tree species under varying climate change scenarios [18], 1km by 1km squares are considered. 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, the second kind concerns some biological aspects of the environment and 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 modelling 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 modelling 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.

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).

In the most general case of habitat modelling, 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.

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 non-trivial

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 areas of activity (home ranges) and sample from these to obtain data for learning habitat suitability models.

Another issue that commonly occurs in habitat modelling, 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 modelling 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 modelling 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 modelling. 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).

2.2. Machine learning for habitat modelling

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.

In the case of habitat modelling, 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 classes represent the target property of interest for each of the organisms, such as presence, abundance or diversity for each organism. This is illustrated in Figure 1.

Sample ID	Descriptive variables						Target variables													
	Temperature	K ₂ Cr ₂ O ₇	NO ₂	Cl	CO ₂	...	<i>Cladophora</i> sp.	<i>Gongrosira</i> incrustans	<i>Oedogonium</i> sp.	<i>Stigeoclonium</i> tenue	<i>Melosira</i> varians	<i>Nitzschia</i> palea	<i>Audouinella</i> chalybea	<i>Erpobdella</i> octoculata	<i>Gammarus</i> fossarum	<i>Baelis</i> rhodani	<i>Hydropsyche</i> sp.	<i>Rhyacophila</i> sp.	<i>Simulium</i> sp.	<i>Tubifex</i> sp.
ID1	0.66	0.00	0.40	1.46	0.84	...	1	0	0	0	0	1	1	0	1	1	1	1	1	1
ID2	2.03	0.16	0.35	1.74	0.71	...	0	1	0	1	1	1	1	0	1	1	1	1	1	0
ID3	3.25	0.70	0.46	0.78	0.71	...	1	1	0	0	1	0	1	0	1	1	1	0	1	1
...

Figure 1: An example of data table for habitat modelling of bioindicator organisms [14]. The descriptive variables are chemical parameters of water samples, while the target variables are the abundances of 14 bioindicator organisms.

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 attribute from the values of other attributes. The target attribute is called the class (dependent variable in statistical terminology). The other attributes 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 modelling refers to both classification and regression.

We further extend the tasks of classification and regression for predicting multiple target attributes. Here, instead of predicting one target attribute, the predictive models make a prediction for multiple target attributes. We formally define the task of multi-target prediction as follows. **Given:**

- A description space X that consists of tuples of values of primitive data types (boolean, discrete or continuous), i.e., $\forall X_i \in X, X_i = (x_{i_1}, x_{i_2}, \dots, x_{i_D})$, where D is the size of the tuple (or number of descriptive variables),
- a target space Y which is a tuple of several variables that can be either

continuous or discrete, i.e., $\forall Y_i \in Y, Y_i = (y_{i_1}, y_{i_2}, \dots, y_{i_T})$, where T is the size of the tuple (i.e., number of target variables),

- a set of examples E , where each example is a pair of tuples from the description and target space, respectively, i.e., $E = \{(X_i, Y_i) | X_i \in X, Y_i \in Y, 1 \leq i \leq N\}$ and N is the number of examples of E ($N = |E|$), and
- a quality criterion q , which rewards models with high predictive accuracy and low complexity.

Find: a function $f : X \rightarrow Y$ such that f maximizes q . In this paper, the function f is represented with decision trees, i.e., predictive clustering trees or ensembles thereof.

The machine learning task of habitat modelling 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 classes corresponding to a target property for each of the species studied. The goal is to learn a predictive model that predicts the target property for each species from the environmental variables (from the given dataset). If we are only looking at presence/absence or suitable/unsuitable as values of the classes, we have a multi-target classification problem. If we are looking at the degree of suitability (density/abundance), we have a multi-target regression problem. Note that, the traditional classification and regression are special cases of multi-target classification and multi-target regression, respectively, when the class concerns a target property of single species.

3. Methodology

In this Section, we present the machine learning methodology used to obtain the habitat models. We first describe the predictive clustering trees for multi-target classification and regression. Then, we present the used ensemble methods.

3.1. Predictive clustering trees

The Predictive Clustering Trees (PCTs) framework sees a decision tree as a hierarchy of clusters: the top-node corresponds to one cluster containing

all data, which is recursively partitioned into smaller clusters while moving down the tree. The PCT framework is implemented in the CLUS system [19], which is available for download at <http://www.cs.kuleuven.be/~dtai/clus>.

CLUS takes as input a set of examples $E = \{(x_i, y_i) | i = 1, \dots, N\}$, where each x_i is a vector of attribute values and y_i are values of a structured (output) datatype T_Y . In this thesis, we consider three different classes of datatypes T_Y : tuples of discrete values, tuples of real values, and hierarchies of classes. For each type T_Y , CLUS needs two functions to be defined. The prototype function returns a representative structured value given a set of such values. The variance function describes how homogeneous a set of structured values is: it is typically based on a distance function on the space of structured values.

PCTs can be induced with a standard *top-down induction of decision trees* (TDIDT) algorithm [6]. The algorithm is presented in Table 1. It takes as input a set of examples (E) and outputs a tree. The heuristic (h) that is used for selecting the tests (t) is the reduction in variance caused by partitioning (\mathcal{P}) the instances (see line 4 of BestTest procedure in Table 1). By maximizing the variance reduction the cluster homogeneity is maximized and it improves the predictive performance. If no acceptable test can be found (see line 6), that is, if the test does not significantly reduces the variance, then the algorithm creates a leaf and computes the prototype of the instances belonging to that leaf.

Table 1: The top-down induction algorithm for PCTs.

procedure PCT(E) returns tree	procedure BestTest(E)
1: $(t^*, h^*, \mathcal{P}^*) = \text{BestTest}(E)$	1: $(t^*, h^*, \mathcal{P}^*) = (\text{none}, 0, \emptyset)$
2: if $t^* \neq \text{none}$ then	2: for each possible test t do
3: for each $E_i \in \mathcal{P}^*$ do	3: $\mathcal{P} =$ partition induced by t on E
4: $\text{tree}_i = \text{PCT}(E_i)$	4: $h = \text{Var}(E) - \sum_{E_i \in \mathcal{P}} \frac{ E_i }{ E } \text{Var}(E_i)$
5: return $\text{node}(t^*, \bigcup_i \{\text{tree}_i\})$	5: if $(h > h^*) \wedge \text{Acceptable}(t, \mathcal{P})$ then
6: else	6: $(t^*, h^*, \mathcal{P}^*) = (t, h, \mathcal{P})$
7: return $\text{leaf}(\text{Prototype}(E))$	7: return $(t^*, h^*, \mathcal{P}^*)$

The main difference between the algorithm for learning PCTs and a standard decision tree learner (for example, see the C4.5 algorithm proposed by Quinlan [20]) is that the former considers the variance function and the prototype function, that computes a label for each leaf, as parameters that can

be instantiated for a given learning task. So far, the PCTs have been instantiated for the following tasks: multiple targets prediction [9, 21], hierarchical-multi label classification [22] and prediction of time-series [23]. In this thesis, we focus on the first two tasks.

PCTs that are able to predict multiple targets simultaneously are called multi-target decision trees (MTDTs). The MTDTs that predict a tuple of continuous variables (regression tasks) are called multi-target regression trees (MTRTs), while the MTDTs that predict a tuple of discrete variables are called multi-target classification trees (MTCTs). The instantiation of the CLUS system that learns multi-target trees is called CLUS-MTDT.

3.1.1. PCTs for multi-target classification

An example of a MTCT is shown in Figures 3 and 4. The internal nodes of the tree contain tests on the descriptive variables (in this case, chemical parameters of the water samples) and the leaves store the predictions (in this case, which species are encountered and which not in a given water sample).

The variance function for the MTCTs is computed as the sum of the Gini indices of the target variables, i.e., $Var(E) = \sum_{i=1}^T Gini(E, Y_i)$. Furthermore, one can also use the sum of the entropies of class variables as a variance function, i.e., $Var(E) = \sum_{i=1}^T Entropy(E, Y_i)$ (this definition has also been used in the context of multi-label prediction [24]). Note that in the single target case, $Var(E) = Entropy(E)$ corresponds to information gain.

The prototype function returns a vector of probabilities that an instance belongs to a given class for each target variable. Using these probabilities, the most probable (majority) class for each target attribute can be calculated. In addition to the two aforementioned instantiations of the variance function for classification problems, the CLUS system also implements other variance functions, such as reduced error, gain ratio and the m -estimate.

3.1.2. PCTs for multi-target regression

An example of a MTRT is shown in Figures 2 and 5. The internal nodes of a PCT for multi-target regression, similar as for multi-target classification, contain tests on the descriptive variables and the leaves store the predictions. The MTRTs look similar as MTCTs, with the difference, that in this case, the prototype in each leaf is the mean value instead of the majority class.

The variance and prototype functions for MTRTs are instantiated as follows. The variance is calculated as the sum of the variances of the target variables, i.e., $Var(E) = \sum_{i=1}^T Var(Y_i)$. The variances of the targets are

normalized, so each target contributes equally to the overall variance. The prototype function (calculated at each leaf) returns as a prediction the tuple with the mean values of the target variables, calculated by using the training instances that belong to the given leaf.

3.2. Ensemble methods

An ensemble is a set of predictive models (called base predictive models). In homogeneous ensembles, such as the ones we consider here, the base predictive models are constructed by using the same algorithm. The prediction of an ensemble for a new instance is obtained by combining the predictions of all base predictive models from the ensemble. In this article, we consider ensembles of PCTs for multi-target prediction [21]. The PCTs in the ensembles are constructed by using bagging and random forests methods that are often used in the context of decision trees. We have adapted these methods to use PCTs.

A necessary condition for an ensemble to have better predictive performance than any of its individual members, is that the base predictive models are accurate and diverse [25]. An accurate predictive model does better than random guessing on new examples. Two predictive models are diverse if they make different errors on new examples. There are several ways to introduce diversity in a set of base predictive models: by manipulating the training set (by changing the weight of the examples [26, 27], by changing the attribute values of the examples [28], by manipulating the feature space [29, 30]) and by manipulating the learning algorithm itself [29, 31].

The prediction of an ensemble for a new instance is obtained by combining the predictions of all the base predictive models from the ensemble. The predictions from the models can be combined by taking the average (for regression tasks) and the majority or probability distribution vote (for classification tasks), as described in [32, 26], or by taking more complex aggregation schemes [11].

We use PCTs as base predictive models for the ensembles for multi-target prediction. To obtain a prediction from an ensemble for multi-target prediction, we accordingly extend the voting schemes. For the task of multi-target regression, as prediction of the ensemble, we take average per target of the predictions of the base predictive models. We obtain the ensemble predictions for the multi-target classification using probability distribution voting (as suggested by Bauer and Kohavi [32]) per target.

We have implemented the bagging and random forests methods within the CLUS system. These two ensemble learning techniques are most widely known and have primarily been used in the context of decision trees. The algorithms of these ensemble learning methods are presented in Table 2. For the random forests method (right in Table 2), the PCT algorithm for multi-target prediction needed changes: A randomized version of the selection of attributes was implemented, which replaced the standard selection of attributes.

3.2.1. *Bagging*

Bagging [26] is an ensemble method that constructs the different classifiers by making bootstrap replicates of the training set and using each of these replicates to construct a predictive model (Table 2, left). Each bootstrap sample is obtained by randomly sampling training instances, with replacement, from the original training set, until an equal number of instances as in the training set is obtained. Breiman [26] showed that bagging can give substantial gains in predictive performance, when applied to an unstable learner (i.e., a learner for which small changes in the training set result in large changes in the predictions), such as classification and regression tree learners.

3.2.2. *Random forests*

A random forest [29] is an ensemble of trees, where diversity among the predictors is obtained by using bootstrap replicates as in bagging, and additionally by changing the set of descriptive attributes during learning (Table 2, right). More precisely, at each node in the decision trees, a random subset of the descriptive attributes is taken, and the best attribute is selected from this subset. The number of attributes that are retained is given by a function f of the total number of descriptive attributes D (e.g., $f(D) = 1$, $f(D) = \lfloor \sqrt{D} + 1 \rfloor$, $f(D) = \lfloor \log_2(D) + 1 \rfloor \dots$). By setting $f(D) = D$, we obtain the bagging procedure. The algorithm for learning a random forest using PCTs as base classifiers is presented in Table 2.

Table 2: The ensemble learning algorithms: bagging and random forests. Here, E is the set of the training examples, k is the number of trees in the forest, and $f(D)$ is the size of the feature subset that considered at each node during tree construction for random forests.

procedure Bagging(E, k)	procedure RForest($E, k, f(D)$)
returns Forest	returns Forest
1: $F = \emptyset$	1: $F = \emptyset$
2: for $i = 1$ to k do	2: for $i = 1$ to k do
3: $E_i = \text{bootstrap}(E)$	3: $E_i = \text{bootstrap}(E)$
4: $T_i = PCT(E_i)$	4: $T_i = PCT_rnd(E_i, f(D))$
5: $F = F \cup \{T_i\}$	5: $F = F \cup \{T_i\}$
6: return F	6: return F

4. Case studies

4.1. Experimental design

4.2. Water quality prediction

4.2.1. Data description

4.2.2. Habitat models

- Water quality paper

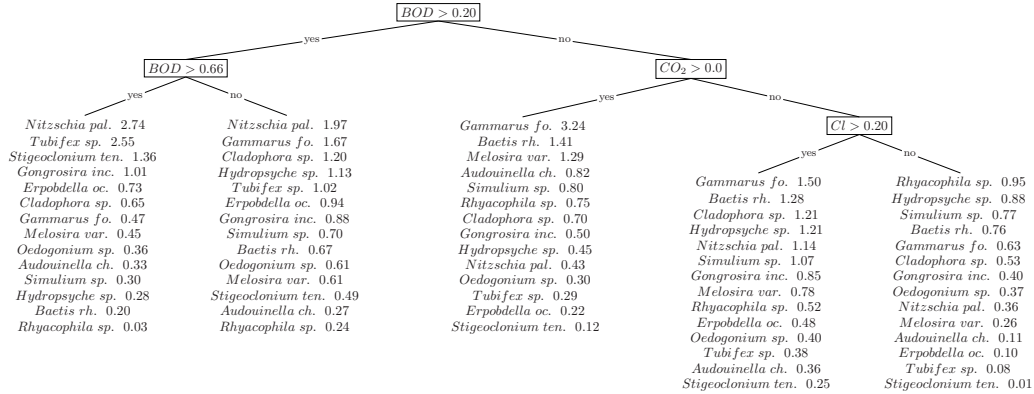


Figure 2: The habitat model for water quality with continuous target variables.

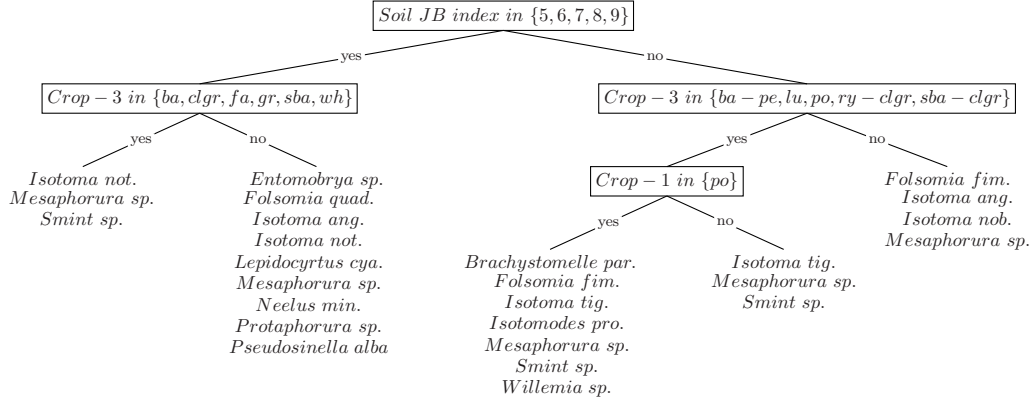


Figure 3: The habitat model for soil quality with discrete target variables.

4.3. Soil quality prediction

4.3.1. Data description

4.3.2. Habitat models

- Soil quality paper

4.4. Prespa diatoms

This case study was published by Kocev et al. [16]. It concerns modelling of the diatom communities in the lake Prespa, Macedonia.

4.4.1. Data description

Lake Prespa is located at the border intersection of Macedonia, Albania and Greece. Monitoring of the state of Lake Prespa was performed during one and a half year period (from March 2005 to September 2006). Samples for analysis were taken from the surface water of the lake at 14 locations. The lake sampling locations are distributed in the three countries as follows: 8 in Macedonia, 3 in Albania and 3 in Greece. The selected sampling locations are representative for determining the eutrophication impact [33].

In total, a total of 218 water samples from the lake Prespa were collected. On these water samples, both physicochemical and biological analyses were performed. The physicochemical properties of the samples provided the environmental variables for the habitat models, while the biological samples provided information on the relative abundance of the studied diatoms. The following physicochemical properties of the water samples were measured: temperature, dissolved oxygen, Secchi depth, conductivity, alkalinity (pH),

nitrogen compounds (NO₂, NO₃, NH₄, inorganic nitrogen), sulphur oxide ions SO₄, and Sodium (Na), Potassium (K), Magnesium (Mg), Copper (Cu), Manganese (Mn) and Zinc (Zn).

The biological variables were the relative abundances of 116 different diatom species. Diatom cells were collected with a planktonic net or as attached growth on submerged objects (plants, rocks or sand and mud). The sample is examined with a microscope, and the diatom taxa and abundance in the samples are obtained by counting 200 cells per sample. The specific specie abundance is then given as the percent of the total diatom count per sampling site.

4.4.2. Habitat models

- Prespa paper

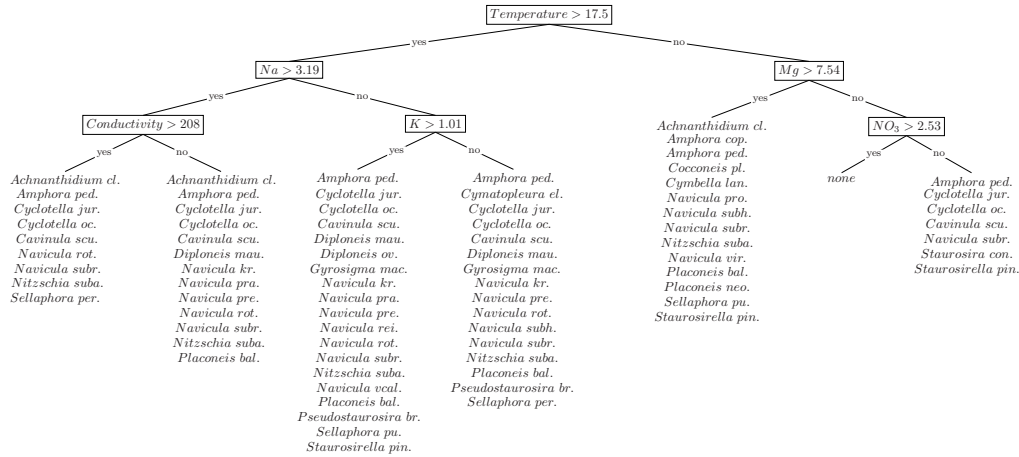


Figure 4: The habitat model for all diatoms with discrete target variables.

4.5. Summary

5. Conclusions

References

- [1] M. Begon, C. Townsend, J. Harper, Ecology: From individuals to ecosystems, Blackwell, 2006.
- [2] S. E. Jørgensen, G. Bendoricchio, Fundamentals of Ecological Modelling, Elsevier, 2001.

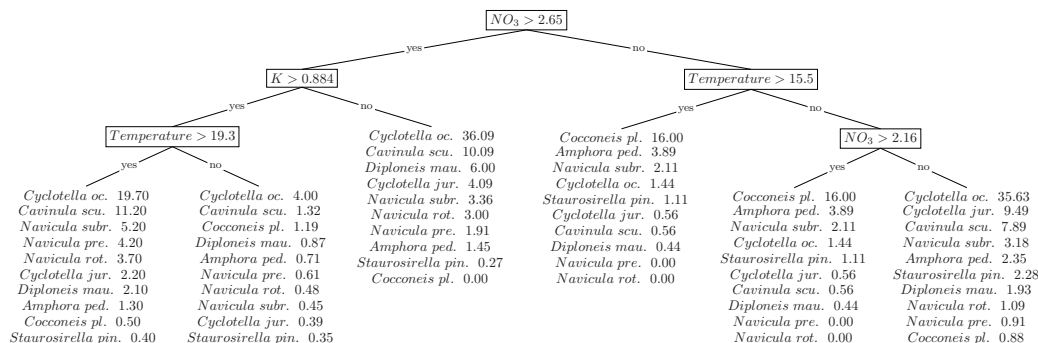


Figure 5: The habitat model for top 10 diatoms with continuous target variables.

- [3] S. Džeroski, Machine learning applications in habitat suitability modeling, in: Artificial Intelligence Methods in the Environmental Sciences, Springer Berlin, 2009, pp. 397–412.
- [4] S. Džeroski, Applications of symbolic machine learning to ecological modelling, Ecological Modelling 146 (1-3) (2001) 263–273.
- [5] T. Mitchell, Machine learning, McGraw Hill, 1997.
- [6] L. Breiman, J. Friedman, R. Olshen, C. J. Stone, Classification and Regression Trees, Chapman & Hall/CRC, 1984.
- [7] P. Clark, R. Boswell, Rule induction with cn2: Some recent improvements, in: Proc. of the European Working Session on Machine Learning, Springer-Verlag, 1991, pp. 151–163.
- [8] H. Blockeel, Top-down induction of first order logical decision trees, Ph.D. thesis, Katholieke Universiteit Leuven, Leuven, Belgium (1998).
- [9] J. Struyf, S. Džeroski, Constraint based induction of multi-objective regression trees, in: Proc. of the 4th International Workshop on Knowledge Discovery in Inductive Databases KDID - LNCS 3933, Springer, 2006, pp. 222–233.
- [10] G. Seni, J. F. Elder, Ensemble methods in data mining: Improving accuracy through combining predictions, Morgan & Claypool Publishers, 2010.

- [11] L. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, Wiley-Interscience, 2004.
- [12] S. Džeroski, P. Panov, B. Ženko, Ensemble methods in machine learning, in: Encyclopedia of complexity and systems science, Springer New York, 2009, pp. 5317–5325.
- [13] H. Blockeel, S. Džeroski, J. Grbović, Simultaneous prediction of multiple chemical parameters of river water quality with tilde, in: Proceedings of the 3rd European Conference on PKDD - LNAI 1704, Springer, 1999, pp. 32–40.
- [14] S. Džeroski, D. Demšar, J. Grbović, Predicting chemical parameters of river water quality from bioindicator data, Applied Intelligence 13 (1) (2000) 7–17.
- [15] D. Demšar, S. Džeroski, T. Larsen, J. Struyf, J. Axelsen, M. Bruns-Pedersen, P. H. Krogh, Using multi-objective classification to model communities of soil, Ecological Modelling 191 (1) (2006) 131–143.
- [16] D. Kocev, A. Naumoski, K. Mitreski, S. Krstić, S. Džeroski, Learning habitat models for the diatom community in lake prespa, Ecological Modelling 221 (2) (2010) 330–337.
- [17] S. Džeroski, D. Drumm, Using regression trees to identify the habitat preference of the sea cucumber (*holothuria leucospilota*) on rarotonga, cook islands, Ecological Modelling 170 (2-3) (2003) 219 – 226.
- [18] N. Ogris, M. Jurc, Potential changes in the distribution of maple species (*acer pseudoplatanus*, *a. campestre*, *a. platanoides*, *a. obtusatum*) due to climate change in slovenia, in: Proc. of the Symposium on Climate Change Influences on Forests and Forestry, Univ. Ljubljana, Slovenia, 2007, pp. 335–358.
- [19] H. Blockeel, J. Struyf, Efficient algorithms for decision tree cross-validation, Journal of Machine Learning Research 3 (2002) 621–650.
- [20] R. J. Quinlan, C4.5: Programs for Machine Learning, 1st Edition, Morgan Kaufmann, 1993.

- [21] D. Kocev, C. Vens, J. Struyf, S. Džeroski, Ensembles of multi-objective decision trees, in: ECML '07: Proceedings of the 18th European Conference on Machine Learning – LNCS 4701, Springer Berlin / Heidelberg, 2007, pp. 624–631.
- [22] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, H. Blockeel, Decision trees for hierarchical multi-label classification, *Machine Learning* 73 (2) (2008) 185–214.
- [23] I. Slavkov, V. Gjorgjioski, J. Struyf, S. Džeroski, Finding explained groups of time-course gene expression profiles with predictive clustering trees, *Molecular BioSystems* 6 (4) (2010) 729–740.
- [24] A. Clare, Machine learning and data mining for yeast functional genomics, Ph.D. thesis, University of Wales Aberystwyth, Aberystwyth, Wales, UK (2003).
- [25] L. K. Hansen, P. Salamon, Neural network ensembles, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12 (10) (1990) 993–1001.
- [26] L. Breiman, Bagging predictors, *Machine Learning* 24 (2) (1996) 123–140.
- [27] Y. Freund, R. E. Schapire, Experiments with a new boosting algorithm, in: Proc. of the Thirteenth International Conference on Machine Learning - ICML, Morgan Kaufman, 1996, pp. 148–156.
- [28] L. Breiman, Using iterated bagging to debias regressions, *Machine Learning* 45 (3) (2001) 261–277.
- [29] L. Breiman, Random forests, *Machine Learning* 45 (1) (2001) 5–32.
- [30] T. K. Ho, The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (8) (1998) 832–844.
- [31] T. G. Dietterich, Ensemble methods in machine learning, in: Proc. of the 1st International Workshop on Multiple Classifier Systems - LNCS 1857, Springer, 2000, pp. 1–15.

lcccccc

Table 3: Testing accuracy on the datasets with discrete target variables.

- [32] E. Bauer, R. Kohavi, An empirical comparison of voting classification algorithms: Bagging, boosting, and variants, Machine Learning 36 (1) (1999) 105–139.
- [33] S. Krstić, Description of sampling sites, Report on baseline data for water (surface and groundwater) including waste related data for the target region - EC-FP6 project TRABOREMA, EC-Project Contract No. INCO-CT-2004-509177, Deliverable 2.2 (2005).

6. Appendix

6.1. Performance of the used machine learning methods

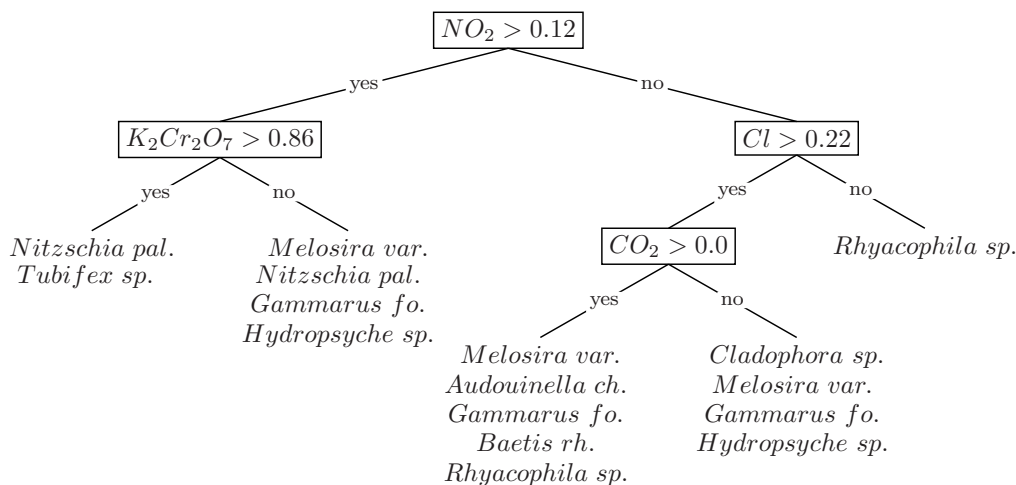


Figure 6: The habitat model for water quality with discrete target variables.

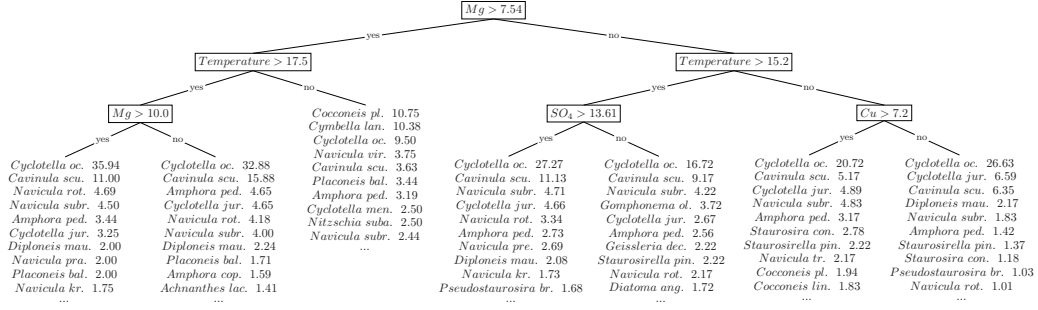


Figure 7: The habitat model for all diatoms with continuous target variables.

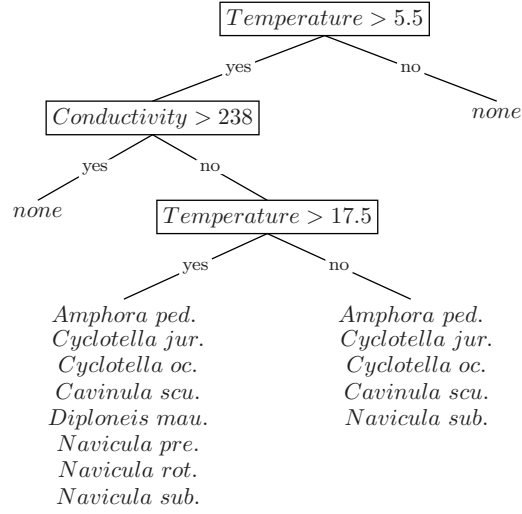


Figure 8: The habitat model for top 10 diatoms with discrete target variables.