Ecological Modelling xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Community structure models are improved by exploiting taxonomic rank with predictive clustering trees

Jurica Levatić[a,b,*], Dragi Kocev[a], Marko Debeljak[a,b], Sašo Džeroski[a,b]

[a] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
[b] Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Community structure modelling
Taxonomic rank
Predictive clustering trees
Classification
Hierarchical multi-label classification

ABSTRACT

Community structure modelling studies the influence of biotic and abiotic factors on the abundance and composition of a given taxonomic group of organisms. With the advancement of measurement and sensor technology, the availability, precision and complexity of environmental data constantly increases. Nowadays, measurements of ecosystems provide a complete snapshot of the state of the system, including information about the community structure of organisms that are present in a given sample. These measurements include multi-species data that are typically analysed by constructing community models as collections of models built for each species separately (local models) without considering the possible (taxonomic) relationships among species.

In this work, we propose to construct a single community structure model for all the species (global model) that is able to exploit the aforementioned relationships. Namely, we investigate whether inclusion of additional information in the form of taxonomic rank or multiple species helps to build better community structure models. More specifically, we use predictive clustering trees (a generalized form of decision trees) to build models for three practically relevant datasets from the task of community structure modelling: microarthopod community living in the agricultural soils of Denmark, organisms living in Slovenian rivers and vegetation found in the State of Victoria, Australia.

On each dataset, we compare the performance of four types of community structure models, which correspond to four machine learning tasks: Single species models without taxonomic rank correspond to single-label classification; single species models with taxonomic rank correspond to hierarchical single-label classification; multi-species models without taxonomic rank correspond to multi-label classification; and multi-species models with taxonomic rank correspond to hierarchical multi-label classification. The results of the experimental evaluation reveal that by using the taxonomic rank and the multi-species aspect of the data, we are able to learn better community structure models.

1. Introduction

One of the most fundamental questions in ecology is: What is the composition of a community of organisms with respect to the environment? Prediction of such composition through modelling of the community structure answers this question when empirical evidence is not attainable. The community is an assemblage of species populations that occur together in space and time. The species that assemble a community are determined by dispersal constraints, abiotic environmental constraints and biotic interactions (Belyea and Lancaster, 1999). To reflect these different constraints, the terms dispersal assembly rules, abiotic assembly rules, and biotic assembly rules are used, respectively (Götzenberger et al., 2012). Community ecology uses the assembly rules approach to investigate the mechanisms that structure biological communities. The objective of assembly rules is to predict species composition in a specified habitat dominated by a set of environmental conditions: (1) to simply predict the presence or absence of species; and (2) to predict the abundance of species (Keddy, 1992; Weiher and Keddy, 2001; Götzenberger et al., 2012).

Abiotic assembly rules are studied with gradient analysis. The choice of the gradients/factors (i.e., environmental variables) can be subjective, because it is based on existing knowledge about the studied species (e.g., elevation and precipitation are gradients for forest communities) and the availability of data about these species that are organised along the gradient of a factor. The fact that the

* Corresponding author at: Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia. Tel.: +386 1 477 3639.
E-mail addresses: jurica.levatic@ijs.si (J. Levatić), dragi.kocev@ijs.si (D. Kocev), marko.debeljak@ijs.si (M. Debeljak), saso.dzeroski@ijs.si (S. Džeroski).

http://dx.doi.org/10.1016/j.ecolmodel.2014.10.023
0304-3800/

species from a community can be arranged in a sequence along a gradient of some environmental factor does not prove that this factor is the most important one. In order to improve the abiotic assembly rules that might contribute to the understanding of community assembly rules, we propose the use of machine learning methods to mitigate the influence of subjective selection of environmental gradients.

There is a plethora of environmental studies that utilize different statistical and machine learning techniques to model ecological communities. 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), Bayesian and mixture models, support vector machines (SVMs) and genetic algorithms (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; Scott et al., 2002). Depending on the context of the study, different methods should be preferred. If the goal is to produce maps of habitat suitability (i.e., species distribution models), 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 ecosystem under consideration, then the methods that yield interpretable models with satisfactory predictive performance should be preferred (such as classification and regression trees/rules). In this work, we focus on classification trees. Classification trees are predictive models particularly suited for the analysis of complex ecological data, since they can deal with non-linear relationships, high-order interactions and missing data, while being easily interpretable at the same time (De'ath and Fabricius, 2000).

Nowadays, data describing ecosystems are available and often include multi-species data (i.e., data on community of organisms encountered at a given site). A typical approach for constructing community structure models is to build a separate (local) model for each of the species in the group (predicting presence/absence), then aggregate the outputs of these models to determine the structure of the entire community. However, such an approach fails to take advantage of the information contained in multi-species data (co-occurrence of species and taxonomic relationships among species) and is not able to model generalized group responses.

An alternative approach is to build a global model that simultaneously predicts the presence/absence of all organisms in the community. Although 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)), research community seldom applies these. The most prominent reason for this is probably the lack of interpretability of the models produced by the aforementioned methods. In this work, we propose to build interpretable community structure models by considering a type of classification task named hierarchical multi-label classification (HMC), where classes are hierarchically organized and each example can belong to more than one class. The HMC approach overcomes the previously mentioned shortcomings by building interpretable global models which exploit taxonomic relationships in the biological community.

The main objective of this work is to explore in detail how the exploitation of multi-species data and information about taxonomic rank affects model interpretability and classification performance. For this purpose, we compare the approach of learning trees for HMC with three other modelling approaches on three practically relevant datasets exemplifying the task of community structure modelling. The datasets at hand describe the microarthopod community in the soils of Denmark, the organisms living in Slovenian rivers, and the vegetation found in the State of Victoria, Australia.

For each dataset, we build four types of models corresponding to the specific modelling approaches depicted in Fig. 1. First, we

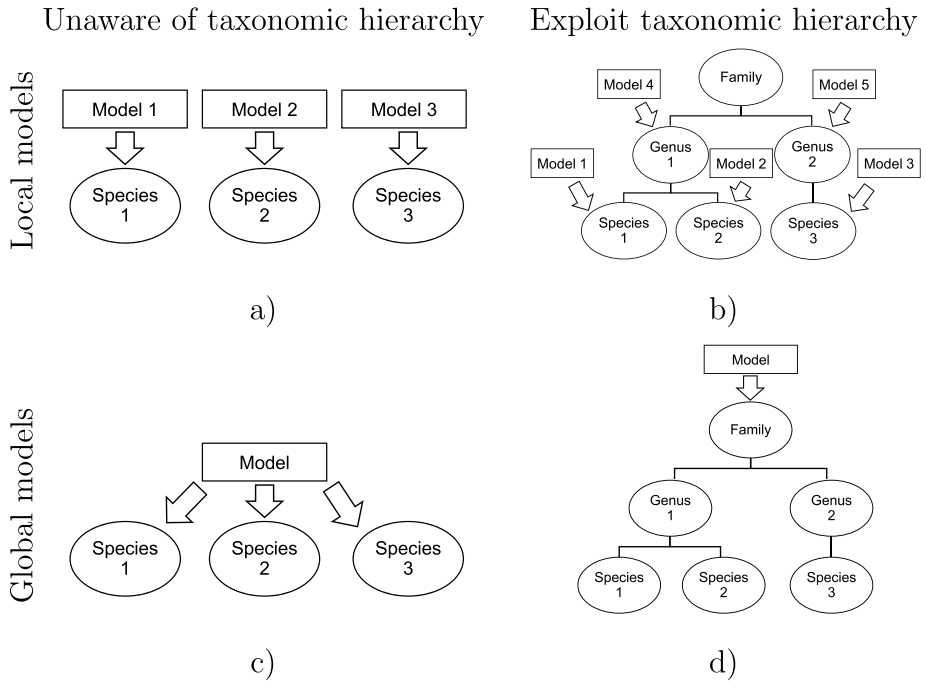

Fig. 1. Schematic representation of the four different modeling tasks we consider to investigate how exploitation of taxonomic rank affects the performance at the task predicting community structure. Single label classification (a), builds a separate model for each of the species, while hierarchical single label classification (b), builds a separate model for each edge of the taxonomic hierarchy (each model is trained by using only data that is relevant to that edge). Multi label classification (c) and hierarchical multi-label classification (d) build one (global) model which considers all of the species at once: the former approach (c) is unaware of the taxonomic hierarchy, while the latter approach (d) exploits information about the taxonomic hierarchy. Each of the models has as input the same environmental variables, while the different kinds of output are given at the pointed ends of the arrows.

Please cite this article in press as: Levatić, J., et al., Community structure models are improved by exploiting taxonomic rank with predictive clustering trees. Ecol. Model. (2014), http://dx.doi.org/10.1016/j.ecolmodel.2014.10.023

construct two types of local models: (1) models for each species separately (single-target classification) and (2) models for each edge of the taxonomic hierarchy (hierarchical single-label classification). Second, we build two types of global models: (3) a model which considers all species at once, but is unaware of the underlying taxonomic hierarchy (multi-label classification), and similarly, (4) a global model for all species, which exploits the information about the taxonomic hierarchy (hierarchical multi-label classification).

We use predictive clustering trees (PCTs) (Blockeel et al., 1998; Struyf and Džeroski, 2006; Vens et al., 2008; Slavkov et al., 2010; Kocev et al., 2013) to construct the four types of models. PCTs are a generalization of ordinary decision trees and they are able to tackle each of the modelling tasks mentioned above. PCTs have been successfully used in the past for a number of modelling tasks in an ecological context, including analysis of time-series on agroecosystem vegetation (Debeljak et al., 2011), vegetation condition prediction (Kocev et al., 2009), soil-quality analysis (Cortet et al., 2011), and habitat models learning (Demšar et al., 2006; Kocev et al., 2010; Kocev and Džeroski, 2013).

The datasets used here have been previously analysed using only simpler methods and not in the context of community structure modelling. First, the previous studies concerned with river communities (Džeroski et al., 2000; Blockeel et al., 1999) used the presence/absence and the abundance of the species to infer indicators of water quality. Next, Demšar et al. (2006) used multi-objective (i.e., multi-target) regression trees to model the overall abundance of the species and the biodiversity of the communities of soil microarthropods. In contrast, we learn models for predicting the structure of the communities, in terms of presence and absence of species in the considered ecosystem. Third, Slovenian river communities and Danish soil microarthropods were also analysed by applying single-label and multi-label classification models to predict the community structures (Kocev and Džeroski, 2013). We consider here a dataset that contains a larger set of organisms encountered in the Slovenian rivers (while only 14 species were considered earlier here we consider 491 species). Finally, the Australian vegetation dataset has not been used for community structure modelling before. This dataset has been thus far used only in the context of clustering the sampling sites (Gjorgjioski et al., 2008). Note that this dataset is very challenging because it contains data on 27,482 sampling sites and 3172 different species. All in all, none of the previous studies exploited taxonomic rank during the model construction and the combination of the taxonomic rank and the multi-species aspect of the data.

2. Materials and methods

2.1. Predictive clustering trees

The PCTs framework views a decision tree as a hierarchy of clusters: the top-node corresponds to one cluster containing all data (i.e., all examples), which is recursively partitioned into smaller clusters while moving down the tree. PCTs are built with a greedy recursive top-down induction algorithm (Table 1). The learning algorithm takes as input a set of examples (E) and outputs a tree. The procedure starts by selecting a test (t) for the root node by using a heuristic function (h) computed on the training examples. The goal of the heuristic (h) is to select a test (t) which maximizes the variance reduction caused by the partitioning ($\mathcal{P}$) of the examples into subsets according to the test outcome (see the BestTest procedure in Table 1). In this way, the cluster homogeneity is maximized and the predictive performance of the tree is improved.

During the construction of the PCTs, all of the predictors (i.e., environmental variables) are considered at each step. However, the final tree contains in its internal nodes only the predictors selected by the heuristic function. These are the variables with the highest scores for the heuristic function at each of the construction steps.

Table 1
The top-down induction algorithm for PCTs.

procedure PCT	**procedure** BestTest
Input: A dataset E	**Input:** A dataset E
Output: A predictive clustering tree	**Output:** the best test (t^*), its heuristic score (h^*) and the partition ($\mathcal{P}^*$) it induces on the dataset (E)
1: $(t^*, h^*, \mathcal{P}^*) = \text{BestTest}(E)$	1: $(t^*, h^*, \mathcal{P}^*) = (none, 0, \emptyset)$
2: **if** $t^* \neq none$ **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: $tree_i = \text{PCT}(E_i)$	4: $h = Var(E) - \sum_{E_i \in \mathcal{P}} \frac{\lvert E_i \rvert}{\lvert E \rvert} Var(E_i)$
5: **return** $\text{node}(t^*, \bigcup_i \{tree_i\})$	5: **if** $(h > h^*) \wedge \text{Acceptable}(t, \mathcal{P})$ **then**
6: **else**	6: $(t^*, h^*, \mathcal{P}^*) = (t, h, \mathcal{P})$
7: **return** leaf(Prototype(E))	7: **return** $(t^*, h^*, \mathcal{P}^*)$

The partitioning of the examples is recursively repeated until a stopping criterion is satisfied, e.g., further partitioning of the examples yields a tree with a lower quality. In that case, the prototype (i.e., the prediction) is calculated and stored in the corresponding leaf of the tree. The prototype is a vector of probabilities that an example belongs to a given class, i.e., probabilities for encountering each of the species at a given site.

The predictions for an example that arrives at a leaf is obtained by applying a user defined threshold τ to the probabilities p_i for each class c_i stored at that leaf. If the probability p_i is above τ, then the examples belong to class c_i. The framework is implemented in the CLUS system (Blockeel and Struyf, 2002), available for download at http://clus.sourceforge.net.

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 the prediction in each leaf) as *parameters* that can be instantiated for a given learning task. So far, PCTs have been instantiated for the following tasks: multi-target prediction which includes multi-label classification, hierarchical multi-label classification, multi-target regression, and prediction of time-series. For further information, we refer the reader to (Struyf and Džeroski, 2006; Vens et al., 2008; Slavkov et al., 2010; Kocev et al., 2013). In this article, we focus on the first two tasks.

2.1.1. Global predictive models

PCTs for multi-label classification (MLC) predict multiple discrete targets simultaneously (i.e., they exploit the multi-species data, but not the taxonomic rank). Therefore, the variance function of a set of examples E for the PCTs for MLC is computed as the sum of the *Gini* indices of the target variables, i.e., $Var(E) = \sum_{i=1}^{T} Gini(E, L_i)$.

CLUS-HMC is the instantiation of the PCT algorithm for hierarchical classification (i.e., considers both the multi-species aspect of the data and the taxonomic rank). The variance function for CLUS-HMC is defined as the average squared distance between each example's class vector (L_i) and the set's mean class vector ($\bar{L}$), i.e.:

$$Var(E) = 1/|E| \cdot \sum_{E_i \in E} d(L_i, \bar{L})^2.$$

The distance measure used in the above formula is a weighted Euclidean distance:

$$d(L_1, L_2) = \sqrt{\sum_{l=1}^{|L|} w(c_l) \cdot (L_{1,l} - L_{2,l})^2},$$

Please cite this article in press as: Levatić, J., et al., Community structure models are improved by exploiting taxonomic rank with predictive clustering trees. Ecol. Model. (2014), http://dx.doi.org/10.1016/j.ecolmodel.2014.10.023

Table 2
Properties of the datasets: N is the number of samples, D/C are the numbers of discrete/continuous attributes, $\mathcal{L}$ is the number of labels, $|\mathcal{H}|$ is the number of nodes in the taxonomic hierarchy, $\mathcal{H}_d$ is the maximal depth of the taxonomic hierarchy, $\overline{\mathcal{L}_L}$ is the average number of labels per example.

Domain [reference]	N	D/C	$\mathcal{L}$	$\lvert\mathcal{H}\rvert$	$\mathcal{H}_d$	$\overline{\mathcal{L}_L}$
Slovenian rivers (Džeroski et al., 2000)	1060	0/16	491	724	5	25
Danish agricultural soils (Demšar et al., 2006)	1944	132/5	35	72	4	7
Australian vegetation (Gjorgjioski et al., 2008)	27,482	0/81	3172	4524	6	28

where $L_{i,l}$ is the lth component of the class vector L_i of an instance E_i, $|L|$ is the size of the class vector, and the class's weight $w(c)$ depends on its depth within the hierarchy.

2.1.2. Local predictive models

Local models for predicting structured outputs use a collection of predictive models, each predicting a component of the overall structure that needs to be predicted. First, we consider collections of models for single-label classification (i.e., models that exploit neither the multi-species data nor the taxonomic rank). Learning a single model of this type is a special case of multi-label classification (with one label), and use the same algorithm as for multi-label classification trees. We call these models single-label classification trees.

Second, we consider the task of hierarchical single-label classification (HSC) (i.e., constructing models that exploit the taxonomic rank but not the multi-species aspect of the data). Note that there are four different approaches to the more general task of HMC used to predict the hierarchy of classes with local models: flat classification, local classifiers per level, local classifiers per node, and local classifiers per parent node (for details, see (Silla and Freitas, 2011)). In this work, we use the last approach (HSC), since it performs better in terms of predictive performance, model complexity and induction time (Vens et al., 2008). In particular, the CLUS-HSC algorithm constructs a decision tree classifier for each edge (connecting a class c with a parent class $par(c)$) in the hierarchy, thus creating an architecture of classifiers.

2.2. Data description

We use three datasets containing information about the structure of communities composed of river water organisms in Slovenian rivers (Džeroski et al., 2000), soil microarthropods in Danish agricultural soils (Demšar et al., 2006), and vegetation in the State of Victoria, Australia (Gjorgjioski et al., 2008). From the original datasets, we removed species which are present at less than three sites, because such a low number of appearances is not sufficient for modelling purposes. The datasets contain information about the presence and absence of species (i.e., both presences and absences are used for learning the predictive clustering tree models). Presence of the species was determined by an expert biologist, and the species not found at a particular sampling site were considered absent.

The main statistical properties of the datasets are given in Table 2. We can observe that the datasets vary in size, including number of sampling sites, the number of species considered, the number of attributes (environmental variables) and the characteristics of the label hierarchy (details of the species taxonomic rank). To understand better the terms related to the properties of the label hierarchy (the number of labels, the number of nodes etc.) consider the toy hierarchy given in Fig. 1d. This hierarchy has three labels ($\mathcal{L} = 3$), six nodes in total ($|\mathcal{H}| = 6$), and a maximal depth of three ($\mathcal{H}_d = 3$). Note that, a label is a leaf node of the hierarchy. This is species in the toy hierarchy, but can be a taxonomic entity of any rank as long as it's terminal node in the hierarchy. Suppose we have a dataset with two examples (i.e., sampling sites), where species 1 and 2 were found at the first site, and species 3 at the second site. The average number of labels per example ($\overline{\mathcal{L}_L}$), i.e., the average number of species per site, of this dataset is 1.5, since the first example has two labels and the second example has one label.

2.2.1. Slovenian rivers dataset

The data for the water organisms from Slovenian rivers was obtained from the Hydrometeorological Institute of Slovenia (now Environment Agency of Slovenia). The data provided cover a 6-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. The monitoring and the sampling were performed in accordance with standards determined with the rules on surface water status monitoring of Republic of Slovenia (Rules on surface water status monitoring, 2009; Amendmends of the rules, 2011).

The physical and chemical samples include the measured values of sixteen different parameters: biological oxygen demand (BOD), chlorine concentration (Cl), CO_2 concentration, electrical conductivity, chemical oxygen demand ($K_2Cr_2O_7$ and $KMnO_4$), concentrations of ammonia (NH_4^+), NO_2, NO_3^-, and dissolved oxygen (O_2), alkalinity (pH), PO_4^{3-}, oxygen saturation, SiO_2, water temperature, and total hardness. The biological samples include a list of all taxa present at the sampling site. The sampled living organism families (or other taxonomical units, referred to as taxa) cover phytobenthos and macrophytes, benthic invertebrates and fish.

2.2.2. Danish agricultural soils dataset

The data for the soil microarthropods from Danish agricultural soils describe 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 20 m × 20 m field area, with a distance of 5 m between the individual samples. Sampling was performed in the upper 5.5 cm soil layer and the sampling containers measured 6 cm in diameter.

The data concern the *Collembola* species community found in the soil samples. These species can be used as indicators of soil quality (in particular soil desiccation) and some are considered as pests for plants. Also, they are the main biological factors responsible for the control of the soil microorganisms (Demšar et al., 2006).

The input attributes describe the field where the microarthropod sample was taken. They mainly include agricultural practices applied to the field (e.g., planted crops, tillage, fertilizer and pesticide use, the history of crops and grazing, etc.) and some soil properties, namely soil type (Greve and Breuning-Madsen, 1999) and treatment of the soil. The complete list of input attributes is given in Table A.4.

2.2.3. Australian vegetation dataset

The data about Australian vegetation were collected across the State of Victoria, Australia, an area of approximately 22,000,000 hectares (Gjorgjioski et al., 2008). The Victoria State is climatically and geologically varied and supports about 4000 indigenous vascular plant species. The area was divided into about 25,000 sampling sites or quadrats, where homogeneous areas of vegetation were sampled over the period of 30 years.

Please cite this article in press as: Levatić, J., et al., Community structure models are improved by exploiting taxonomic rank with predictive clustering trees. Ecol. Model. (2014), http://dx.doi.org/10.1016/j.ecolmodel.2014.10.023

Quadrat sizes depend on the type of plant community and its minimal area (i.e., a point where further increase in quadrat size result in only sporadic addition of novel species). Generally, quadrats in grassland and shrubland are 100 m^2 in size and quadrats in mallee and woodland are 900 m^2 in size.

All vascular plants growing in or extending over the sample space were recorded. The geographic coordinates of all the sites were recorded, and were used to extract many environmental (climatic, radiometric, topographic) and spectral (remote sensing) variables from a 'stack' of data themes stored in a GIS. The complete list of input attributes is given in Table A.5.

2.3. Experimental setup

We constructed four types of predictive models, as described in Fig. 1, for each of the three datasets.

We used *F*-test pruning to ensure that the produced models are not overfitted and have better predictive performance (Vens et al., 2008). The *F*-test pruning uses the exact Fisher test to check whether a given split/test in an internal node of the tree yields a statistically significant reduction in variance. If there is no such a split/test, the node is converted to a leaf. A significance level is selected from a set of *p*-values (0.125, 0.1, 0.05, 0.01, 0.005 and 0.001) to optimize the predictive performance by using internal 3-fold cross validation.

We evaluated the predictive performance of the models on the leaf classes/labels in the target hierarchy (i.e., at the species level). We made this choice in order to ensure a fair comparison across the different tasks. Namely, if we consider all labels (the leaf labels and the inner node labels), the single-label classification task will be very close to the one of hierarchical single-label classification; similarly, the task of multi-label classification becomes very close to the one of hierarchical multi-label classification.

By evaluating only the performance on leaf labels, we can also measure more precisely the influence of the inclusion of the different types of information in the learning process on the predictive performance of the models. To further ensure this, we set the w_0 parameter for the weighted Euclidean distance for HMC as 1: all labels in the hierarchy contribute equally. By doing this, we measure only the effect of integrating the multi-label information (considering the multiple labels simultaneously) and the hierarchy information (considering taxonomic rank).

2.4. Model performance measures

We evaluated the models by using the area under the Precision-Recall curve (AUPRC) as predictive performance measure, and in particular, the area under the average Precision-Recall curve ($AU\overline{PRC}$) as suggested by Vens et al. (2008). The points in the Precision-Recall (PR) space were obtained by changing the value of the threshold τ from 0 to 1 with step 0.02. For each value of the threshold τ, precision and recall values are micro-averaged as follows: $\overline{Prec} = \sum_i TP_i / \left(\sum_i TP_i + \sum_i FP_i\right)$, and $\overline{Rec} = \sum_i TP_i / \left(\sum_i TP_i + \sum_i FN_i\right)$, where i ranges over all classes that are leafs in the output hierarchies.

AUPRC is a threshold independent performance measure. Closely related to it is the area under the receiver operating characteristic curve (AUROC). However, AUROC rewards the correctly predicted negative examples, which can give an overly optimistic performance when there is a large skew in the class distribution (i.e., the number of positive and negative examples are imbalanced) (Davis and Goadrich, 2006). Since this is the case in the datasets considered here, we have chosen to evaluate the studied methods by using the AUPRC measure.

We measured the performance of the predictive models along several dimensions. First, we estimated the predictive performance of the models on unseen cases by using 10-fold cross-validation. The 10-fold cross validation 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 then repeated 10 times, where each disjoint part is used for testing exactly once. 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 performance on unseen cases (estimated by 10-fold cross-validation), we also present the performance of the methods on the complete dataset (i.e., training set performance) to assess the descriptive power of the methods. For the next evaluation dimension, we measured how much the different models over-fit the training data. To this end, we use the relative decrease of training set performance to the one estimated by 10-fold cross-validation. We define this as an over-fitting score $\left(O_s = \left(AU\overline{PRC}_{train} - AU\overline{PRC}_{test}\right) / AU\overline{PRC}_{train}\right)$. Smaller values of this score mean that the approach at hand over-fit the data less.

Finally, we measured the model complexity and the time efficiency of learning predictive models. The complexity of the global models is the number of nodes in a given tree, while the complexity of the local models is the sum of all nodes from all trees. Similarly, the learning time for the global approaches is the time needed to construct the single global model, while the learning time for the local approaches is the sum of the times needed to construct all of the local models.

We adopt the recommendations by Demšar (2006) for the statistical evaluation of the results. We used the corrected non-parametric test for statistical significance on the per-fold-data for the folds of 10-fold cross validation for each dataset separately. Afterwards, to detect where the statistically significant differences appear (between which methods), we used the Nemenyi post-hoc test (Nemenyi, 1963). We present the result from the Nemenyi post-hoc test with an average ranks diagram (Demšar, 2006). The ranks are depicted on an axis, in such a manner that the best ranking algorithms are at the right-most side of the diagram (i.e., the best performing algorithms have lower values for the average ranks). The algorithms that do not differ statistically significantly (in performance) are connected with a line.

3. Results

In this section, we present the results of the experimental evaluation. First, we discuss the predictive performances and the efficiency of learning the obtained models. We then analyze the interpretability of the models.

3.1. Performance of the models

The performance of the community structure models is given in Table 3. Below we discuss the results, first focusing on the predictive performance and then on the efficiency of learning the different community structure models. In terms of predictive performance, we are concerned with three main questions: (1) Which is the best community structure modelling method? (2) Is it preferable to learn local or global models, i.e., does exploiting the multi-species data help? and (3) Does the inclusion of information about the taxonomic hierarchy improve the predictive performance?

Let us start by identifying the best and the worse performing methods. The best performing method is the HMC method that includes information both on the taxonomic rank and the multi-species aspect of the data. On the other hand, the simplest method, i.e., the one performing single-label classification, by learning the local models unaware of the taxonomic hierarchy

Please cite this article in press as: Levatić, J., et al., Community structure models are improved by exploiting taxonomic rank with predictive clustering trees. Ecol. Model. (2014), http://dx.doi.org/10.1016/j.ecolmodel.2014.10.023

Table 3
Comparisons of the performances of four modeling methods in terms of predictive power ($\overline{AUPRC}$), the relative decrease of performance between the training set and test set (O_S), the learning time (in seconds) and the model complexity (the number of nodes in the decision trees). The best predictive performance for each dataset is shown in bold.

Dataset	Method	$\overline{AUPRC}$	O_S	Learning time	Complexity
Slovenian rivers	Single-label	0.239	0.692	23.3	15,336
	HSC	0.309	0.591	10.2	25,035
	Multi-label	0.322	0.007	9.4	1
	HMC	**0.374**	0.132	0.6	37
Danish farms	Single-label	0.790	0.099	3.7	2605
	HSC	0.808	0.083	1.3	2873
	Multi-label	0.801	0.112	0.7	265
	HMC	**0.815**	0.065	0.4	259
Australian vegetation	Single-label	0.232	0.715	14,888.2	482,745
	HSC	0.306	0.591	76,023.2	648,970
	Multi-label	0.278	0.684	4639.5	23,699
	HMC	**0.376**	0.180	313.5	1279

and the multi-species information, is clearly outperformed by all the other methods.

Next, global models outperform their local counterparts across the three datasets. Multi-label classification is better than single-label classification, while hierarchical multi-label classification outperforms hierarchical single-label classification. Moreover, the global models tend to overfit less than the local counterparts thus achieving better generalization.

Finally, the results clearly demonstrate that the use of the taxonomic rank improves the performance of both local and global models. We can thus conclude that the inclusion of the information on the multi-species aspect of the data and/or the taxonomic rank improves the predictive performance of the models; however, our results do not provide clear evidence which of the two sources of information is more beneficial.

To test whether the observed differences in predictive (modelling) performance are statistically significant, we use the Friedman test (Demšar et al., 2006). The results from the test show that the difference in performance is statistically significant for each dataset with p-value <0.0001. Fig. 2 presents the average ranks from the Nemenyi post-hoc test for all types of models. The diagrams show that the HMC models perform statistically significantly better than the single-label models in all three cases. The HMC approach is significantly better than HSC on the Slovenian rivers dataset and multi-label classification on the Australian vegetation dataset. Furthermore, the HSC method is statistically significantly better than the single-label method on the Danish soils and the Australian vegetation, while it is better (not statistically significantly) than the single-label method on the Slovenian rivers.

The efficiency of the different community structure modelling methods is expressed by the time needed to construct the models and the size of the models. Note that the latter is of greater importance, since it is directly related to the models' interpretability. The results show that learning the global models is much faster (more efficient) than learning the local models in both time and size: The HMC models are constructed fastest and are smallest (except for the Slovenian rivers dataset, where the selection of the significance level for the F-test pruning was too stringent for the multi-label classification method).

3.2. Interpretation of the obtained models

The interpretability of a model is a highly desired property in the field of community structure modelling. The interpretable

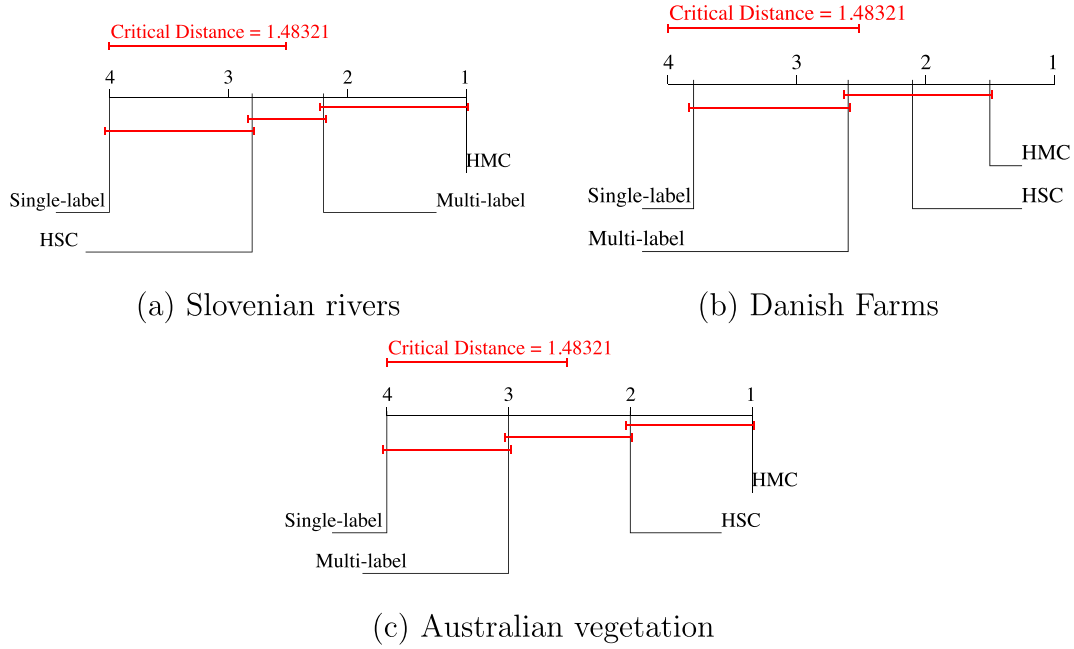

Fig. 2. Average ranks diagrams for the performance of the four methods in terms of $\overline{AUPRC}$ for each of the three datasets. Better algorithms are positioned on the right-hand side, and those that differ in performance by less than the critical distance at p-value = 0.05 are connected with a line.

Please cite this article in press as: Levatić, J., et al., Community structure models are improved by exploiting taxonomic rank with predictive clustering trees. Ecol. Model. (2014), http://dx.doi.org/10.1016/j.ecolmodel.2014.10.023

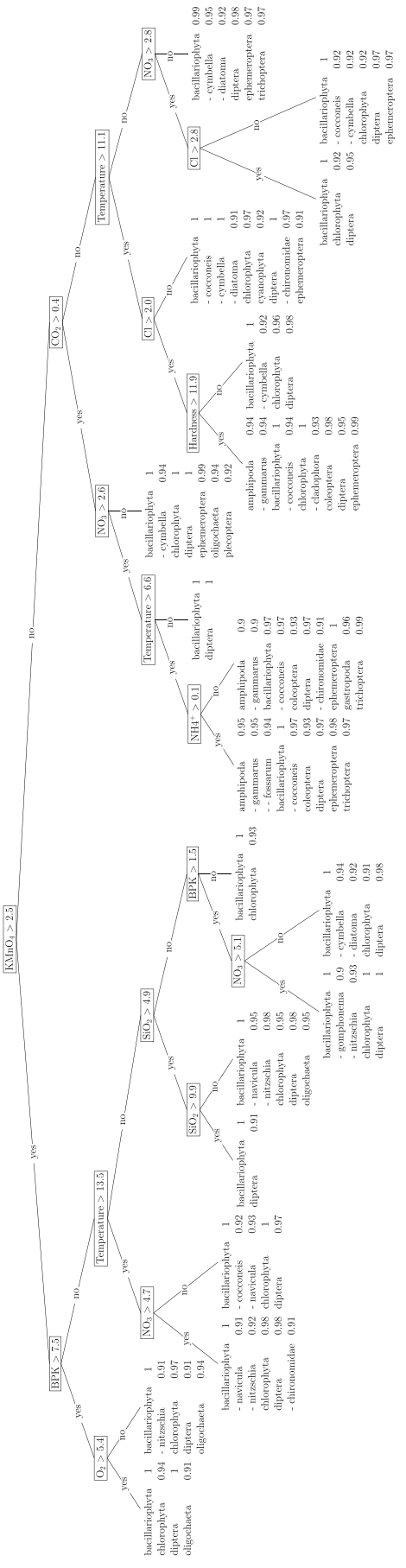

Fig. 3. HMC decision tree for the Slovenian rivers dataset. The numbers beside the taxa represent the probabilities of presence. Taxa present with probability greater than 0.9 are listed.

Please cite this article in press as: Levatić, J., et al., Community structure models are improved by exploiting taxonomic rank with predictive clustering trees. Ecol. Model. (2014), http://dx.doi.org/10.1016/j.ecolmodel.2014.10.023

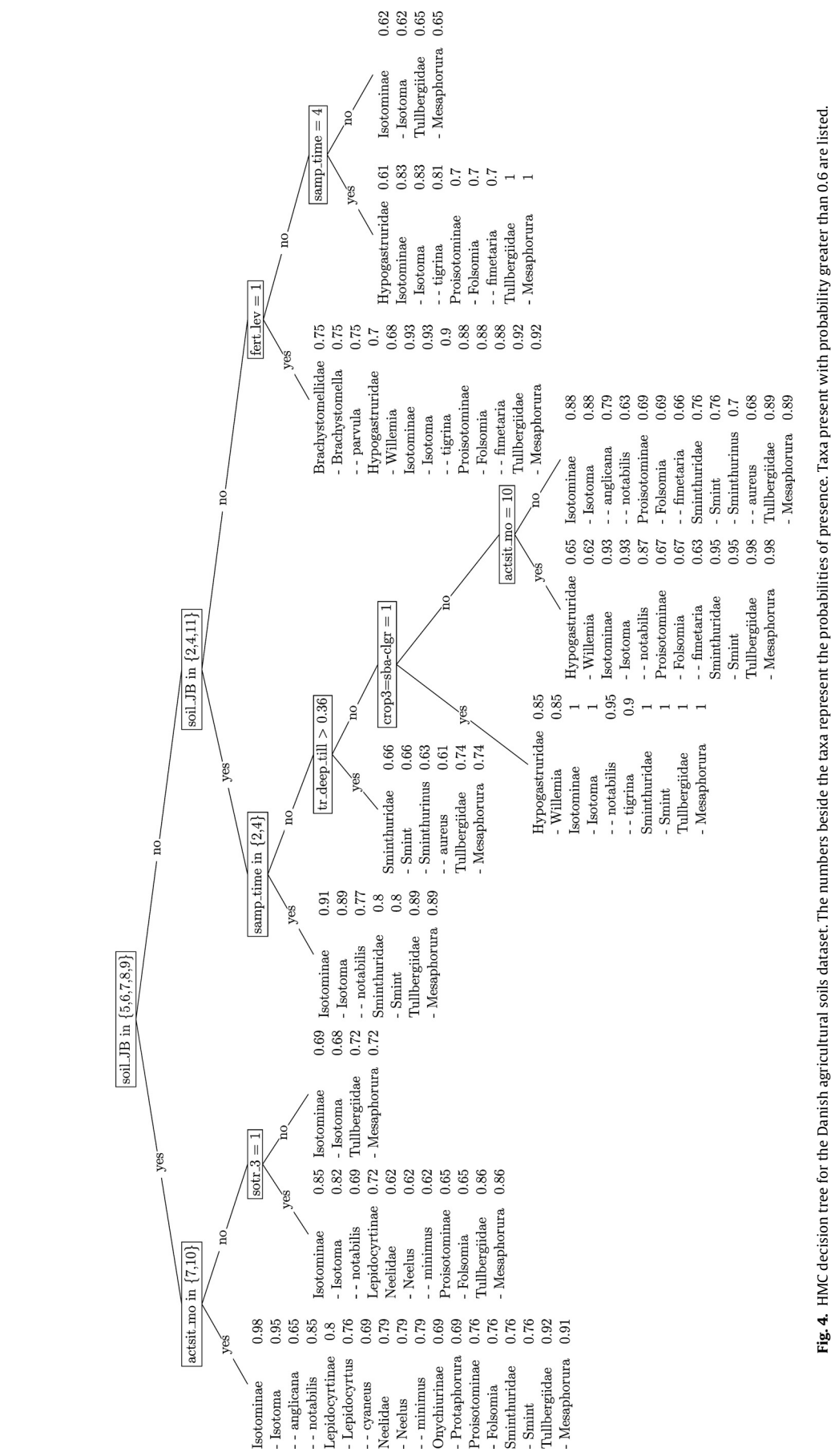

Fig. 4. HMC decision tree for the Danish agricultural soils dataset. The numbers beside the taxa represent the probabilities of presence. Taxa present with probability greater than 0.6 are listed.

Please cite this article in press as: Levatić, J., et al., Community structure models are improved by exploiting taxonomic rank with predictive clustering trees. Ecol. Model. (2014), http://dx.doi.org/10.1016/j.ecolmodel.2014.10.023

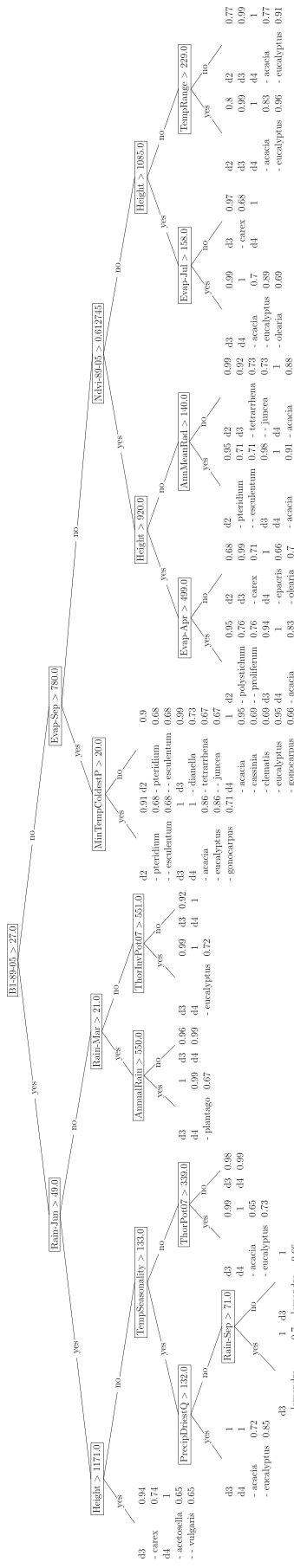

Fig. 5. HMC decision tree for the Australian vegetation dataset. The numbers beside the taxa represent the probabilities of presence. Taxa present with probability greater than 0.65 are listed.

models offer an insight into the processes that are going on within the observed ecosystem. All the predictive models that we consider here (PCTs) are readily interpretable. However, there are two major differences in interpretability between local and global PCT models. Firstly, global models, especially HMC trees, are much smaller than local models (Table 3), and allow for easier analysis and interpretation. Secondly, each local model provides information only for a small part of the output space, i.e., they are valid just for a single species (single-label trees) or taxon (HSC).

The HSC models are similar to the single-label classification models, but the former have the added complexity that they are organized into a hierarchical architecture and cannot be interpreted out of context and independently from other trees. This makes the interpretation of the HSC models an even more difficult task: To reconstruct the complete community model, one needs to look at the separate models and then try to make some overall conclusions. However, this could be very tedious or even impossible for sites with high biodiversity where there are hundreds of species present, such as the sites we consider here – Slovenian rivers or vegetation in the State of Victoria, Australia.

The single global model is valid for the complete structured output, i.e., for the whole community of species present in the ecosystem. In Figs. 3–5, we present the HMC models for community structure of Slovenian rivers, Danish agricultural soils and Australian vegetation, respectively. The global models are able to capture the interactions present among species, i.e., which species can co-exist at a location with given environmental properties. Moreover, the HMC models, as compared to the multi-label models, offer additional information about the higher taxonomic rank. For example, the HMC model given in Fig. 4 states that there is a low probability (0.65) that the species *Isotoma anglicana* is present under the given environmental conditions (the soil is clayey, sandy clayey, heavy clayey or silty, and the age of current situation is 7 or 10 months), while there is a high probability (0.95) that the genus *Isotoma* is present (left-most leaf of the HMC tree).

4. Discussion

The obtained models for the community structure of Slovenian rivers, Danish agricultural soils and Australian vegetation identify the most important abiotic factors determining the presence or absence of species in the corresponding biological communities. The community structure model for Slovenian rives (Fig. 3) identifies the chemical and biological oxygen demand, the concentration of carbon dioxide, temperature and the concentration of nutrients as the most important abiotic factors influencing the structure of the water community. The highest number of taxonomic groups is encountered in environmental conditions with low chemical oxygen demand, high temperature and high availability of nutrients (NO_3^-), while the conditions with high chemical and biological oxygen demands, low temperature and low concentration of nutrients restrict the number of taxonomic groups in Slovenian waters. These critical abiotic factors have also been identified in the previous studies of this dataset (Kocev and Džeroski, 2013; Džeroski et al., 2000).

The HMC tree for Danish agricultural soils (Fig. 4) shows that the community structure depends mostly on the properties of the soil, the applied crop management practices and the time since application. Namely, it differentiates the communities that develop in clay (soils with JB index between 5 and 9) and other types of soil. In particular, the time dimension plays the second most important role because the number of taxonomic groups in the predicted communities is increasing with the time since the last disturbance of the soil (e.g., by plowing, sowing and harvesting) and the duration of the bioactive season of the year (from spring to autumn). The structure of the model shows the negative impact of shortage or surplus of nutrients in the soil. Similar patterns of abiotic effects on *Collembola* and *Acarina* species have been identified by Demšar et al. (2006), who predicted the abundance of selected species separately.

The community structure model for the Australian vegetation dataset (Fig. 5) indicates that environmental factors related to water supply and air temperature have the most severe effects. Vegetation communities exposed to water shortages and high temperature are composed of small number of taxonomic groups, while communities with good water supply show high taxonomic diversity. The indicated pattern of interaction between the abiotic factors and the vegetation community structure is expected because the sampling sites have a very large area. At such a spatial scale, the attributes related to water and temperature always have the most dominant effects on vegetation structure. If more specific data from smaller areas are collected, the structure of the model should reveal interactions that predict more specific assemblages of species adapted to the local specific growing conditions.

The structures of all three HMC models (Figs. 3–5) presented in this paper are explainable and in accordance with results of previous analyses of these datasets (Demšar et al., 2006; Kocev and Džeroski, 2013; Džeroski et al., 2000; Gjorgjioski et al., 2008; Blockeel et al., 1999). However, their predictive power is higher than the predictive power of the previous models. Accordingly, our models are of high scientific relevance in terms of their contributions to the objective determination of assembly rules for structuring the studied biological communities, and at the same time their high performance makes them more reliable.

5. Conclusions

Characterizing the relationships between environmental variables and the presence/abundance of plants and animals is one of the most fundamental tasks in ecology. It contributes towards understanding the influence of environmental factors on the structure of the community of species co-existing at a given spatial unit.

In this study, we investigate whether the information on the taxonomic rank and the multi-species aspect of community data help to learn better community structure models. We compare the models obtained by several approaches for predicting community structure in terms of their predictive performance, over-fitting score, time needed to construct the model, and the size of the model. The results show that the exploitation of the taxonomic rank for learning global or local models of community structure improves the predictive performance of the models. The use of the multi-species aspect of the data also improves the predictive performance of the models, i.e., global models have better performance than local models and over-fit less. The performance gains resulting from each of the two sources of information are comparable. Finally, the best performing method is hierarchical multi-label classification (which exploits both the taxonomic rank and the multi-species data), while the worst performing method is single-label classification (which does not exploit the taxonomic rank and the multi-species data).

The evaluation of the models efficiency revealed the following. First, the global models are much faster to construct than the local models. Second, the global models are much smaller than the local models in size, and thus are much easier to interpret. The global models provide an overview of the complete community structure, while inferring the community structure from the local models is a laborious, tedious, expensive and sometimes intractable task. For example, for the Australian vegetation data an expert would have to build and analyse the models for as many as 3172 species and then try to infer some overall conclusions. However, if the diversity of

Please cite this article in press as: Levatić, J., et al., Community structure models are improved by exploiting taxonomic rank with predictive clustering trees. Ecol. Model. (2014), http://dx.doi.org/10.1016/j.ecolmodel.2014.10.023

organisms in the observed ecosystem is so high (i.e., 3172 species) then finding and describing some general community structure from the many individual models could be impossible.

All in all, integrating the information about the taxonomic rank and multi-species data improves the predictive performance and the interpretability of the community structure models. Using both of these sources of additional information yields the best results, both in terms of predictive performance and model size.

Acknowledgments

We would like to thank *Matt White* from the Arthur Rylah Institute for Environmental Research, Department of Sustainability and Environment, Heidelberg, Victoria, Australia for providing the dataset on the Australian vegetation; *Jasna Grbović* from the Environmental Agency of Slovenia for providing the dataset on Slovenian rivers; and *Paul Henning Krogh* from the National Environmental Research Institute Roskilde, Denmark for providing the dataset on soil microarthropods.

We would also like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (grant number ICT-2013-612944).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2014.10.023.

References

Amendmends of the rules on surface water status monitoring, 2011. Offcial Gazette of the Republic of Slovenia. 81, 10417-10419.
Araújo, M.B., New, M., 2007. Ensemble forecasting of species distributions. Trends Ecol. Evol. 22 (1), 42–47.
Belyea, L.R., Lancaster, J., 1999. Assembly rules within a contingent ecology. Oikos 86 (3), 402–416.
Blockeel, H., Struyf, J., 2002. Efficient algorithms for decision tree cross-validation. J. Mach. Learn. Res. 3, 621–650.
Blockeel, H., Raedt, L.D., Ramon, J., 1998. Top-down induction of clustering trees. In: Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann, pp. 55–63.
Blockeel, H., Džeroski, S., Grbović, J., 1999. Simultaneous prediction of multiple chemical parameters of river water quality with TILDE. In: Żytkow, J., Rauch, J. (Eds.), Principles of Data Mining and Knowledge Discovery, Vol. 1704 of Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp. 32–40.
Cortet, J., Kocev, D., Ducobu, C., Džeroski, S., Debeljak, M., Schwartz, C., 2011. Using data mining to predict soil quality after application of biosolids in agriculture. J. Environ. Qual. 40 (6), 1972–1982.
Džeroski, S., Demšar, D., Grbović, J., 2000. Predicting chemical parameters of river water quality from bioindicator data. Appl. Intell. 13 (1), 7–17.
Davis, J., Goadrich, M., 2006. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240.
De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81 (11), 3178–3192.
Debeljak, M., Squire, G.R., Kocev, D., Hawes, C., Young, M.W., Džeroski, S., 2011. Analysis of time series data on agroecosystem vegetation using predictive clustering trees. Ecol. Model. 222 (14), 2524–2529.
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. Ecol. Model. 191 (1), 131–143.
Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.
Drew, C.A., Wiersma, Y.F., Huettmann, F., 2011. Predictive Species and Habitat Modelling in Landscape Ecology. Springer, New York.
Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC, J., Overton, M., Townsend Peterson, A., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 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, Cambridge.
Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19 (1), 1–67.
Götzenberger, L., de Bello, F., Brathen, K.A., Davison, J., Dubuis, A., Guisan, A., Lepš, J., Lindborg, R., Moora, M., Prtel, M., Pellissier, L., Pottier, J., Vittoz, P., Zobel, K., Zobel, M., 2012. Ecological assembly rules in plant communities – approaches, patterns and prospects. Biol. Rev. Camb. Philos. Soc. 87 (1), 111–127.
Gjorgjioski, V., Džeroski, S., White, M., 2008. Clustering Analysis of Vegetation Data, Tech. Rep. 10065. Jožef Stefan Institute.
Greve, M.H., Breuning-Madsen, H., 1999. Soil Mapping in Denmark, Tech. Rep. European Soil Bureau.
Kampichler, C., Wieland, R., Calmé, S., Weissenberger, H., Arriaga-Weiss, S., 2010. Classification in conservation biology: a comparison of five machine-learning methods. Ecol. Inf. 5 (6), 441–450.
Keddy, P.A., 1992. Assembly and response rules: two goals for predictive community ecology. J. Veg. Sci. 3 (2), 157–164.
Kocev, D., Džeroski, S., 2013. Habitat modelling with single-and multi-target trees and ensembles. Ecol. Inf. 18, 79–92.
Kocev, D., Džeroski, S., White, M., Newell, G., Griffioen, P., 2009. Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. Ecol. Model. 220 (8), 1159–1168.
Kocev, D., Naumoski, A., Mitreski, K., Krstić, S., Džeroski, S., 2010. Learning habitat models for the diatom community in lake Prespa. Ecol. Model. 221 (2), 330–337.
Kocev, D., Vens, C., Struyf, J., Džeroski, S., 2013. Tree ensembles for predicting structured outputs. Pattern Recognit. 46 (3), 817–833.
Lek, S., Scardi, M., Verdonschot, P.F., Descy, J.-P., Park, Y.-S., 2005. Modelling Community Structure in Freshwater Ecosystems. Springer, Heidelberg.
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. Biol. Conserv. 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. Environ. Model. Softw. 25 (7), 826–836.
Rules on surface water status monitoring, 2009. Offcial Gazette of the Republic of Slovenia. 10, 832-840.
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, Washington D.C.
Silla, C., Freitas, A., 2011. A survey of hierarchical classification across different application domains. Data Mining Knowl. Discov. 22 (1–2), 31–72.
Slavkov, I., Gjorgjioski, V., Struyf, J., Džeroski, S., 2010. Finding explained groups of time-course gene expression profiles with predictive clustering trees. Mol. Biosyst. 6 (4), 729–740.
Struyf, J., Džeroski, S., 2006. Constraint based induction of multi-objective regression trees. In: Bonchi, F., Boulicaut, J.-F. (Eds.), Knowledge Discovery in Inductive Databases, vol. 3933 of Lecture Notes in Computer Science. Springer, Berlin/Heidelberg, pp. 222–233.
Vens, C., Struyf, J., Schietgat, L., Džeroski, S., Blockeel, H., 2008. Decision trees for hierarchical multi-label classification. Mach. Learn. 73 (2), 185–214.
Weiher, E., Keddy, P., 2001. Ecological Assembly Rules: Perspectives, Advances, Retreats. Cambridge University Press, Cambridge.

Please cite this article in press as: Levatić, J., et al., Community structure models are improved by exploiting taxonomic rank with predictive clustering trees. Ecol. Model. (2014), http://dx.doi.org/10.1016/j.ecolmodel.2014.10.023