Available online at www.sciencedirect.com
ScienceDirect
00 (2013) 1–21
Ecological Informatics

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

Dragi Kocev, Sašo Džeroski*

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

Abstract

Habitat modelling studies the influence of abiotic factors on the abundance of a given taxonomic group of organisms. In this work, we investigate the effect of environmental conditions on communities of organisms in three different ecosystems. Namely, we consider the diatom community in Lake Prespa, Macedonia, the *Collembola* community in the soils of Denmark and 14 organisms living in Slovenian rivers. The data for these case studies consist of physical and chemical properties of the environment as well as the relative abundances or presence of the organisms under investigation.

The multi-species data are analyzed by constructing habitat models for each species separately (single-target decision trees) or by constructing a single habitat model for all the species (multi-target predictive clustering trees). Typically, habitat models are constructed for each species individually and thus do not exploit the interactions between/among species. While approaches for building a single habitat model of a group of organisms exist, they typically construct models that are not readily interpretable and, thus, are seldom used by the research community. In this work, we explore in detail the construction of interpretable models of both types. Furthermore, we construct ensembles of decision trees and ensembles of predictive clustering trees to increase the predictive performance of the models.

The key outcomes of the interpretation and discussion of the obtained models for each case study are as follows. First, we show that multi-target predictive clustering trees are a very useful method for the analysis of multi-species data and that they are more efficient and produce more concise models than single-target decision trees. The obtained multi-target habitat models are readily interpretable and identify the environmental conditions that influence the composition and structure of a given community of organisms. Second, we conclude that the temperature and magnesium are the most important factors influencing the complete diatom community in Lake Prespa, while the nitrates and the temperature influence more the most abundant species. Third, the biological oxygen demand is the most influential factor for the abundance of river dwelling species, while the river community structure is mostly influenced by the NO_2 concentration. Finally, the structure of the community of soil microarthropods is mostly influenced by the soil type and the crop history.

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

*Corresponding author (telephone: +386 1 477 3217)
Email addresses: Dragi.Kocev@ijs.si (Dragi Kocev), Saso.Dzeroski@ijs.si (Sašo Džeroski)

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 (Begon et al., 2006). The distribution can be considered along the spatial dimension(s) and/or the temporal dimension. Within ecology, the topic of ecological modeling (Jørgensen 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 (temporal dimension) and habitat modelling for a given species (spatial dimension). We further focus on the latter, but consider a set (community) of species rather than a single one.

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. 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 (Džeroski, 2009, 2001) 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 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). An example dataset for habitat modelling is illustrated in Figure 1.

	Descriptive variables						**Target variables**													
Sample ID	Temperature	$K_2Cr_2O_7$	NO_2	Cl	CO_2	...	*Cladophora sp.*	*Gongrosira incrustans*	*Oedogonium sp.*	*Stigeoclonium tenue*	*Melosira varians*	*Nitzschia palea*	*Audouinella chalybea*	*Erpobdella octoculata*	*Gammarus fossarum*	*Baetis rhodani*	*Hydropsyche sp.*	*Rhyacophila sp.*	*Simulim sp.*	*Tubifex 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 data table for habitat modelling of bioindicator organisms for river water quality (Džeroski et al., 2000). The descriptive variables are chemical parameters of water samples, while the target variables are the presence/absence of the 14 bioindicator organisms.

Machine learning (and in particular predictive modeling) is increasingly often used to automate the construction of ecological models (Džeroski, 2001). There is a plethora of environmental studies that utilize different statistical and machine learning techniques to perform habitat modelling. The main techniques used

in this context are generalized linear models, generalized additive models, classification and regression trees, tree ensembles (random forests, bagging and boosting), fuzzy models, artificial neural networks (ANNs), support vector machines (SVMs) and genetic algorithms. Depending of the context of the study, different methods should be preferred. If the goal is to produce maps of the habitat suitability then the methods with best predictive performance should be preferred (such as, ensembles and SVMs). On the other hand, if the goal is to obtain further understanding concerning the eco-system under consideration then the interpretable methods with satisfactory predictive performance should be preferred (such as, classification and regression trees/rules). For further details we refer the reader to the work of Elith et al. (2006), Araújo and New (2007), Kampichler et al. (2010), Pino-Mejías et al. (2010), Oppel et al. (2012), Drew et al. (2011), Franklin (2009) and Scott et al. (2002).

In the most general case of habitat modelling, we are interested in the relation between the environmental variables and the structure of the population (absolute and relative abundances of the organisms in the group studied) at the spatial unit of analysis. One approach to solving this task is to build habitat models for each of the species 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. Habitat modelling studies typically focus on the former approach, i.e., habitat modelling for individual species. While there exist methods that take the latter approach, i.e., directly exploit the multi-species data (such as multivariate adaptive regression splines (Friedman, 1991), ANNs or clustering (Lek et al., 2005)), these are seldom applied by the modelling community. The most prominent reason for this is probably the lack of interpretability of the models produced by the aforementioned methods.

Considering all this, the main focus of the work presented here is to investigate in more detail the later approach: predicting the presence of all organisms from a group, i.e., community modelling. Our main hypothesis is that we can produce readily interpretable models of the community structure (i.e., models valid for the whole community) and thus increase the insight into the observed eco-system. To this end, we propose to use predictive clustering trees (Blockeel, 1998), and in particular their instantiation for predicting multiple targets (called multi-target predictive clustering trees) for modelling the community structure. We then compare this approach to standard single-target decision trees (Breiman et al., 1984). Our approach has several advantages over constructing a model for each species separately and then aggregating/combining their outputs. To begin with, it exploits the relations and interactions that may exist between the species from the group. Next, the considered models are smaller and faster to learn. Furthermore, from a clustering point of view, the PCTs are unique in the sense that they provide cluster descriptions while constructing the clusters. Finally, it is easier to interpret a single multi-target predictive clustering tree (PCT) than set of single-target decision trees (DTs).

To further increase the predictive performance of the PCTs, we construct ensembles. Ensembles are sets of predictive models, called base predictive models. The predictions of the base predictive models are combined by some combination scheme (such as voting or averaging) 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 (Seni and Elder, 2010; Kuncheva, 2004; Džeroski et al., 2009).

We explore the potential and demonstrate the advantages of using predictive clustering trees and ensembles thereof for community structure modelling on three case studies. First, we model the diatom community in Lake Prespa, Macedonia (Kocev et al., 2010). Then, we model the community structure of organisms living in Slovenian rivers (Blockeel et al., 1999; Džeroski et al., 2000). Finally, we model the community of *Collembola* species living in the soil of experimental farming systems in Denmark (Demšar et al., 2006).

The remainder of this paper is organized as follows. In Section 2, we describe the methodology for community structure modelling with single- and multi-target trees and ensembles thereof. In Section 3, we present the datasets, the experimental design and results. Next, we discuss the habitat models for the communities of organisms living in a lake, a river and in the soil in Section 4. Finally, Section 5 concludes.

2. Methodology

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

2.1. Predictive clustering trees

Predictive clustering trees (PCTs) (Blockeel, 1998) are a generalization of decision trees (Breiman et al., 1984) and can be used for a variety of learning tasks, such as multi-target prediction, time-series prediction and (hierarchical) multi-label classification. The PCT framework views a decision tree as a hierarchy of clusters: the top-node of a PCT corresponds to one cluster (group) containing all data, which is recursively partitioned into smaller clusters while moving down the tree. The leaves represent the clusters at the lowest level of the hierarchy and each leaf is labeled with its clusters prototype (prediction). PCTs can be learned by the system CLUS available at `http://clus.sourceforge.net`.

PCTs are built with a greedy recursive top-down induction algorithm, similar to that of CART (Breiman et al., 1984). The learning algorithm starts by selecting a test for the root node by using a heuristic function computed on the training examples. The goal of the heuristic is to guide the algorithm towards small trees with good predictive performance. Based on the selected test, the training set is partitioned into subsets according to the test outcome. This is recursively repeated to construct the subtrees. The partitioning process stops when a stopping criterion is satisfied. In that case, the prototype (i.e., the prediction) is calculated and stored in a leaf.

The different stopping criteria can be used to deal with noise and other types of imperfection in the data. This is also know as tree pruning. Three types of pruning of PCTs are implemented: *maximal depth*, *minimal instances in a leaf* and *F-test pruning*. The *maximal depth* algorithm takes as input a user defined integer value that specifies the maximally allowed depth for the leaves in the tree. The *minimal instances in a leaf* algorithm also takes as input a user defined integer value that specifies the minimal number of instances that each leaf of the tree must contain. Finally, the *F-test pruning* algorithm checks whether the addition of a split at a given leaf of the tree significantly reduces the intra-cluster variance for the examples in that leaf: The significance level is specified by the user. These pruning algorithms increase the interpretability of PCTs, while maintaining (or increasing) their predictive performance.

The main difference between the algorithm for learning PCTs and a standard decision tree learner is that the former considers the variance function and the prototype function, that computes a prediction for each leaf, as parameters that can be instantiated for a given learning task. For further information, we refer the reader to (Kocev et al., 2013). In this paper, we use PCTs instantiated for the first of these tasks, i.e., PCTs for multi-target classification and PCTs for multi-target regression.

Examples of *PCTs for multi-target classification* are shown in Figures 5, 12, 9 and 11. The variance function for the PCTs for multi-target classification is computed as the sum of the Gini indices of the target variables, i.e., $Var(E) = \sum_{i=1}^{T} Gini(E, Y_i)$. 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.

Examples of *PCTs for multi-target regression* are shown in Figures 7, 8, 10 and 6. The variance and prototype functions for PCTs for multi-target regression 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.

2.2. Ensemble methods

An ensemble is a set of predictive models (called base predictive models). The prediction of an ensemble for a new instance is obtained by combining the predictions of all base predictive models from the ensemble. The predictions from the models can be combined by taking the average (for regression tasks) and the

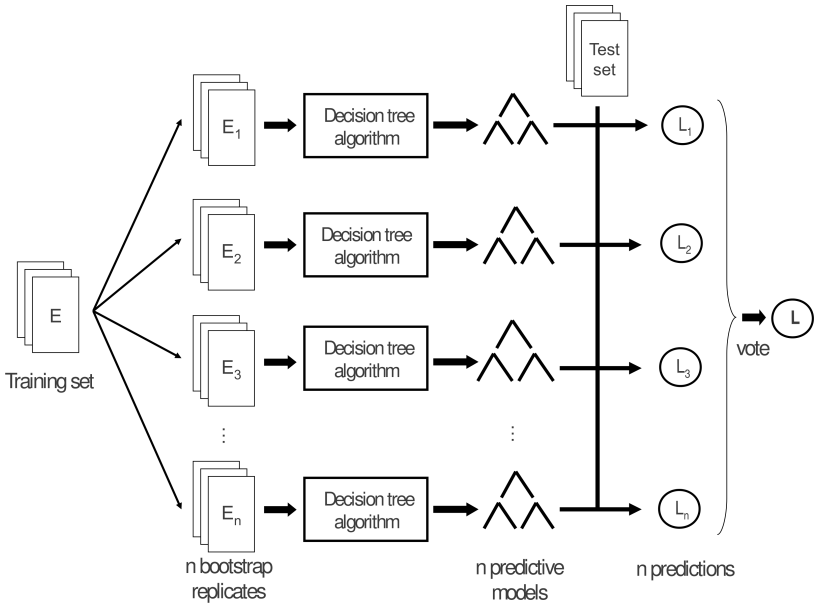


Figure 2: An illustration of the ensemble learning method of bagging. From the training set of examples E, n bootstrap replicates are created (E_1, E_2 ... E_n). Predictive models are then constructed (using a tree construction algorithm) on each of the n replicates. The predictions of the base predictive models (L_1, L_2 ... L_n) are combined by a voting (or averaging) scheme into a final prediction L of the ensemble.

majority or probability distribution vote (for classification tasks) (Bauer and Kohavi, 1999; Breiman, 1996), or by taking more complex aggregation schemes (Kuncheva, 2004).

In this article, we consider ensembles of PCTs for multi-target prediction (Kocev et al., 2007; Kocev, 2011). The PCTs in the ensembles are constructed by using the bagging and random forests methods that are often used in the context of decision trees. We have adapted these methods for use in PCTs. The bagging ensemble learning method is illustrated in Figure 2. For the random forests method, 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. 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 a prediction of the ensemble, we take the averages per target of the predictions of the base predictive models. We obtain the ensemble multi-target predictions for the multi-target classification using probability distribution voting (as suggested in (Bauer and Kohavi, 1999)) per target.

Bagging (Breiman, 1996) 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. 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 (1996) 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 learned model and its predictions), such as classification and regression tree learners.

A *random forest* (Breiman, 2001) is an ensemble of trees, where diversity among the predictors is obtained by both using bootstrap replicates as in bagging, and by dynamically changing the set of descriptive attributes during learning. 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., $f(D) = 1$, $f(D) = \lfloor\sqrt{D}+1\rfloor$, $f(D) = \lfloor log_2(D)+1\rfloor \ldots$) (Breiman, 2001). By setting $f(D) = D$, we obtain the bagging procedure.

3. Experimental results

In this section, we present the results of the use of the proposed methodology for modelling the communities of organisms that live in three complex ecosystems: a lake, a river and the soil. First, we present the data that describe the respective ecosystems. We then give the experimental design that was used to construct the habitat models. Finally, we discuss the predictive performance of the obtained habitat models.

3.1. Data overview

We used datasets from three separate studies that constructed habitat models for the following communities: lake diatoms (Kocev et al., 2010), river water organisms (Blockeel et al., 1999; Džeroski et al., 2000), and soil microarthropods (Demšar et al., 2006). Table 1 summarizes the properties of all seven considered variants of the three datasets. In the remainder, we describe the data concerning these communities.

Table 1: Short summary information on the datasets. For each dataset, the type of the targets, number of targets (T), number of samples (N), and the number of discrete and continuous descriptive attributes (D/C) are given.

Case study	Dataset	Target Type	T	N	D/C
Lake Prespa	DiatomsAll	*continuous*	111	218	0/18
Lake Prespa	DiatomsTop10	*continuous*	10	218	0/18
Lake Prespa	DiatomsAll-nom	*discrete*	111	218	0/18
Lake Prespa	DiatomsTop10-nom	*discrete*	10	218	0/18
Slovene Rivers	WaterQuality	*continuous*	14	1060	0/16
Slovene Rivers	WaterQuality-nom	*discrete*	14	1060	0/16
Danish soil	SoilQuality-nom	*discrete*	39	1944	41/7

3.1.1. Diatom communities from lake Prespa

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 (Krstić, 2005).

A total of 218 water samples from the lake were collected. On these water samples, both physico-chemical and biological analyses were performed. The physico-chemical 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 physico-chemical properties of the water samples were measured: temperature, dissolved oxygen, Secchi depth, conductivity, alkalinity (pH), nitrogen compounds (NO_2, NO_3, NH_4, organic and total nitrogen), sulphur oxide ions SO_4, phosphorus P, sodium (Na), potassium (K), magnesium (Mg), copper (Cu), manganese (Mn) and zinc (Zn).

The biological variables were the relative abundances of 111 different diatom species (given in Table 2). 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 relative abundance of a given species is then the respective percent of the total diatom count per sampling site.

We analyze four variants of the dataset. The first variant includes the relative abundances of all diatom species (multi-target regression, *DiatomsAll*). The second variant contains the presence of all diatom species (multi-target classification, *DiatomsAll-nom*). The third variant includes the relative abundances of the top 10 diatom species (the ones with highest average abundance across all samples, i.e., multi-target regression, *DiatomsTop10*). The fourth variant contains the presence of the top 10 diatom species (multi-target classification, *DiatomsTop10-nom*). More details about the data are given in (Kocev et al., 2010).

Table 2: List of all diatom species encountered in the measurements for Lake Prespa, Macedonia. The species in italics are the top 10 most abundant species across all samples.

Achnanthes lac.	*Cyclotella oc.*	Gomphonema min.	Navicula pro.	Planothidium ros.
Achnanthes sp.	Cymatopleura el.	Gomphonema ol.	Navicula rei.	Pseudostaurosira bre.
Achnanthidium cl.	Cymbella aff.	Gomphonema ol.	*Navicula rot.*	Reimeria sin.
Achnanthidium cl. bal.	Cymbella lan.	Gomphonema par.	Navicula subh.	Rhoicosphenia abb.
Achnanthidium min.	Cymbella neo.	Gomphonema pum.	*Navicula subr.*	Rhopalodia gib.
Amphora aeq.	Diatoma ang.	Gomphonema sar.	Navicula tri.	Sellaphora perb.
Amphora cop.	*Diploneis mau.*	Gomphonema ter.	Navicula vir.	Sellaphora pu.
Amphora fog.	Diploneis mod.	Gyrosigma mac.	Navicula vircl.	Stauroneis gra.
Amphora in.	Diploneis ov.	Hannea ar.	Neidium du.	Stauroneis pho.
Amphora ov.	Encyonema cae.	Hantzschia amp.	Nitzschia alp.	Stauroneis sm.
Amphora ped.	Encyonema min.	Hippodonta ros.	Nitzschia dis.	Staurosira con.
Amphora th.	Encyonema sil.	Martyana mar.	Nitzschia lin.	Staurosira con. bin.
Amphora ven.	Encyonopsis mic.	Meridion cir.	Nitzschia rec.	Staurosira con. ven.
Aulacoseira gra.	Epithemia ad.	Meridion cir. con.	Nitzschia suba.	*Staurosirella pin.*
Caloneis sch.	Epithemia so.	Navicula ant.	Nupela la.	Surirella ang.
Cavinula scu.	Fallacia och.	Navicula cap.	Orthoseira ros.	Surirella min.
Cocconeis dis.	Fragilaria cap.	Navicula cry.	Pinnularia bor.	Tabellaria floc.
Cocconeis neo.	Fragilaria cap. va.	Navicula gre.	Pinnularia subc.	Tryblionella ang.
Cocconeis pl.	Fragilaria par.	Navicula has.	Placoneis bal.	Ulnaria ul.
Cocconeis pl. eug.	Frustulia vul.	Navicula krs.	Placoneis elg.	
Cocconeis pl. li.	Geissleria dec.	Navicula lan.	Placoneis min.	
Cyclotella jur. nud.	Gomphonema cl.	Navicula pra.	Placoneis neo.	
Cyclotella men.	Gomphonema it.	*Navicula pre.*	Planothidium lan.	

3.1.2. Water organisms from Slovenian rivers

The data for this case study come from the Hydro-meteorological Institute of Slovenia (now Environmental Agency of Slovenia) that performs water quality monitoring for Slovenian rivers and maintains a database of water quality samples. The data provided cover a six year period of monitoring, starting from 1990 until 1995. Biological samples were taken twice a year, once in summer and once in winter, while physical and chemical samples were taken several times a year for each sampling site.

In total, there are 1061 samples, each described with its physico-chemical and biological properties. The physico-chemical properties of the samples include the measured values of 16 parameters: biological oxygen demand (BOD), electrical conductivity, chemical oxygen demand ($K_2Cr_2O_7$ and $KMnO_4$), concentrations of Cl, CO_2, NH_4, PO_4, SiO_2, NO_2, NO_3 and dissolved oxygen (O_2), alkalinity (pH), oxygen saturation, water temperature, and total hardness. The considered organisms include 14 species, 7 animal and 7 plant species, selected as strong bioindicators of water quality. The animal species include *Erpobdella octoculata*, *Gammarus fossarum*, *Baetis rhodani*, *Hydropsyche species*, *Rhyacophila species*, *Simulium species* and *Tubifex species*, while the plant species include *Cladophora species*, *Gongrosira incrustans*, *Oedogonium species*, *Stigeoclonium tenue*, *Melosira varians*, *Nitzschia palea* and *Audouinella chalybea*.

Each sample was examined and the density/frequency of the occurrence of each species was recorded by an expert biologist. In particular, this frequency is recorded at three different qualitative levels, where 1 means that the species occurs incidentally, 3 frequently, and 5 abundantly (multi-target regression, *WaterQuality*). Furthermore, we create one more variant of the dataset, where the presence of each species is recorded (1 if there are some organisms of a given species present in a given sample and 0 otherwise, i.e., multi-target classification, *WaterQuality-nom*). More details about the data are given in (Blockeel et al., 1999) and (Džeroski et al., 2000).

3.1.3. Soil microarthropods from Danish farms

The data used in this study describes four experimental farming systems (observed during the period 1989-1993) and a number of organic farms in Denmark (observed during the period 2002-2003). Soil samples

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

The data concerns the *Collembola* species community in the soil samples. These species can be used as indicators of the soil quality (in particular soil dessication) and some are considered as pests for the plants. Also, they are one of the main biological factors responsible for the control of the soil microorganisms (Ponge, 1991).

Table 3: List of *Collembola* species in the soil samples from Denmark.

Anurida pyg.	Friesea mir.	Isotomodes bis.	Neelus min.	Sminthurinus el.
Brachystomelle par.	Heteromurus nit.	Isotomodes prod.	Orchesella cin.	Sminthurus vir.
Ceratophysella den.	Hypogastrua sp.	Isotomurus pal.	Orchesella vil.	Stenaphorura quad.
Ceratophysella suc.	Isotoma ang.	Isotomurus sp.	Protaphorura sp.	Tomocerus fl.
Entomobrya sp.	Isotoma not.	Lepidocyrtus cy.	Pseudosinella al.	Tomocerus min.
Folsomia fim.	Isotoma tig.	Lepidocyrtus lan.	Pseudosinella sex.	Tomocerus sp.
Folsomia quad.	Isotomiella min.	Mesaphorura sp.	Smint sp.	Willemia sp.
Folsomia spi.	Isotomodes arm.	Neanura fam.	Sminthurinus au.	

The dataset contains 1944 samples, where each sample is described by 48 field properties and the presence of 39 *Collembola* species (given in Table 3). The field properties mainly include agricultural measures, such as crops planted, packing, tillage, fertilizer and pesticide use, etc. For several of the properties, a history of 3 years is also recorded, i.e., there is an attribute for the year in which the sample was collected, one for the past year, and one for 2 and 3 years ago. The data about the *Collembola* species contain information whether a species is present or absent in a sample (multi-target classification, *SoilQuality-nom*). More details about the data are given in (Demšar et al., 2006).

3.2. Experimental design

We have constructed two types of habitat models: multi-target and single-target models. The multi-target models are valid for the complete community, while the single-target models are valid for a single species from the community. We then analyzed the data along the lines of two scenarios.

The first (knowledge discovery) scenario focuses on the knowledge discovery process where the goal is to obtain better understanding about the conditions that govern the presence and abundance of the species in a given community. To this end, we constructed multi-target PCTs and single-target DTs. The second (predictive power) scenario is more focused on obtaining state-of-the-art predictive performance and thus obtaining models that are more reliable. In this context, we constructed ensembles of multi-target PCTs and ensembles of single-target DTs.

In the following, we first give the parameter instantiations used by the algorithms for the two scenarios. We then give a brief summary information on the datasets used in this study and present the evaluation measures used to assess the predictive performance and the efficiency of the models. Finally, we review the procedure for statistical evaluation of the experimental results.

3.2.1. Parameter instantiation

In the 'knowledge discovery' scenario, the trees were pruned using three pruning algorithms: *maximal depth*, *minimal instances in a leaf* and *F-test pruning*. In this study, we set the value for maximal depth to 3 (this means that trees with up to 15 nodes can be constructed), the trees have a minimum of 16 instances in each leaf, and the splits in the trees reduce the intra-cluster variance statistically significantly at the 0.05 level. We would like to note that we did not employ F-test pruning on the *DiatomsAll* and *DiatomsAll-nom* datasets, because they have a large number of species and a small number of samples. The usage of this type of pruning for these two datasets showed to be very restrictive, yielding trees that were inappropriately small for the problems at hand.

In the 'predictive power' scenario, we constructed ensembles by using the two approaches of bagging and random forests. Ensembles of multi-target PCTs were constructed for each dataset and were valid for the complete community. Ensembles of DTs were constructed for each dataset and for each species separately.

In both cases, the base predictive models (PCTs and DTs) were unpruned (as suggested in (Bauer and Kohavi, 1999)). The ensembles consist of 100 base predictive models. The random forests require also that the user specifies the size of the attribute subset that is considered at each node in the tree-construction. In this study, we set this value to $f(D) = \lfloor log_2(D) + 1 \rfloor$, where D is the number of attributes, as suggested in (Breiman, 2001).

3.2.2. Predictive performance measures

Empirical evaluation is the most widely used approach for assessing the performance of machine learning algorithms. The performance of a machine learning algorithm is assessed using some evaluation measure. We first describe the evaluation measures for multi-target regression, then for multi-target classification. At the end, we present the measures used to assess the efficiency and the complexity of the proposed methods.

For the task of predicting multiple continuous targets (regression), we employed three well known measures: the correlation coefficient (CC), root mean squared error ($RMSE$) and relative root mean squared error ($RRMSE$). For each of these measures, we performed tests of statistical significance. We present here only the results in terms of $RRMSE$, but the same conclusions hold for the other two measures (for which results are given in the Supplementary information).

We used five evaluation measures for classification: accuracy, precision, recall, F-score, the Matthews correlation coefficient and balanced accuracy (also known as Area Under the Curve) (Sokolova and Lapalme, 2009). Since we are interested in correctly predicting both the presence and the absence of a given species, we aggregate the values from the confusion matrix. We use two averaging approaches to adapt these measures for multi-class problems: micro and macro averaging. We give more details on these evaluation measure in the Supplementary information.

Finally, we compare the algorithms by measuring their efficiency in terms of time consumption and size of the models. We measure the processor time needed to construct the models: in the case of predicting the individual species separately, we sum the times needed to construct the separate models. In a similar way, we calculate the sizes of the models as the total number of nodes (internal nodes and leafs) of the single multi-target tree and all single-target trees taken together.

3.2.3. Statistical evaluation

We estimated the predictive performance of the methods by using 10-fold cross-validation. This evaluation procedure starts by splitting the complete dataset into 10 disjoint parts. Next, a predictive model is constructed using nine of these parts of the dataset and the performance of this model is tested on the tenth part. This is repeated 10 times. The performance on unseen cases of the model learned from the entire dataset is then estimated by aggregating the performance results from the 10 runs. In addition to the testing performance on unseen cases (estimated by 10-fold cross-validation), we also present the performance of the methods on the complete dataset to assess the descriptive power of the methods. We then compare the performance measures for different methods by using tests for statistical significance.

We adopt the recommendations by Demšar (2006) for the statistical evaluation of the results. We use the non-parametric Friedman test (Friedman, 1940) for statistical significance with the correction from Iman and Davenport (1980). Afterwards, to check where the statistically significant differences appear (between which methods), we use the Nemenyi post-hoc test (Nemenyi, 1963). We present the result from the Nemenyi post hoc test with an average ranks diagram as suggested by Demšar (2006). Example average ranks diagrams can be seen in Figures 3 and 4. The ranks are depicted on the axis, in such a manner that the best ranking algorithms are at the right-most side of the diagram. The algorithms that do not differ significantly (in performance) are connected with a line.

We have performed two statistical analyses. The first considers all target variables from all datasets together for the assessment of the statistical significance: The results are shown in Figures 3 and 4, for the $RRMSE$ and micro averaged balanced accuracy, respectively. For the remaining evaluation measures the results are given in the Supplementary information. The second statistical analysis considers the target

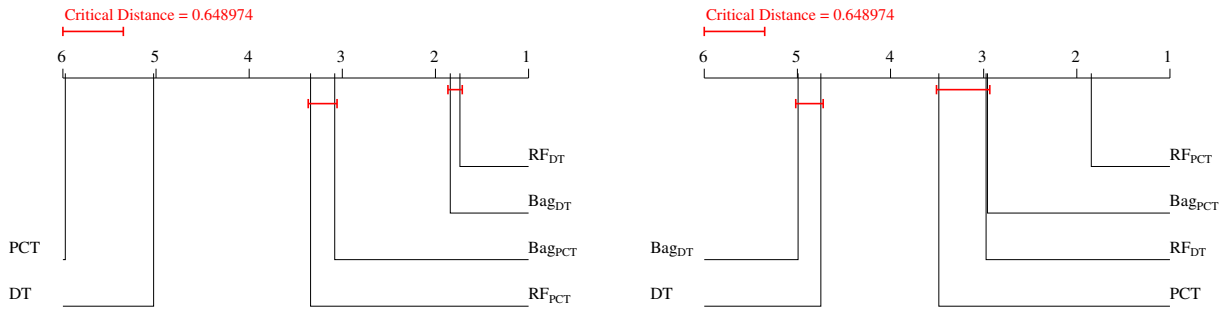
variables for each dataset separately. The results of this analysis for all evaluation measures are given in the Supplementary information.

3.3. Performance of the proposed habitat modelling methodology

In this section, we show the results for both predictive and descriptive performance of the obtained models. We first present the results from the modelling of the species abundance (multi-target regression task). We then discuss the results from the modelling of the species presence/absence (multi-target classification). Finally, we summarize the results for the performance and efficiency of the proposed methods.

3.3.1. Modelling the abundance of species

Figure 3 shows the average rank diagrams for the methods used for modelling of the species abundance (i.e., multi-target regression) as evaluated by using the RRMSE measure. The average rank diagram for the evaluation on the training set (descriptive power) shows that the single-target models achieve statistically significantly better average ranks (i.e., better performance) than the multi-target models. However, when the evaluation is performed on unseen data by using 10-fold cross-validation (predictive power) the multi-target models have statistically significantly better average ranks. This means that the single-target models tend to overfit the data at hand and do not capture the more general dependencies that may exist between the observed environmental variables and the organisms. In other words, the single-target models capture specific knowledge from the data, while the multi-target models capture more general relations present in the observed ecosystems. Similar conclusions can be made if we look at the results for the RMSE and correlation coefficient measures, which are given in the Supplementary information.



(a) Train (b) Cross-validation

Figure 3: The average rank diagrams for the RRMSE evaluation measure on the datasets that contain information about the species abundance. The 'Train' performance is obtained from the complete training set, while the 'Cross-validation' gives the performance assessed using 10-fold cross-validation.

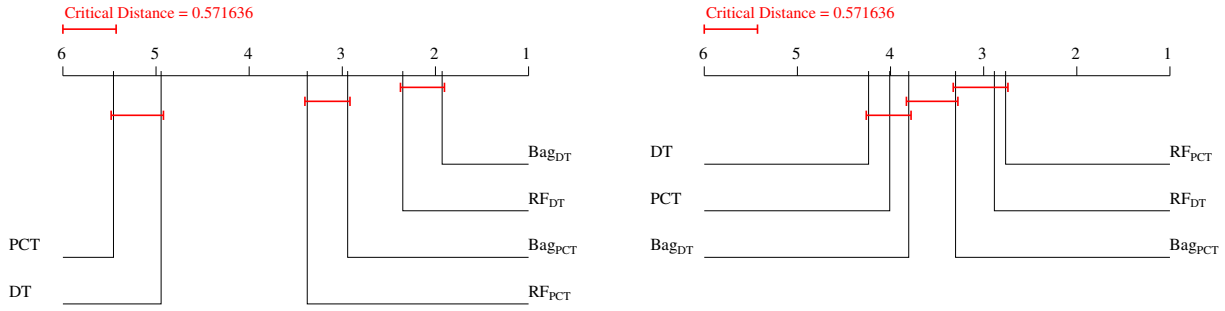
Table 4: Efficiency of the proposed methods for modelling the abundance of the species. The *time* needed to construct the models is given in seconds. For the single-target models, the time shown is the total time needed to construct the models for all targets. The *size* of the habitat models is calculated as follows. For the multi-target PCTs, it is the total number of nodes (internal nodes and leaves). For the single-target DTs, the size of the models is the sum of all nodes in the trees for all targets. The model sizes for the ensembles is the total number of nodes in all models in the ensemble.

		PCT	DT	Bag_{PCT}	Bag_{DT}	RF_{PCT}	RF_{DT}
Time	DiatomsAll	0.06	0.42	14.89	43.59	3.97	17.43
	DiatomsTop10	0.01	0.07	2.35	7.22	1.06	4.08
	WaterQuality	0.05	0.25	12.86	41.15	4.64	23.65
Size	DiatomsAll	13	971	15402	420616	15450	476804
	DiatomsTop10	11	118	15120	107052	15448	117110
	WaterQuality	9	150	82842	451822	83036	524486

Next, let us compare the multi-target and single-target models by their efficiency as measured by the time needed to construct the models and the size of the obtained models. The results are given in Table 4. To begin with, the (multi-target) PCTs are $\sim$ 6.3 times faster to construct and have $\sim$ 34 times smaller models than the (single-target) decision trees. Next, the multi-target ensembles are more efficient than the single-target ensembles. Namely, multi-target bagging is $\sim$ 3.1 times faster and has $\sim$ 13.3 times smaller models than single-target bagging. Similarly, multi-target random forests are $\sim$ 4.4 times faster and have $\sim$ 14.9 times smaller models than single-target random forests. All in all, the multi-target models are far superior to the single-target models when compared by their efficiency. Moreover, if we consider the interpretability, it is much easier to interpret a single model with 9 nodes, for example, than 14 models with total of 150 nodes (as learned for the *WaterQuality* dataset).

3.3.2. Modelling the presence/absence of species

We evaluated the predictive performance of the methods used for modelling of the species presence/absence (i.e., multi-target classification) by using the measures mentioned in Section 3.2.2. For simplicity, in Figure 4, we present the average ranks for the methods used for modelling of the species presence/absence as evaluated by their micro balanced accuracy, since this measure is able to avoid inflated performance estimates in datasets with imbalanced classes (Brodersen et al., 2010). The balanced accuracy is actually the mean of sensitivity and specificity. We provide the evaluation using the other measures in the Supplementary information of this manuscript. For the modelling of the presence/absence of species, the average ranks show that the single-target models achieve better performance when evaluated on the training set (descriptive power), while the multi-target models have better performance when evaluated by 10-fold cross-validation (predictive power). This means that when it comes to capturing more general knowledge and relationships that may exist between the environmental conditions and the structure of a given community of species, the multi-target models should be preferred.



(a) Train (b) Cross-validation

Figure 4: The average rank diagrams for the micro balanced accuracy evaluation measure on the datasets that contain information about the species presence/absence. The 'Train' performance is obtained from the complete training set, while the 'Cross-validation' gives the performance assessed using 10-fold cross-validation.

We further compare the multi-target and single-target models by their efficiency as measured by the time needed to construct the models and the size of the obtained models. The results are given in Table 5. First, the PCTs are $\sim$ 5.9 times faster to construct and have $\sim$ 18.5 times smaller models than the decision trees. Next, the multi-target ensembles are more efficient than the single-target ensembles. Namely, multi-target bagging is $\sim$ 2 times faster and has $\sim$ 7.1 times smaller models than single-target bagging. Similarly, multi-target random forests are $\sim$ 3.6 times faster and have $\sim$ 8.6 times smaller models than single-target random forests. In sum, the multi-target models are much more efficient than the single-target models.

3.3.3. Summary of the results

A quick inspection of the results shows that the prediction problem is very difficult: the predictive performance is low. We first constructed the PCTs (multi-target models) and DTs (single-target models) and

Table 5: Efficiency of the proposed methods for modelling the presence/absence of the species. The *time* needed to construct the models is given in seconds. For the single-target models, the time shown is the total time needed to construct the models for all targets. The *size* of the habitat models is calculated as follows. For the multi-target PCTs, it is the total number of nodes (internal nodes and leaves). For the single-target DTs, the size of the models is the sum of all nodes in the trees for all targets. The model sizes for the ensembles is the total number of nodes in all models in the ensemble.

		PCT	DT	Bag_{PCT}	Bag_{DT}	RF_{PCT}	RF_{DT}
Time	DiatomsAll-nom	0.07	0.40	23.46	38.57	4.66	15.41
	DiatomsTop10-nom	0.02	0.09	2.87	5.77	1.06	3.11
	SoilQuality-nom	0.10	0.76	46.09	80.55	4.07	19.73
	WaterQuality-nom	0.04	0.25	14.03	38.88	5.43	19.03
Size	DiatomsAll-nom	13	371	15206	261988	15372	305552
	DiatomsTop10-nom	7	76	13450	37430	14064	45578
	SoilQuality-nom	11	239	51640	249090	35196	252934
	WaterQuality-nom	9	116	81244	288080	81802	335656

assessed their predictive performance. The poor predictive performance motivated us to further investigate the limits of predictive performance on the data at hand. To this end, we employed ensembles (bagging and random forests) of both PCTs (Kocev et al., 2013) and decision trees (Breiman, 1996, 2001). It is well known that ensemble methods perform better than individual trees and are amongst the top performing methods for predictive modelling Breiman (2001). The results in our study show that the ensembles do lift the predictive performance, both for PCTs and DTs as base models.

However, the predictive performance of the ensembles, as estimated by 10-fold cross-validation, is still quite low (i.e., the correlation coefficients are all smaller than 0.5). This leads us to the conclusion that the low predictive performance of the PCTs and DTs is not due to the modelling algorithm, but to the limited data quality and quantity. The datasets at hand aim at describing very complex eco-systems with a limited number of attributes and samples. These datasets are small in terms of the number of samples and the number of environmental conditions considered. Thus, they do not consider important factors that surely influence the outcome. Increase of the predictive performance of the models can be expected once measurements of a wider range of attributes are performed at additional sampling locations. We confirmed this also by estimating the variable importance using the random forests mechanism (Breiman, 2001). We used the extension of the method for estimating variable importance for the task of multi-target prediction, proposed by Kocev (2011). The results (presented in the Supplementary information) show that the values for the variable importances are quite small and that the values for all of the variables are close to each other. We hypothesize that this is due to fact that a wider range of attributes is needed to describe these complex eco-systems.

A comparison of the performance of single-target DTs and multi-target PCTs shows that single-target DTs perform better on the training data. But, on unseen data, the PCTs have better performance (as assessed by cross-validation). This means that the single-target DTs tend to over-fit the training data. This can be observed by a quick inspection of the trees constructed for each fold from the 10-fold cross-validation. Namely, the structure and the variables in the nodes of the multi-target PCTs across the ten folds are quite similar to each other and to the tree constructed on the whole training set. On the other hand, the structure and the variables in the nodes of the single-target DTs vary more across the ten folds. This is because the multi-target PCTs are able to capture the more general knowledge that is present in the data, while the single-target DTs capture knowledge specific for the individual species. The over-fitting issue is also reflected in the sizes of the models: the multi-target models describe the eco-system with trees of 7 to 13 nodes, while the single-target models need 76 to 971 nodes. Moreover, for the domains *DiatomsTop10*, *DiatomsTop10-nom* and *Water quality*, the multi-target PCTs are smaller than the average size of the single-target DTs. For example, the multi-target PCT for *DiatomsTop10* has 11 nodes, while the 10 single-target DTs have 118 nodes in total or 11.8 nodes on average. This shows that the multi-target PCTs are much easier to interpret than the multiple single-target DTs. Furthermore, the multi-target PCTs consider the interactions between the different target species, while the single-target DTs are focused on a single species, thus making overall

conclusions for the complete community structure is difficult.

The multi-target PCTs and single-target DTs are a reasonable choice for habitat models. They are very easy to interpret, with reasonable size and predictive performance. Moreover, the multi-target PCTs capture the interactions that might exist between the different species and try to explain them. Also, they offer overall predictions for the structure of the complete community that is being modelled.

The ensembles trade-off interpretability for predictive performance and thus offer state-of-the-art predictive performance. The multi-target ensembles have slightly better predictive performance than the single-target ensembles on the datasets at hand, while being much more efficient. All in all, multi-target PCTs and single-target DTs can be used if the goal is to discover some novel knowledge about the underlying eco-system. The ensembles, on the other hand, can be used when high predictive performance is required.

4. Discussion of the habitat models

In this section, we present and discuss the habitat models obtained from the three case studies: modelling the diatom community in Lake Prespa, Macedonia, modelling plant and animal organisms in rivers in Slovenia, and modelling *Collembola* species in soil samples from Denmark.

4.1. Habitat models for diatom communities

We constructed two types of habitat models for this case study. The first type of habitat models are the regression models, which aim at predicting the relative abundances of the species. The second type of habitat models are the classification models, which predict the presence of the species in a given sample. In other words, the latter type of habitat models will answer the question which species are present in the given sample, while the former type of habitat models answers how abundant (relatively) the species is in the given sample. Furthermore, we construct habitat models that predict the complete diatom community (single habitat model for all diatom species) and habitat models that are valid just for a single species.

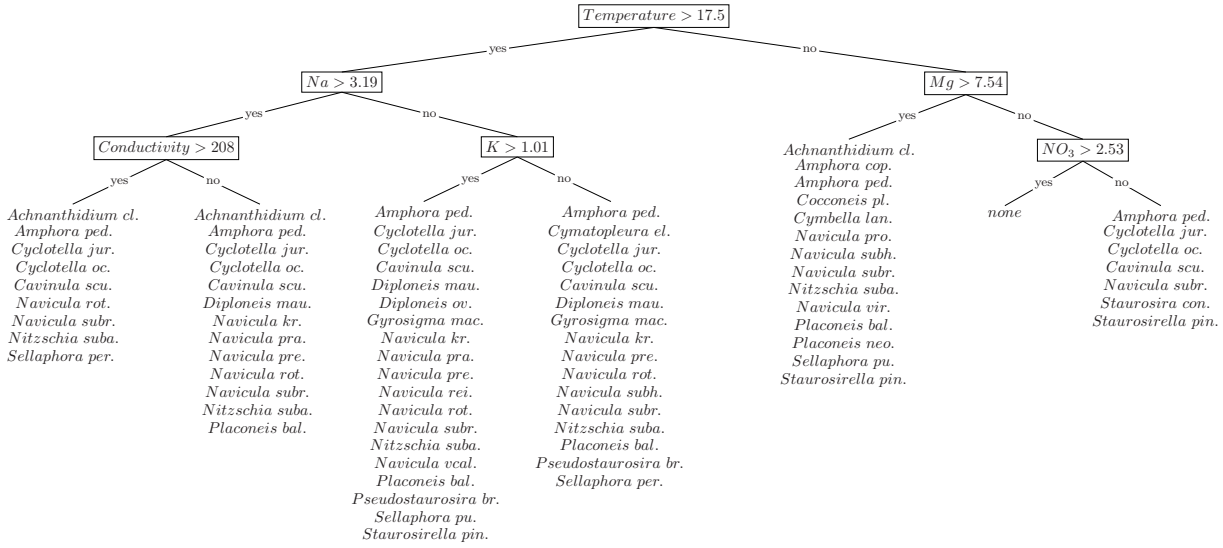


Figure 5: The habitat model for all diatoms with discrete target variables. A complete list of all diatom species in the data is given in Table 2.

We begin by discussing the habitat model for the complete diatom community (Figure 5) that predicts the presence of all species. Then, we present the habitat model for the relative abundances of the top 10 species (Figure 7). Finally, we give the habitat models for each (of the top 10 populated) diatom species separately (Figure 8). We also present the habitat models for the relative abundances of all diatom species and the presence of the top 10 species in the Appendix (Figures 6 and 9, respectively).

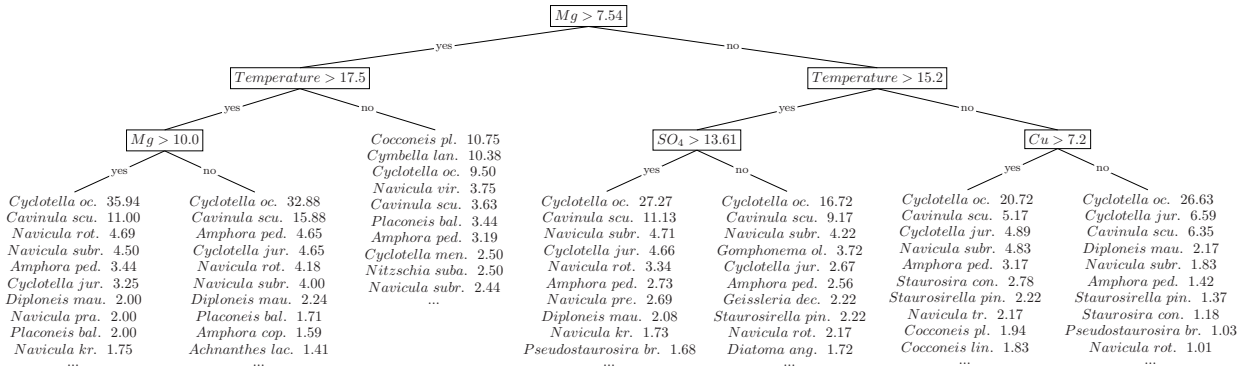


Figure 6: The habitat model for all diatoms with continuous target variables.

The habitat model for the complete diatom community is shown in Figure 5. This model describes the environmental conditions that influence and shape the diatom communities encountered in Lake Prespa. More specifically, it describes the conditions for encountering 6 types of diatom communities and the environmental conditions that limit or favor the presence of all diatom species examined here. We can note that most important factors for the diatom community structure are the temperature and the metals. Higher temperatures result in more diverse diatom communities consisting of a larger number of different diatom species. The metallic ions (potassium, magnesium and sodium) are a crucial part of the enzymes that play an important role in the life of diatoms. Similar findings have been found by Gold et al. (2002): metallic pollution (in particular, cadmium and zinc) affected and changed the diatom community.

The most diverse diatom community (with most diatom species) is encountered at higher temperatures, lower concentrations of sodium and higher concentrations of potassium (the third leaf in the tree, counting from left to right). The most harsh conditions that are unsuitable for most diatom species are low temperatures, low concentrations of magnesium and high concentrations of nitrates (the sixth leaf in the tree, counting from left to right). The same factors also influence the relative abundances of all diatom species (see Figure 6).

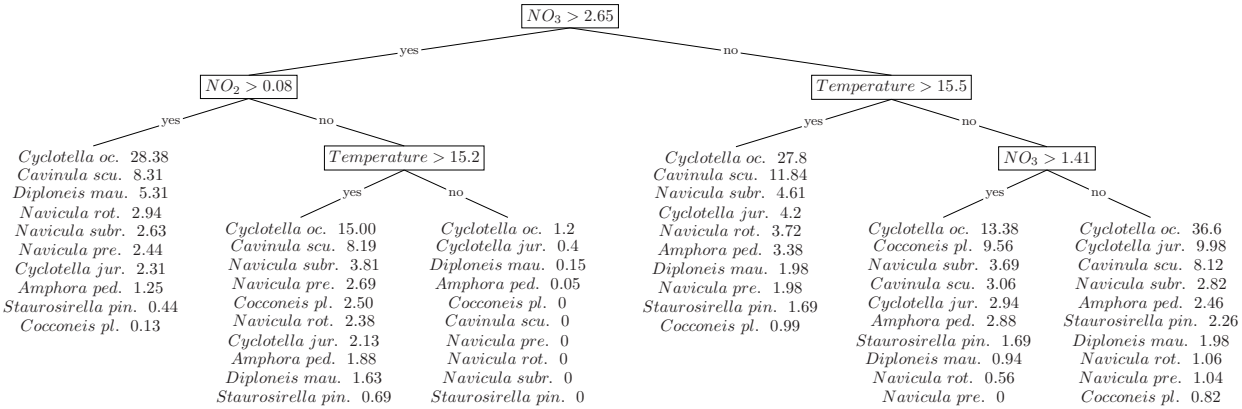


Figure 7: The habitat model for the top 10 diatoms with continuous target variables.

Figure 7 depicts the habitat model for the top 10 most abundant diatoms in the lake samples. It defines the environmental conditions that favor or limit the abundance of the most abundant diatom species, i.e., explains the relations that dominant diatoms have with the physico-chemical properties according to their trophic preference. This model identifies the temperature and the nutrients (presented through the concentrations of nitrates and nitrites) as the most influential factors. Higher temperatures lead to larger

abundances of the diatoms: a comparison of the diatom abundances in the second and the third leaf of the tree reveals that lower temperatures (accompanied with a low concentration of nitrites and a high concentration of nitrates) lead to reduction of the abundance of the top 10 diatom species. Nitrates, as one of the basic external nutrients for the diatoms and a chemical indicator of higher trophic levels, influence positively the relative abundances of the diatoms given the higher temperatures (see the second, third and fourth leaf of the tree).

Note the difference in the most important factors for predicting the structure of the entire community (Figures 5 and 6) versus the top 10 diatom species (Figures 9 and 7). For the top 10 diatom species, the nutrients (NO_3 concentrations) are most important. For the structure of the entire community, however, metal ions play a key role. While the nutrients are components of proteins, metals such as potassium and magnesium are parts of (co)-enzymes that also play an important role in cellular processes.

We further discuss the habitat models for the two most abundant diatom species (shown in Figure 8): *Cyclotella ocellata* and *Cavinula scutetelloides*. The most abundant diatom species, *Cyclotella ocellata*, is mostly influenced by the nitrogen compounds (NO_3 and NH_4), the temperature and the conductivity of the water and the potassium concentration. The concentration of nitrates (i.e., nutrients) influences positively the abundance: the leafs of the tree with higher nitrate concentration (the first three leafs) contain samples with higher abundances of *Cyclotella ocellata* than the leafs on the right-hand side of the tree.

The habitat model for *Cavinula scutetelloides* shows that the temperature, nitrates and metal concentrations are most influential for this diatom species. Higher temperatures favor this specific diatom species, while optimal concentration of magnesium (between 6.13 and 9.44 $\mu g/dm^3$) are needed for the highest abundance. This habitat model also reveals the limiting role of manganese and copper for this species: the higher values of these metals at lower temperatures reduce the abundance of the species.

The habitat models per species shown in Figure 8 do offer interesting insights about the specific species. However, it is very difficult to draw conclusions that concern the diatom community as a whole. This problem is even more emphasized if the community consists of many different species, as is the case here. Namely, constructing a habitat model for each of the 111 species from this case study, and, moreover, interpreting these models is a very labor intensive task. Furthermore, combining all the 111 separate habitat models into general conclusions concerning the complete community is not feasible. On the other hand, the multi-target PCTs offer general conclusions about the relations between the environmental factors and community structure by using a single model. Hence, they provide unified knowledge about the complete diatom community.

4.2. Habitat models for river organisms

We constructed two habitat models for the community structure, i.e., the 7 plants and 7 animals in water samples from Slovenian rivers. The first concerns the relative abundances of the 14 species (regression model with multiple continuous target variables shown in Figure 10), while the second concerns the presence of the species (classification model with multiple discrete target variables shown in Figure 11).

In this section, we focus on the habitat model that predicts the (relative) abundances of the 14 species (Figure 10). This model identifies the biological oxygen demand, the concentration of carbon dioxide CO_2 and the concentration of chlorine Cl as the most important environmental factors influencing the structure of the community. The higher values for BOD (higher than 0.66) favors the plant species. Namely, the plant species *Nitzschia palea* is most abundant when the content of BOD is highest (the first and the second leaf from the tree), while it is encountered rarely at the lower content of BOD (the last three leaves of the tree). The lowest abundance of all species (the right-most leaf of the tree) is encountered with low content of BOD, absence of CO_2 and low chlorine concentrations (smaller than 0.20 mg/l).

A brief inspection of the habitat model for the presence of the 14 species (Figure 11) shows that the most important factors influencing the structure of the community are the concentrations of nitrites NO_3, i.e., nutrients, potassium dichromate ($K_2Cr_2O_7$), which measures the chemical oxygen demand, chlorine (Cl) and carbon dioxide (CO_2). The habitat model for the presence of the species, however, does not identify BOD content as an important factor for the community structure. The cleanest waters (rightmost leaf), which lack nutrients, support *Rhyacophila species*, an indicator of unpolluted waters. The most polluted

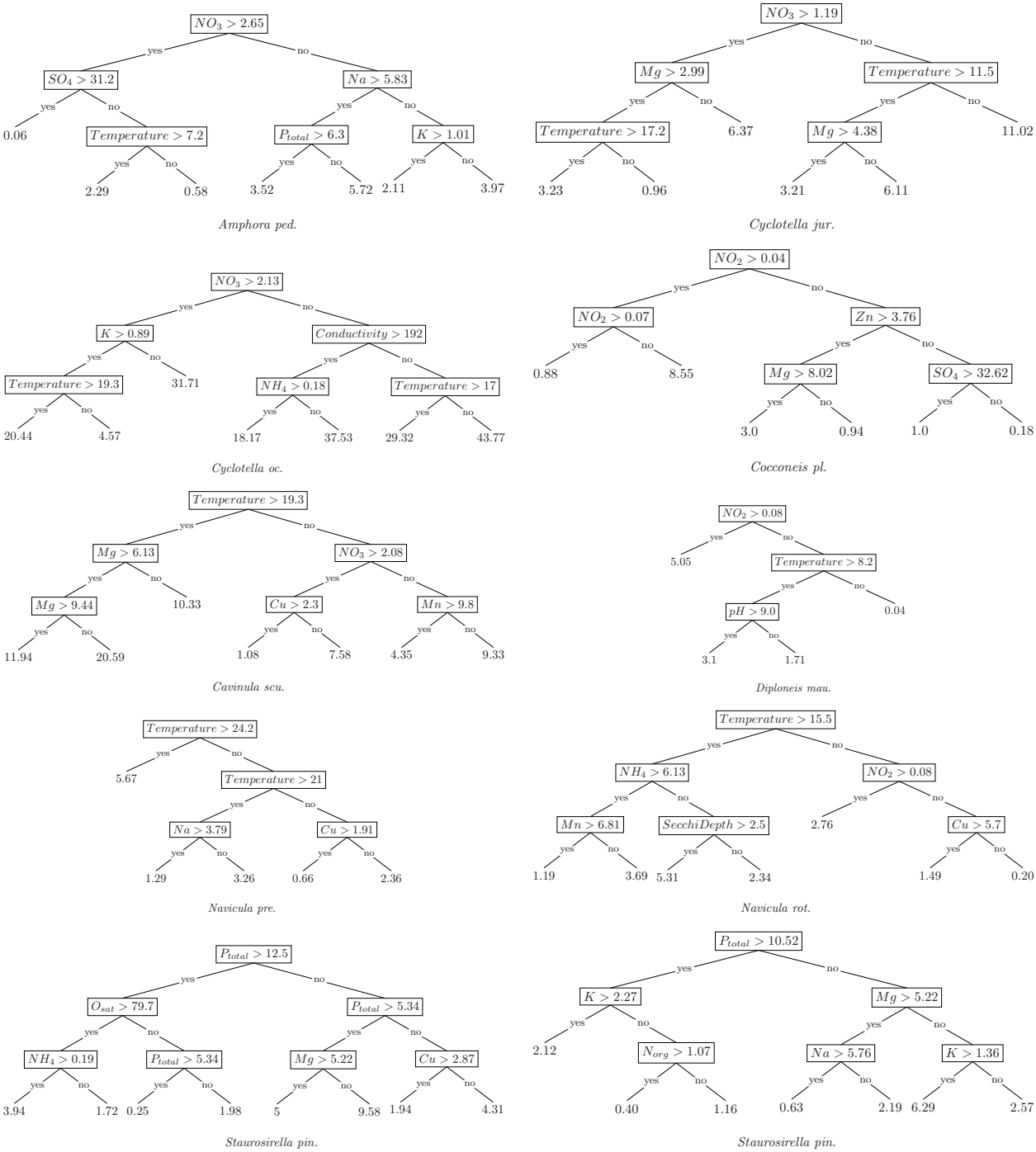


Figure 8: Individual habitat models for each of the top 10 diatom species in Lake Prespa.

waters (leftmost leaf) allow only for the most pollution-resistant organisms (*Nitzschia palea* and *Tubifex species*). Moderate clean/polluted waters support larger, more diverse communities.

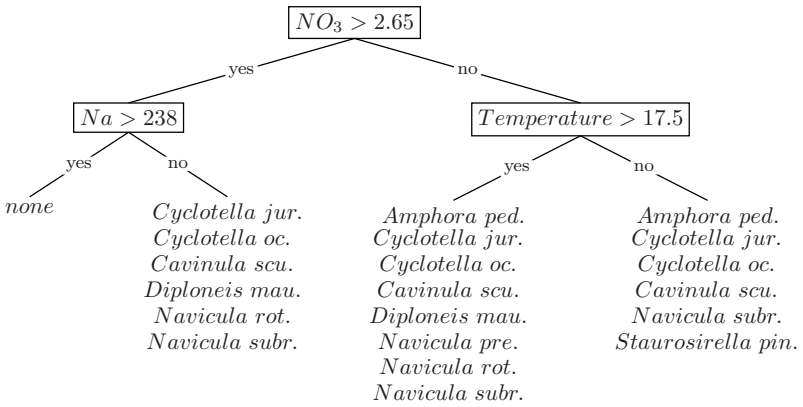


Figure 9: The habitat model for top 10 diatoms with discrete target variables.

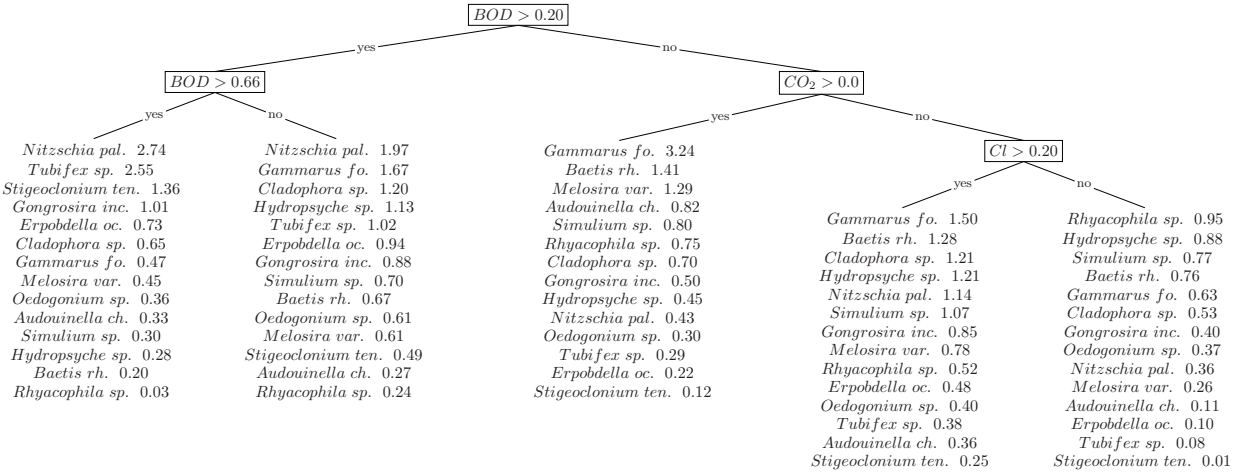


Figure 10: The habitat model for communities in Slovenian rivers.

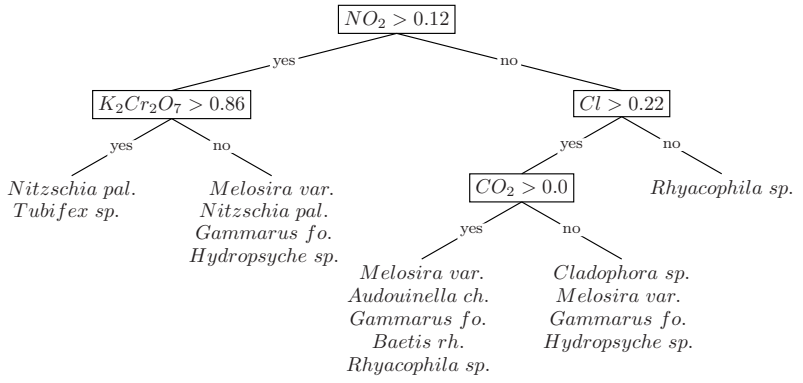


Figure 11: The habitat model for communities in Slovenian rivers described with discrete target variables.

4.3. Habitat models for soil microarthropods

We constructed a habitat model for all 39 *Collembola* species that predicts the presence of a given group of species under given environmental conditions. The habitat model is depicted in Figure 12. The habitat

model of the *Collembola* community identifies the type of the soil as the most important environmental factor. Namely, it differentiates the communities that develop in clay (soils with JB index between 5 and 9) and other types of soil.

Other than the soil type, the type of crop, the one currently on the field and the crop types from the previous years are most important. This means that certain types of crop contribute to the increase of the population of some *Collembola* species and other types of crop to the decrease of the population. In particular, if the soil type is clay and the crop three years ago (the first and second leaf in the tree) was winter barley, clover-grass, fallow, grass, spring barley or winter wheat, then *Collembola* species *Isotoma notabilis*, *Mesaphorura species* and *Smint species* are present. Otherwise, there is a much larger community of *Collembola* species shown in the second leaf of the tree.

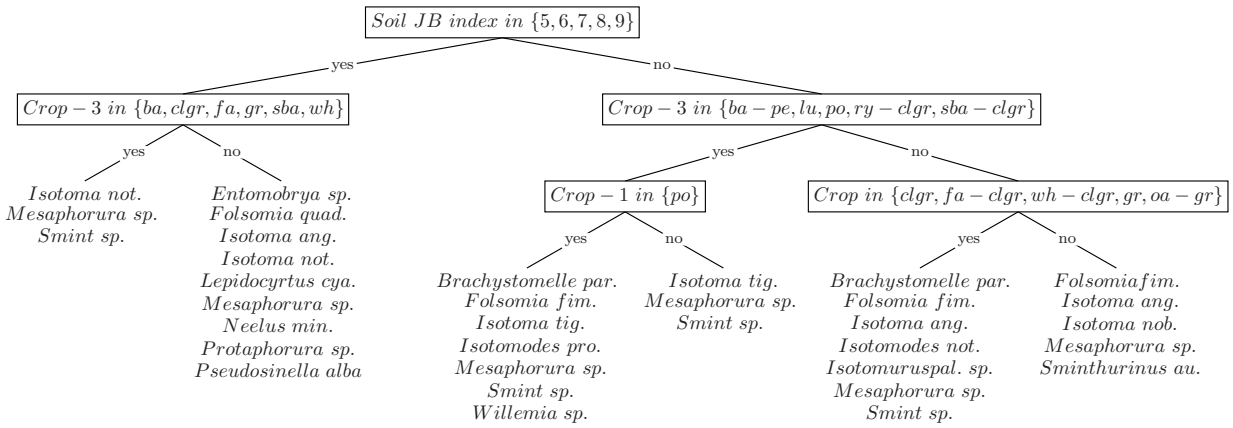


Figure 12: The habitat model for soil - modelling *Collembola* species (discrete target variables). A complete list of all *Collembola* species used in this case study is given in Table 3, while a list of abbreviations and crop types is given in Table 6.

Table 6: List of abbreviations for the crop types of the soil samples from Denmark.

Abbr.	Crop	Abbr.	Crop	Abbr.	Crop
ba	winter barley	fa-gr	fallow, grass	ry-clgr	rye, clover, grass
ba-ch	winter barley, chicory	gr	grass	sba	spring barley
ba-clgr	winter barley, clover, grass	le	leeks	sba-clgr	spring barely, clover, grass
ba-gr	winter barley, grass	lu	lupin	sba-gr	spring barely, grass
ba-pe	winter barley, peas	lu-gr	lupin, grass	swh	spring wheat
be	beets/carrots	oa	oates	tc	triticale
cc	catch crop	oa-clgr	oates, clover, grass	wc	whole crop
ch	chicory	oa-gr	oates, grass	wc-gr	whole crop, grass
chgr	chicory, grass	pe	Peas	wh	winter wheat
clgr	clover, grass	po	Potatoes	wh-chgr	winter wheat, clover, grass
clgr-wc	clover, grass, wholecrop	ra	rape	wh-gr	winter wheat, grass
fa	fallow	rd	radish		
fa-clgr	fallow, clover, grass	ry	rye		

4.4. Improvements over the analyses from the original studies

We improve upon the studies where the data and initial results from their analyses were first published along several directions. In the first study of the diatom communities, Kocev et al. (2010) construct models for the diatom abundances (multi-target regression), while here we additionally construct models for the presence/absence data (i.e., the community structure, multi-target classification). Moreover, in the current study, we apply additional pruning algorithms that yield more stable trees. Next, in the previous studies of the river communities, Džeroski et al. (2000) and Blockeel et al. (1999) use the abundance information to

infer the values of the chemical parameters of the water, i.e., the presence/absence and the abundance of the species was used as an indicator of the water quality. Here we learn predictive models in the other direction, i.e., we construct habitat models for the communities of river organisms (both multi-target regression and multi-target classification). Finally, in the first study of the communities of soil microarthropods, Demšar et al. (2006) construct models for the overall abundance of the species and the biodiversity of the community. Here we focus on predicting the presence/absence of all species and model the community structure itself (multi-target classification).

5. Conclusions

In this paper, we presented a novel type of habitat models that are based on predictive clustering trees. In particular, we show how trees that are able to predict multiple target variables simultaneously can be successfully used for modelling community structure. Different groups of species that live in different ecosystems can be modelled in this fashion.

While the constructed habitat models do not exhibit as high predictive performance as one would expect, this is not a limitation of the methodology itself (since it is known that ensembles of predictive clustering trees achieve state-of-the-art predictive performance), but rather a consequence of the difficulty of the problem addressed. The modelling problem at hand is very difficult, because the ecosystems under consideration are complex and the data available are of limited quantity and quality. For obtaining models with better predictive performance, more measurements are necessary. These measurements should include additional sampling locations, a longer period of observation, and a wider range of measured environmental variables.

To model community structure one would traditionally construct a separate habitat model for each species. This approach results in many single-target habitat models that are related to the same ecosystem. Although these models do offer insight into the environmental conditions influencing the abundance of individual species, they do not convey information about the conditions influencing the structure of the community. Moreover, these models do not exploit the interactions that exist between the different species. Our experiments have shown that these models capture only specific knowledge and tend to overfit the training data at hand.

The constructed multi-target habitat models have high expressive power. They exploit the relations between the different species to build a more representative and comprehensible habitat model. Furthermore, the multi-target habitat models can be readily interpreted by a domain expert. They offer insight into the community structures encountered within the different ecosystems under different environmental conditions. These models can be used, for example, to develop a decision support system to control the abundance of pests and harmful species without disrupting the complete community, e.g., decreasing the abundance of some beneficial species.

We have illustrated the use of multi-target predictive clustering trees in three case studies: modelling the diatom community in Lake Prespa, modelling bioindicator species in Slovenian rivers and modelling the *Collembola* community in Danish soils. For each of these case studies, we have constructed several habitat models and provided discussion concerning the models. Besides showing the environmental preferences of the individual species, each of the constructed models offers knowledge about the structure of the community of organisms encountered in the given ecosystem: it shows the environmental preferences of the species.

The presented methodology can be further employed in the context of using the species as bioindicators to determine the level of environmental pollution/quality of a given ecosystem. Namely, one can use the abundance of the species to predict the environmental conditions in the ecosystem. These environmental conditions can reflect the level of pollution in the water or in the soil, i.e., model the quality of the water or the soil. Moreover, this approach can be integrated with a system for automatic taxa identification from images (such as the one presented by Dimitrovski et al. (2012)) and a decision support system for preventing water or soil pollution. Such an integrated system could facilitate fast, efficient and automated environmental quality control.

References

Araújo, M. B., New, M., 2007. Ensemble forecasting of species distributions. Trends in ecology & evolution 22 (1), 42–47.

Bauer, E., Kohavi, R., 1999. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning 36 (1), 105–139.

Begon, M., Townsend, C., Harper, J., 2006. Ecology: From individuals to ecosystems. Blackwell.

Blockeel, H., 1998. Top-down induction of first order logical decision trees. Ph.D. thesis, Katholieke Universiteit Leuven, Leuven, Belgium.

Blockeel, H., Džeroski, S., Grbović, J., 1999. Simultaneous prediction of multiple chemical parameters of river water quality with Tilde. In: Zytkow, J. M., Rauch, J. (Eds.), Proc. of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases - LNCS 1704. Springer, pp. 32–40.

Breiman, L., 1996. Bagging Predictors. Machine Learning 24 (2), 123–140.

Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32.

Breiman, L., Friedman, J., Olshen, R., Stone, C. J., 1984. Classification and Regression Trees. Chapman & Hall/CRC.

Brodersen, K. H., Ong, C. S., Stephan, K. E., Buhmann, J. M., 2010. The Balanced Accuracy and Its Posterior Distribution. In: Proceedings of the 20th International Conference on Pattern Recognition - ICPR2010. IEEE Computer Society, pp. 3121–3124.

Demšar, D., Džeroski, S., Larsen, T., Struyf, J., Axelsen, J., Bruns-Pedersen, M., Krogh, P. H., 2006. Using multi-objective classification to model communities of soil. Ecological Modelling 191 (1), 131–143.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30.

Dimitrovski, I., Kocev, D., Loskovska, S., Džeroski, S., 2012. Hierarchical classification of diatom images using ensembles of predictive clustering trees. Ecological Informatics 7 (1), 19–29.

Drew, C. A., Wiersma, Y. F., Huettmann, F., 2011. Predictive Species and Habitat Modelling in Landscape Ecology. Springer.

Džeroski, S., 2001. Applications of symbolic machine learning to ecological modelling. Ecological Modelling 146 (1-3), 263–273.

Džeroski, S., 2009. Machine learning applications in habitat suitability modeling. In: Haupt, S. E., Pasini, A., Marzban, C. (Eds.), Artificial Intelligence Methods in the Environmental Sciences. Springer, pp. 397–412.

Džeroski, S., Demšar, D., Grbović, J., 2000. Predicting chemical parameters of river water quality from bioindicator data. Applied Intelligence 13 (1), 7–17.

Džeroski, S., Panov, P., Ženko, B., 2009. Ensemble methods in machine learning. In: Meyers, R. (Ed.), Encyclopedia of complexity and systems science. Springer, pp. 5317–5325.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., J. Phillips, S., Richardson, K., Scachetti-Pereira, R., E. Schapire, R., Soberón, J., Williams, S., S. Wisz, M., E. Zimmermann, N., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29 (2), 129–151.

Franklin, J., 2009. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge University Press.

Friedman, J. H., 1991. Multivariate adaptive regression splines. Annals of Statistics 19 (1), 1–67.

Friedman, M., 1940. A comparison of alternative tests of significance for the problem of m rankings. Annals of Mathematical Statistics 11, 86–92.

Gold, C., Feurtet-Mazel, A., Coste, M., Boudou, A., 2002. Field transfer of periphytic diatom communities to assess short-term structural effects of metals (Cd, Zn) in rivers. Water Research 36 (14), 3654–3664.

Iman, R. L., Davenport, J. M., 1980. Approximations of the critical region of the Friedman statistic. Communications in Statistics - Theory and Methods 9 (6), 571–595.

Jørgensen, S. E., Bendoricchio, G., 2001. Fundamentals of Ecological Modelling. Elsevier.

Kampichler, C., Wieland, R., Calmé, S., Weissenberger, H., Arriaga-Weiss, S., 2010. Classification in conservation biology: A comparison of five machine-learning methods. Ecological Informatics 5 (6), 441 – 450.

Kocev, D., 2011. Ensembles for predicting structured outputs. Ph.D. thesis, International postgraduate school Jožef Stefan, Ljubljana, Slovenia.

Kocev, D., Naumoski, A., Mitreski, K., Krstić, S., Džeroski, S., 2010. Learning habitat models for the diatom community in Lake Prespa. Ecological Modelling 221 (2), 330–337.

Kocev, D., Vens, C., Struyf, J., Džeroski, S., 2007. Ensembles of multi–objective decision trees. In: Proc. of the 18th European Conference on Machine Learning ECML'07 - LNCS 4701. Springer, pp. 624–631.

Kocev, D., Vens, C., Struyf, J., Džeroski, S., 2013. Tree ensembles for predicting structured outputs. Pattern Recognition 46 (3), 817–833.

Krstić, S., 2005. 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.

Kuncheva, L., 2004. Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience.

Lek, S., Scardi, M., Verdonschot, P. F., Descy, J.-P., Park, Y.-S., 2005. Modelling Community Structure in Freshwater Ecosystems. Springer.

Nemenyi, P. B., 1963. Distribution-free multiple comparisons. Ph.D. thesis, Princeton University, Princeton, NY, USA.

Oppel, S., Meirinho, A., Ramírez, I., Gardner, B., O'Connell, A. F., Miller, P. I., Louzao, M., 2012. Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. Biological Conservation 156, 94–104.

Pino-Mejías, R., Cubiles-de-la-Vega, M. D., Anaya-Romero, M., Pascual-Acosta, A., Jordán-López, A., Bellinfante-Crocci,

N., 2010. Predicting the potential habitat of oaks with data mining models and the r system. Environmental Modelling & Software 25 (7), 826 – 836.

Ponge, J.-F., 1991. Food resources and diets of soil animals in a small area of Scots pine litter. Geoderma 49 (1-2), 33–62.

Scott, J. M., Heglund, P. J., Morrison, M. L., Haufler, J. B., Raphael, M. G., Wall, W. A., Samson, F. B., 2002. Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press.

Seni, G., Elder, J. F., 2010. Ensemble methods in data mining: Improving accuracy through combining predictions. Morgan & Claypool Publishers.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45 (4), 427–437.

Supplementary information. The supplementary information for this manuscript is available at the personal web pages of the authors or at the following direct link: `http://kt.ijs.si/DragiKocev/wikipage/lib/exe/fetch.php?media=2013_habmod_supplementary.pdf`.