Development and selection of decision trees for water management: Impact of data preprocessing, algorithms and settings

Article *in* Ai Communications · November 2016

DOI: 10.3233/AIC-160711

CITATIONS
7

READS
277

4 authors, including:

Gert Everaert
Vlaams Instituut Voor De Zee
82 PUBLICATIONS **1,487** CITATIONS

SEE PROFILE

Ine Pauwels
Research Institute for Nature and Forest (INBO) Belgium
36 PUBLICATIONS **344** CITATIONS

SEE PROFILE

Peter Goethals
Ghent University
434 PUBLICATIONS **8,108** CITATIONS

SEE PROFILE

AI Communications 29 (2016) 711–723
DOI 10.3233/AIC-160711
IOS Press

Development and selection of decision trees for water management: Impact of data preprocessing, algorithms and settings

Gert Everaert [a,*], Ine Pauwels [b], Elina Bennetsen [a] and Peter L.M. Goethals [a]

[a] *Ghent University, Laboratory of Environmental Toxicology and Aquatic Ecology, Coupure Links 653, B-9000 Ghent, Belgium*

[b] *Research Institute for Nature and Forest (INBO), Kliniekstraat 25, B-1070, Brussels, Belgium*

Abstract. In the present research, we found that different preprocessing options and parameterizations of classification and regression trees alter their model fit and have a direct effect on their applicability for end-users. We found that, in terms of applicability, classification trees react different to pruning than regression trees. Indeed, in case of high pruning levels, classification focus on the extreme values of the response variable, whereas regression tree are more likely to predict the intermediate values. Furthermore, when applying cross-validation with a high number of folds, modellers are likely to find one model that outperforms the other models in terms of reliability. Models were assessed based on the determination coefficient, the percentage of Correctly Classified Instances and the Cohen's Kappa statistic for each parameterization. We found positive correlations ($R^2 > 0.70$) between the statistical criteria and we found a non-linear negative relation between the model fit and the level of pruning. Therefore, environmental modellers should make use of an exhaustive list of model parameterizations to develop and compare environmental models in a transparent and objective manner. General methodological guidelines derived from the present research may help modellers to efficiently select statistical and ecological relevant models that are meeting the needs of users. The validity of our conclusion should be further tested for other datasets and scientific domains as our findings are based on one set of freshwater data.

Keywords: Classification and regression tree, parameterization, applicability, field data

1. Introduction

Models have been increasingly used in ecology as an instrument to understand the properties of ecosystems [30]. Ecological models are representations of the real ecosystem [44] and can characterize ecosystems' responses to changing environmental conditions. As they can help to manage and preserve natural resources for future generations, ecological models have been often used in decision making [19,33,37,42]. For example, classification and regression trees [6] have been successfully implemented to quantify species-environment relations [10,13,18], to predict species occurrences [28,45,46], to predict dispersal of exotic species [1,4], to mark out protected areas [14,48], to determine ecological quality [12,31] or to produce habitat maps [40]. Classification and regression trees, are black-box models that quantitatively describe predictor-response relations. They explain variations in a response variable by splitting predictor variables at certain thresholds [10]. The trees consist of a transparent set of knowledge rules, which are deduced directly from the data, i.e. no expert knowledge is involved [10,35]. This implies that the rules found mainly depend on the quality of the information contained within the data [15]. However, ecological field data are often incomplete (missing values), noisy (containing errors or outliers) or have a skewed distribution [19]. As such, although it may result in loss of information [21], data preprocessing is an essential part of the statistical analysis when interpreting ecological field data by means of classification and regression trees [27,50].

Different parameterizations of the same modelling technique may produce considerably different outcomes [2], which on his turn has consequences for the applicability of the models for the end-users. Classification trees can be large and difficult to interpret [7],

*Corresponding author. Tel.: +32 92 64 38 97; E-mail: gert.everaert@ugent.be.

but little research has been undertaken to increase their comprehensibility for end-users. Although it has already been suggested by Araujo and Guisan [2] and Goethals et al. [26], only few publications are available in literature [17,32,45] that use multiple implementations of the same technique to assess and compare model predictions and performances. Indeed, classification and regression trees are often induced via a trial-and-error approach, rules of thumb or the software's default settings. The parameterization of the algorithm used to develop the trees, such as the selection of the number of cross-validations, the degree of pruning and the minimum number of observations per leaf is often done extemporaneously. For example, 10-fold cross-validation has become the standard [49], but it has been already shown that the optimal number of folds depends on the size of the dataset and the aim of the modelling process [38]. Moreover, the selection of the final model is often driven by the best model fit and highest statistical reliability [3]. However, maximizing statistical indicators does not always result in the most optimal model in terms of applicability by end-users [19,34]. For example, Osei-Bryson [39] stated that when only assessing the statistical reliability, a decision tree with an accuracy rate of 0.959 and 29 leaves would be selected over a decision tree with a rate of 0.958 and 5 leaves. Therefore, apart from the statistical reliability, also the comprehensibility and applicability of the decision trees merit attention in the assessment of the end-product [17,47]. To date, the interplay between model parameterization, model fit and applicability for end-users remains understudied [17,39].

A better understanding is needed of why and when different parameterizations of the same modelling technique provide different results [2]. In the present research we aim: (1) to study the effect of model parameterization on the model fit of classification and regression trees; and (2) to illustrate consequences of the model parameterization on the applicability of the models for end-users. To address these research questions, we assessed the effect of 335 unique model parameterization on the model fit and on their applicability for end-users. The classification and regression trees were based on 863 field samples in the freshwater environment and aimed to relate the ecological water quality (assessed based on the macroinvertebrate community) with the physical-chemical water quality.

2. Material and methods

The methodology of the present research is illustrated in a conceptual diagram (Fig. S1). This scheme consists of general methodological guidelines for environmental data mining. Multiple conceptual schemes have been introduced earlier [23,29], but in the present research focus is on the linkage between the data preprocessing, model parameterization, and the applicability of the models for end-users.

2.1. Data

The dataset was compiled in the scope of the Water Framework Directive (WFD), which is one of the most important guidelines for European river managers. The main objective of the WFD is to reach a good ecological status of the European water courses by 2027 [16]. Currently, only a limited number of Flemish surface waters comply with this request. Therefore, the Flemish government, which is responsible for the environmental management in the northern part of Belgium, planned several rehabilitation projects to restore the ecological river quality. However, the impact of these restoration actions on the ecological status of the watercourses could not be quantified yet. Therefore, classification and regression trees were constructed to gain insight into the relations between chemical and ecological surface water conditions and, in the end, to help river manager decide where to allocate the limited resources for ecological restoration. Overall, the aim of the classification and regression trees is to characterize the ecological water quality (assessed based on the macroinvertebrate community) using the surrounding abiotic conditions, such as for example, physical-chemical variables.

The dataset contained 863 samples comprising of the ecological water quality based on the macroinvertebrate community and the physical-chemical conditions of Flemish water courses. These data cover various points (sampling locations in large streams, small rivers and large rivers) and several years (from 1989 to 2009) (Table 1). The physical-chemical data were available in the form of statistical derivatives and were calculated over one year. The physical-chemical data available were, in line with Schneiders et al. [43] and Everaert et al. [20]; maximum Biological Oxygen Demand (BOD_{max}, mg O_2/L), maximum Chemical Oxygen Demand (COD_{max}, mg O_2/L), median Kjeldahl nitrogen concentration (KjN_{med}, mg N/L), median ni-

Table 1

Summary of the original (A) and stratified (B) dataset showing the first quartile, median, mean and third quartile value based on 863 (A) and 240 (B) samples for the Ecological Quality ratio (EQR), maximum Biological Oxygen Demand (BOD_{max}), maximum Chemical Oxygen Demand (COD_{max}), median Kjeldahl nitrogen concentration (KjN_{med}), median nitrate concentration ($NO_{3,\ med}$), minimum oxygen concentration (DO_{min}), average orthophosphate concentration ($oPO_{4,\ avg}$) and average total phosphorous concentration (Pt_{avg})

	First quartile	Median	Mean	Third quartile	Unit
(A) Original dataset (863 samples)					
EQR	0.25	0.35	0.40	0.55	–
BOD_{max}	4.6	7.0	8.9	11.0	mg O_2/L
COD_{max}	39	53	63	78	mg O_2/L
KjN_{med}	1.75	2.60	2.98	4.00	mg N/L
$NO_{3,\ med}$	2.18	3.66	3.78	5.05	mg N/L
DO_{min}	2.40	3.80	3.80	5.09	mg O_2/L
$oPO_{4,\ avg}$	0.12	0.31	0.38	0.52	mg P/L
Pt_{avg}	0.30	0.54	0.63	0.81	mg P/L
	bad	poor	moderate	good_high	
Prevalence of EQR_{class}	249	385	169	60	–
(B) Stratified dataset (240 samples)					
EQR	0.28	0.53	0.50	0.71	–
BOD_{max}	3.9	6.6	7.7	10.0	mg O_2/L
COD_{max}	38	52	59	70	mg O_2/L
KjN_{med}	1.03	2.24	2.47	3.40	mg N/L
$NO_{3,\ med}$	2.04	3.39	3.69	5.05	mg N/L
DO_{min}	2.63	4.06	4.02	5.60	mg O_2/L
$oPO_{4,\ avg}$	0.09	0.21	0.33	0.44	mg P/L
Pt_{avg}	0.22	0.42	0.53	0.72	mg P/L
	bad	poor	moderate	good_high	
Prevalence of EQR_{class}	60	60	60	60	–

trate concentration ($NO_{3,\ med}$, mg N/L), minimum dissolved oxygen concentration (DO_{min}, mg O_2/L), average orthophosphate concentration ($oPO_{4,\ avg}$, mg P/L) and average total phosphorous concentration (Pt_{avg}, mg P/L). All substances were analysed in accordance to the standards of ISO 17025. The ecological status of the surface waters is assessed based on the macroinvertebrate community as discussed by Gabriels et al. [25] and is quantified using an Ecological Quality Ratio (EQR), ranging from 0 to 1. In the context of the WFD, and for transparency towards decision makers, these continuous scores are converted to five ecological quality classes ("bad, "poor", "moderate", "good" and "high"). In the present research, the quality classes 'good' and 'high' were combined in one class, named "good_high". In the data set all variables were quantified, i.e. there were no missing values. A summary of the data set can be found in Table 1. The Pearson correlation coefficients were calculated to explore the relations between physical-chemical characteristics and the ecological quality (Table 2).

2.2. Data preprocessing

We compiled two sets of data consisting of physical-chemical information and the ecological status of the watercourses. One set of data was identical to the description above, thus consisted of all 863 samples and had a skewed distribution of the response variable (Table 1). The second set of data was sub-sampled from the first set, and each class of the response variable (i.e. "bad", "poor", "moderate", "good_high") was presented in the same proportion. To do so, we identified the number of cases in the least represented ecological water quality class and selected the same amount of cases randomly in the other quality classes. By doing so, the skewed distribution of the response variable was removed and all quality classes (i.e. "bad", "poor", "moderate", "good_high") were equally represented in the second set of data. The second set of data, i.e. the stratified dataset, contained 240 (= 4 classes × 60 samples/class) out of 863 samples (Table 1). Both sets of data were used to develop classification and regression

Table 2
Correlation matrix of the original dataset showing the Pearson correlation between the Ecological Quality ratio (EQR), maximum Biological Oxygen Demand (BOD_{max}), maximum Chemical Oxygen Demand (COD_{max}), median Kjeldahl nitrogen concentration (KjN_{med}), median nitrate concentration ($NO_{3,\,med}$), minimum oxygen concentration (DO_{min}), average orthophosphate concentration ($oPO_{4,\,avg}$) and average total phosphorous concentration (Pt_{avg})

	EQR	BOD_{max}	COD_{max}	KjN_{med}	$NO_{3,\,med}$	DO_{min}	$oPO_{4,\,avg}$	Pt_{avg}
EQR	1.00	−0.38	−0.32	−0.64	−0.11	0.56	−0.54	−0.58
BOD_{max}		1.00	0.63	0.46	−0.10	−0.42	0.44	0.53
COD_{max}			1.00	0.38	−0.06	−0.33	0.37	0.48
KjN_{med}				1.00	−0.03	−0.61	0.73	0.74
$NO_{3,\,med}$					1.00	0.05	0.10	−0.01
DO_{min}						1.00	−0.55	−0.56
$oPO_{4,\,avg}$							1.00	0.86
Pt_{avg}								1.00

trees aiming to characterize the ecological water quality based on the physical-chemical conditions (see further).

2.3. Model development

The aim of the classification and regression trees was to quantify the ecological status based on the physical-chemical conditions and to assess the importance of the model parameterization on the applicability of the resulting models for end-users. The basic idea of classification and regression trees (CART) is very simple, i.e. predicting a response variable (e.g. Y) from predictor variables (e.g. $X_1, X_2, \ldots, X_p$). To do so, a tree is grown that is constructed by splitting a node into child nodes repeatedly, beginning with the root node that contains the whole learning sample. To decide on which split is selected all the possible splits at each node are considered and the one that results in child nodes that are the "purest" is selected. In this selection procedure univariate splits are considered, i.e. each split depends on the value of one predictor variable.

Classification and regression trees were produced for 335 unique parameterizations of the number of cross-validations (n_{cv}), the minimum number of observations (min_{object}) and the complexity parameter (cp_{value}). Cross-validation has been very valuable in environmental modelling because all data are used to train and validate the model [10,26]. In k-fold cross-validation, k subsets of equal size are randomly generated from the original dataset. In the first round of the cross-validation, a single subset is retained as the validation data for testing the models, and the remaining $k - 1$ subsets are used as training data. This process is then repeated k times, with each of the subsets used exactly once for validation. The result of a k-fold cross-validation are k models which are combined to produce a single overall performance estimation [18]. Usually, between 10 and 50% of the available observations are used for validation purposes [32]. The number of cross validations is quantified using the n_{cv} parameter. Large, complex classification and regression trees tend to overfit the data and overestimate the predictive power of the model [22,49]. Pruning, or the removal of knowledge rules that are not beneficial according to the complexity parameter, prevents the model from overfitting [22]. The degree of pruning of a decision tree is regulated based on the complexity parameter (cp_{value}). An alternative and more pragmatic way to prune a tree is by means of the minimum number of observations (min_{object}). The latter parameter quantifies the minimum number of observations that must exist in a node in order for a split to be attempted.

In the present research, the n_{cv} varied from 2, 4, 6, 8, 10, 20, 40, 80; the cp_{value} varied from 0.01, 0.05, 0.10, 0.15, 0.20, 0.25 and min_{object} varied from 2, 4, 6, 8, 10, 20, 40. To assess the model fit, we quantified the determination coefficient (R^2), the percentage of Correctly Classified Instances (CCI) and the Cohen's Kappa statistic (K) for each unique parameterization. The CCI was calculated as the percentage of true positive and true negative predictions. The K measured the percentage of true positive and true negative predictions, but adjusted these values for the amount of agreement that could be expected due to chance [9,28]. Values for K are between 0 and 1, a value close to 1 indicates a better model prediction. The CCI should reach at least 70% in order to have a satisfactory model performance [24]. Additionally, the model complexity was quantified as the number of nodes embedded in the model [8] and further referred to as $depth_{tree}$.

For each of the 335 unique parameterizations, we stored the mean, standard deviation and maximum value of the R^2, CCI, K and $\text{depth}_{\text{tree}}$. Pairplots visualized the univariate patterns between the model parameterizations (n_{cv}, $\text{min}_{\text{object}}$ and cp_{value}) and the performance criteria (R^2, CCI and K) [51]. Each dot in the pairplots deals with one unique model parameterization, so in total 335 dots are shown per panel. The effect of sub-sampling on the performance criteria was assessed using a paired t-test (with $\alpha = 0.05$). All calculations were performed in R [41]. The R package *rpart* was implemented to develop the classification and regression trees (Figs S2–S4).

2.4. Simulations

A selection of the 335 models that were developed in the previous section were used to simulate future ecological water quality under different management scenarios. We selected those classification trees that were developed using 5-fold cross-validation ($n_{\text{cv}} = 5$) and three different values of the complexity parameters ($\text{cp}_{\text{value}} = 0.01, 0.05$, and 0.15). Each classification tree was built five times and the data were reshuffled in between each run. As such, each of the settings tested (three in total), resulted in five different classification trees. Next, we compiled an external data set using the water quality model 'Planification Et Gestion de l'ASsainissement des Eaux' (PEGASE), French acronym for "Planning and management of water purification". PEGASE simulates the effect of the proposed river restoration actions on the physical-chemical water quality [11]. In the present research, by implementing the classification and regression trees on the PEGASE data, the effects of the restoration actions on the physical-chemical water quality were translated towards their effect on the ecological water quality. The aim of these simulations is to illustrate the impact of different model parameterization on the applicability of the resulting classification trees for end-users. More information on the PEGASE data can be found in Deliege et al. [11] and Boets et al. [5].

2.5. Regression tree analyses

We compared whether regression trees do react similarly (in terms of model fit and applicability) to changing model parameterization as classification trees. To do so, the 335 parameterizations identical to those descripted earlier were used to build regression trees (cfr. *Model development*). The simulations, used to assess the impact of the model parameterization on the applicability were performed in the same way as for classification trees (cfr. *Simulations*). As our main aim was to verify whether regression trees react in the same way as classification trees in terms of applicability, focus in the results and discussion will be on the impact of the model parameterization on the model applicability. The pairplots between the model parameterizations and the model fit are shown in supportive information.

3. Results and discussion

3.1. Effect of stratification

Stratification results in decreasing quartile values of the environmental variables which are negatively correlated with the ecological water quality (i.e. BOD_{max}, COD_{max}, KjN_{med}, etc.). However for the DO_{min} that is positively related with the ecological water quality, the quartile values increase (Table 1). In terms of model fit, as quantified based on the R^2, CCI and K, the stratified set of data (i.e. with 240 samples) result in an improved model fit compared to the original dataset (i.e. with 863 samples). Indeed, the R^2 increased from 0.52 ± 0.06 to 0.67 ± 0.03 ($p < 0.001$) and the K increased from 0.18 ± 0.12 to 0.29 ± 0.08 ($p < 0.001$) (Table 3). The CCI did not change significantly (Table 3). These findings corroborate with those of McPherson et al. [36] and Araujo and Guisan [2], who found that the best classification results are obtained when each re-

Table 3

Classification trees were developed for 335 unique model parameterizations. For each model parameterizations we quantified the model fit based on the determination coefficient (R^2), the correctly classified instances (CCI), the Kappa statistic (K). The number of nodes in the tree ($\text{depth}_{\text{tree}}$) was a measure of the complexity of the tree. Models were constructed for both the original dataset (863 samples) and the stratified dataset (240 samples)

	Min.	Max.	Mean ± SD
Initial dataset			
$\text{depth}_{\text{tree}}$	0.00	3.00	1.11 ± 0.81
R^2	0.42	0.60	0.52 ± 0.06
K	0.00	0.34	0.18 ± 0.12
CCI	0.45	0.59	0.52 ± 0.04
Stratified dataset			
$\text{depth}_{\text{tree}}$	0.50	3.00	1.67 ± 0.74
R^2	0.56	0.74	0.67 ± 0.03
K	0.08	0.41	0.29 ± 0.08
CCI	0.25	0.57	0.47 ± 0.08

sponse variables' category is equally represented. Indeed, McPherson et al. [36] state that a prevalence of 50% is optimal when predicting presence-absence data. So, if we pursue this reasoning, four quality classes are best predicted if each class account for 25% of the dataset. As such, every class of the response variable has an equal chance to be selected in the cross-validation procedure. Araujo and Guisan [2] refer to stratification as sub-sampling and define it as the selection of samples from a larger set to reduce an existing bias. Overall, we found that stratification is beneficial for the model fit (Table 3), which is in line with results of Everaert et al. [19].

3.2. *Effect of model parameterization on model fit*

R^2, CCI and K were used to quantify the statistical reliability of classification trees and were positively correlated with each other. The correlation between the R^2 and K, and between R^2 and CCI was 0.7 (Fig. 1). The correlation between CCI and K was 1.0 (Fig. 1). We found that the smaller the size of the cross-validation folds (i.e. high n_{cv}), the higher the variability of the performance criteria (Fig. 2). Indeed, the n_{cv} is positively related with: (1) the standard deviation of the R^2 (0.8); (2) the standard deviation of K (0.7); and (3) the standard deviation of the CCI (0.9; Fig. 2). Furthermore, we found that the smaller the size of the cross-validation folds, the higher the probability to hit upon an excellent model (Fig. 3). The latter is illustrated in Fig. 3 showing the maximum R^2, CCI and K in function of the parameterizations. Note the positive relation between the n_{cv} and R^2 (0.9), between n_{cv} and K (0.7), and between n_{cv} and the CCI (0.8) (Fig. 3). Furthermore, we found that higher values of the complexity parameter (cp_{value}) result in more simple models (i.e. relation of -0.9 between the cp_{value} and $depth_{tree}$; Fig. 1). The latter is visualized in two examples of classification trees (Fig. S5).

Although low pruning levels (i.e. low cp_{value}) often result in complex trees, with a better model fit, the relation between the model complexity and model fit is not linearly increasing. Indeed, at a certain level of complexity adding even more complexity is no longer beneficial for the model fit (Fig. 1). When the complexity parameter (cp_{value}) increases from 0.01 and 0.10, average performance criteria are stable, but the model complexity ($depth_{tree}$) declines. This indicates that less complex models may result in similar statistical reliabilities (Fig. 1). This finding has important consequences as large and complex classification trees tend to overfit the data and overestimate the predictive power of the model [22,49]. To prevent the model from overfitting, an optimal balance must be found between the statistical performance and the complexity of the model (Fig. 1). To do so, pruning, or the removal of knowledge rules, is often applied. Most optimal pruning levels can be identified if a series of values of the complexity parameter are implemented. An ad hoc method for selecting the complexity parameter seems efficient in the short run but it is likely to pick out an unrepresentative model [10]. Overall, based on an exhaustive list of model parameterizations, general trends can be derived and the optimal pruning level selected. In the present research, optimal pruning levels, parameterized by means of the cp_{value}, varied around 0.1 (Fig. 1). In case that more pruning is applied, values of CCI, K and R^2 decrease as they are negatively correlated with the CCI (-0.8), K (-0.8), and R^2 (-0.4) (Fig. 1.) Finally, we found only a weak influence of the minimum number of observations per leaf (min_{object}) on the model fit (Figs 1, 2 and 3). The latter probably relates to the fact that the pruning level (cp_{value}) was more important than the min_{object}.

3.3. *Effect of model parameterization on model applicability*

Previous paragraphs dealt with technical aspects of the modelling process. However, model evaluation involves more than comparing predictions with observations and calculating statistical performance criteria [17,34]. The link between the derived models and their applicability is often forgotten. We found that classification trees based on the stratified dataset and at low pruning levels (i.e. with few knowledge rules deleted) are most complex (Fig. 1), but succeed to predict all four water quality classes (Fig. 4(A)). At high pruning levels (i.e. several knowledge rules are deleted) the classification trees tend to predict extreme values (Fig. 4(C)). Indeed, when cp_{value} is low (i.e. 0.01) all four water quality classes can be predicted. However, in case that cp_{value} increases to 0.15 only the bad and good_high water quality classes could be predicted (i.e. each with a prevalence of ca. 50%). Also note that the standard deviation of the prediction increases, with increasing levels of pruning (Fig. 4).

Although it has already been suggested by Araujo and Guisan [2] and Goethals et al. [26], it rarely happens that researchers use multiple implementations of the same technique to assess and compare model pre-

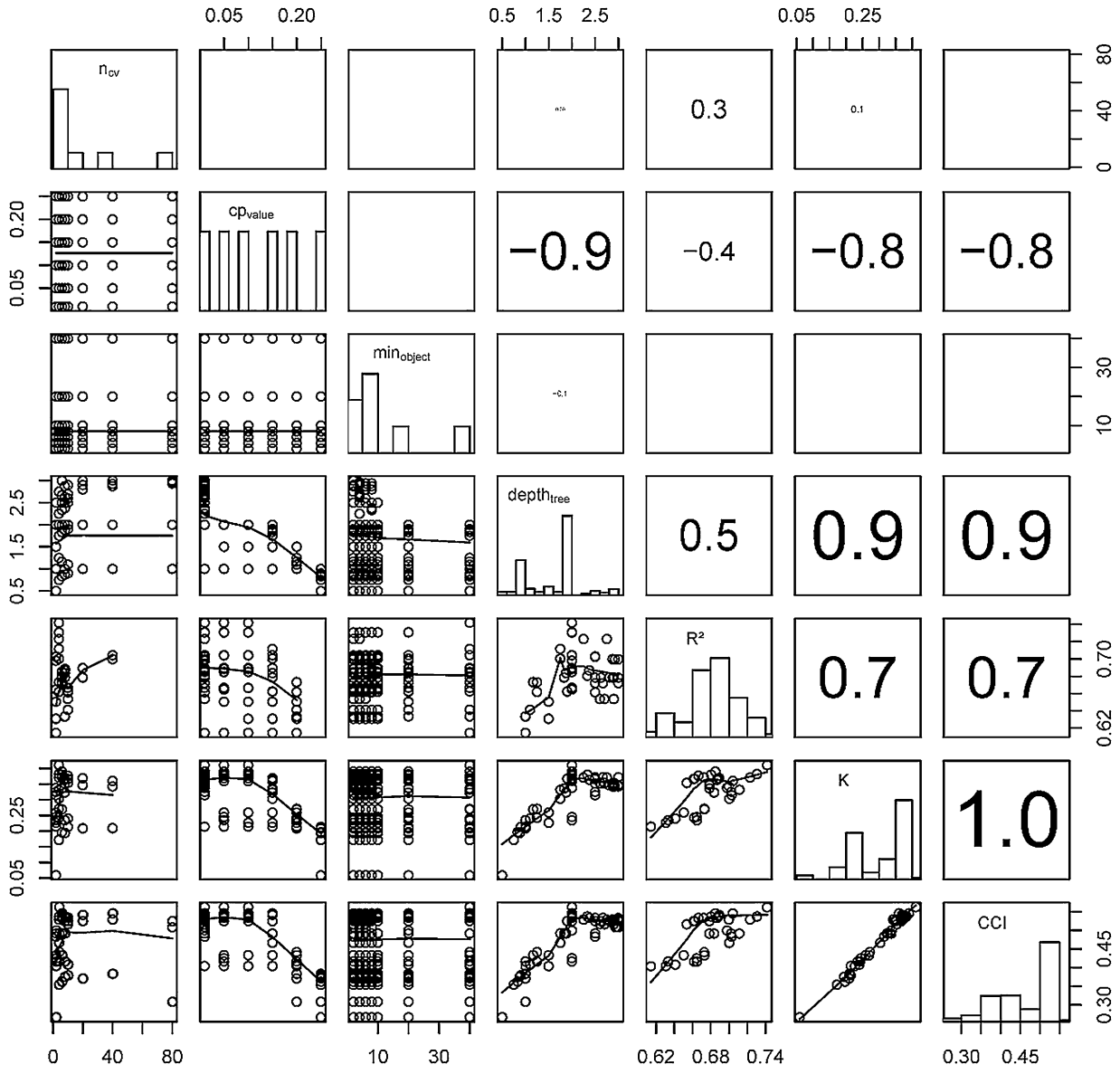

Fig. 1. Pairplot showing univariate interactions between the model development settings (n_{cv} = number of cross-validations, cp_{valuee} = complexity parameter, min_{object} = minimum number of observations per leaf) and mean statistical criteria for classification trees, based on the stratified dataset. The model fit of each parameterization (i.e. 335 in total) was quantified based on the determination coefficient (R^2), the correctly classified instances (CCI), the Kappa statistic (K). The number of nodes in the tree ($depth_{tree}$) was a measure of the complexity of the tree. Each panel in the pairplot consists of 335 dots and each dot corresponds to one unique model parameterization. In the lower left panels we visualize the relation between the model parameterizations and the corresponding CCI, K and R^2. The upper right panels quantify these relations by means of a correlation analysis.

dictions and performances. One example from earlier years is the research of Kozak and Kozak [32] in which discrete steps are used to put aside observations for model validation. Recently, Tirelli and Pessani [45] varied the complexity parameter transparently from 0.15 to 0.25 and statistically compared the performance of unpruned and pruned models. However, influence of other model development settings was not discussed in detail.

In the present study, models that could not predict a good ecological water quality were, unesteemed their predictive performance, not applicable by end-

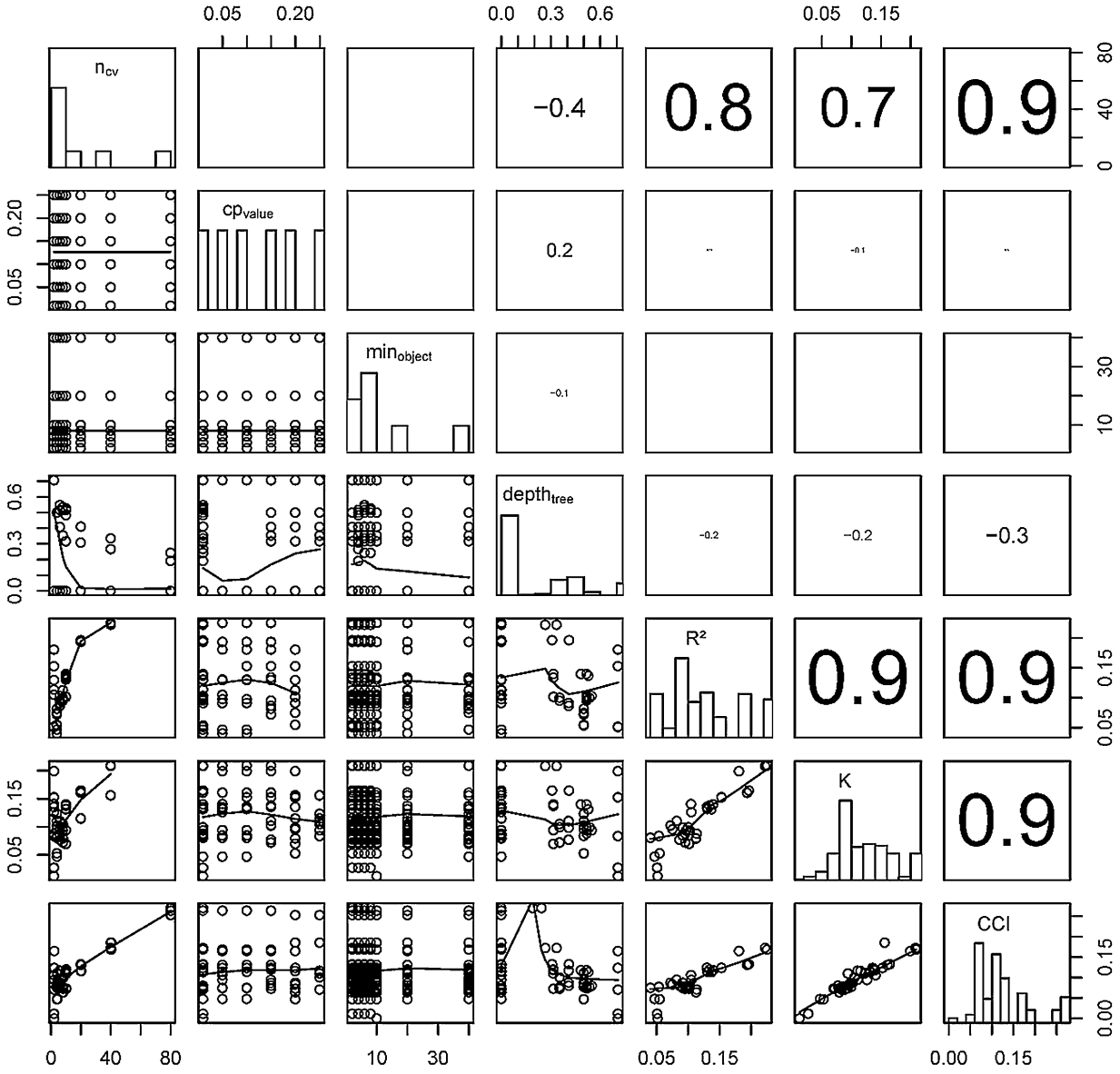

Fig. 2. Pairplot showing univariate interactions between the model development settings (n_{cv} = number of cross-validations, cp_{valuee} = complexity parameter, min_{object} = minimum number of observations per leaf) and the standard deviation of the statistical criteria for classification trees, based on the stratified dataset. The model fit of each parameterization (i.e. 335 in total) was quantified based on the determination coefficient (R^2), the correctly classified instances (CCI), the Kappa statistic (K). The number of nodes in the tree ($depth_{tree}$) was a measure of the complexity of the tree. Each panel in the pairplot consists of 335 dots and each dot corresponds to one unique model parameterization. In the lower left panels we visualize the relation between the model parameterizations and the corresponding CCI, K and R^2. The upper right panels quantify these relations by means of a correlation analysis.

users (cfr. *Data*). In this context, data stratification was very helpful to ensure that our model could predict the whole biological quality range. Without this technique it would have been difficult to predict the shift from the moderate to the good biological water quality and consequently to help stakeholders to decide which was the most efficient restoration plan (cfr. *Data*). These findings support the conclusions formulated by Larocque et al. [34] stating that other quality aspects than statistical reliability are equally important in the model selection. Although a trial-and error approach seems more efficient in the short run, select-

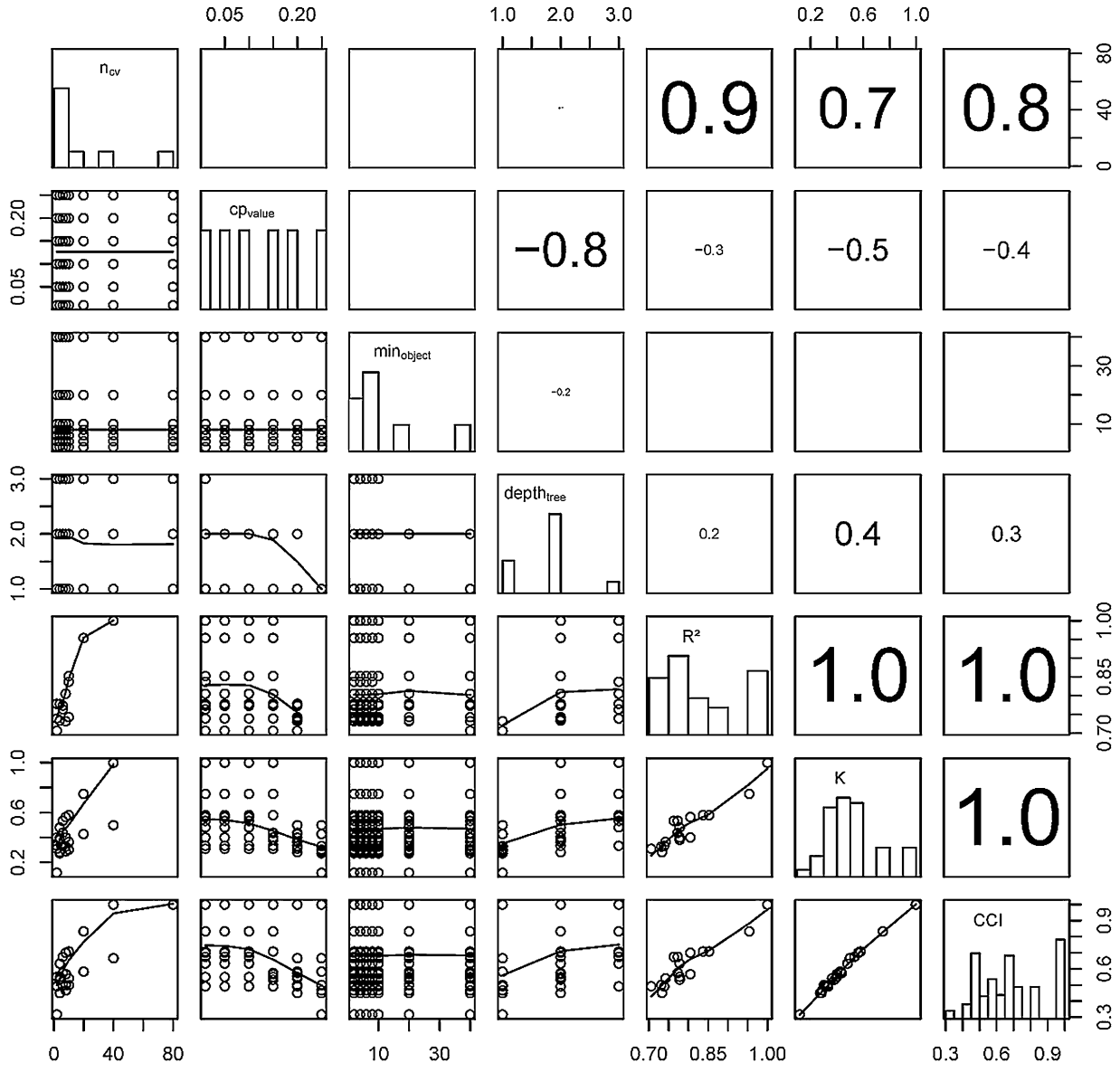

Fig. 3. Pairplot showing univariate interactions between the model development settings (n_{cv} = number of cross-validations, cp_{valuee} = complexity parameter, min_{object} = minimum number of observations per leaf) and the maximum value of the statistical criteria for classification trees, based on the stratified dataset. The model fit of each parameterization (i.e. 335 in total) was quantified based on the determination coefficient (R^2), the correctly classified instances (CCI), the Kappa statistic (K). The number of nodes in the tree ($depth_{tree}$) was a measure of the complexity of the tree. Each panel in the pairplot consists of 335 dots and each dot corresponds to one unique model parameterization. In the lower left panels we visualize the relation between the model parameterizations and the corresponding CCI, K and R^2. The upper right panels quantify these relations by means of a correlation analysis.

ing a model that per accident performs well, is less transparent than comparing a list of models in an automated, consistent and transparent way. Stability cannot be guaranteed if a trial-and error approach is applied. So, different model parameterizations may improve the understanding of the sensitivity of models and allow more robust models comparisons [2]. Overall, we found that model parameterization does not only alter the model fit, but also the conclusions drawn from and the applicability of environmental models for end-users.

The effect of stratification (Table S1) and the effect of the model parameterizations on the model fit (Figs S2–S4) is similar for regression trees as for clas-

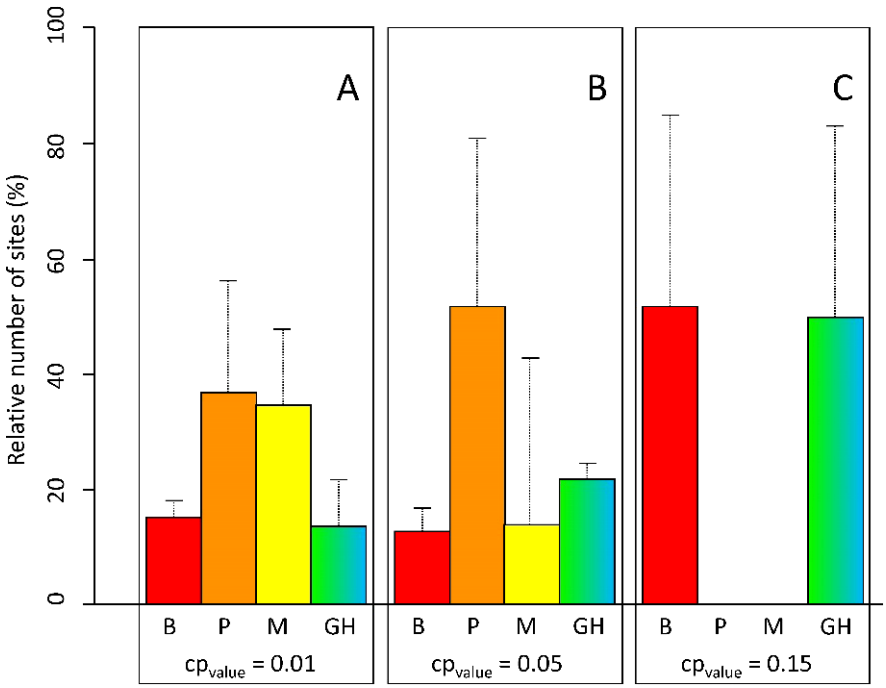

Fig. 4. Visualisation of the predicted ecological water quality in Flanders in 2027 based on classification trees. Four ecological water quality classes are shown, being bad (B, red), poor (P, orange), moderate (M, yellow) and good_high (GH, gradient green to blue). Classification trees were developed with 5-fold cross-validation and three complexity parameters (cp_{value}) were applied; 0.01 (A), 0.05 (B), and 0.15 (C). Each model was built 5 times, average and standard deviation of the predictions per fold were calculated.

sification trees. The main difference between classification trees and regression trees relates to the consequences of the model parameterizations on the applicability of the regression trees (Fig. 5(A)–(C)). In case of low pruning ($cp_{value} = 0.01$) or moderate pruning ($cp_{value} = 0.05$), results are similar between regression and classification trees (Fig. 4(A) and 4(B); Fig. 5(A) and (B)). Indeed, similar as for classification trees all four water quality classes can be predicted if low or moderate pruning is applied. However, in case high pruning levels ($cp_{value} = 0.15$), regression trees tend to focus on the intermediate water quality (poor, moderate) (Fig. 5(C)). By contrast, classification trees focus on the extreme water quality classes at high pruning levels (Fig. 4(C)).

4. Conclusion

The main objective of this study was to address the knowledge gap regarding the relation between model parameterization, model fit and applicability for end-users. We show that data stratification, number of cross-validation folds and pruning may impact the classification trees' reliabilities and, more importantly, alter the applicability of, and the conclusions drawn from, environmental models. We found that three statistical criteria were positively correlated and that there is a non-linear trade- off between the pruning of the model and the model fit. Based on our findings environmental modellers should be stimulated to develop models in a systematic way and trained with well-defined parameter settings to guarantee reliable, stable and reproducible models. Our findings should be further tested for other datasets and scientific domains.

Supplementary data

Online supplement consisting of supplementary Figures S1–S5 and Table S1 is available at: http://dx.doi.org/10.3233/AIC-160711.

Acknowledgements

Gert Everaert is supported by a post-doctoral fellowship from the Special Research Fund of Ghent Univer-

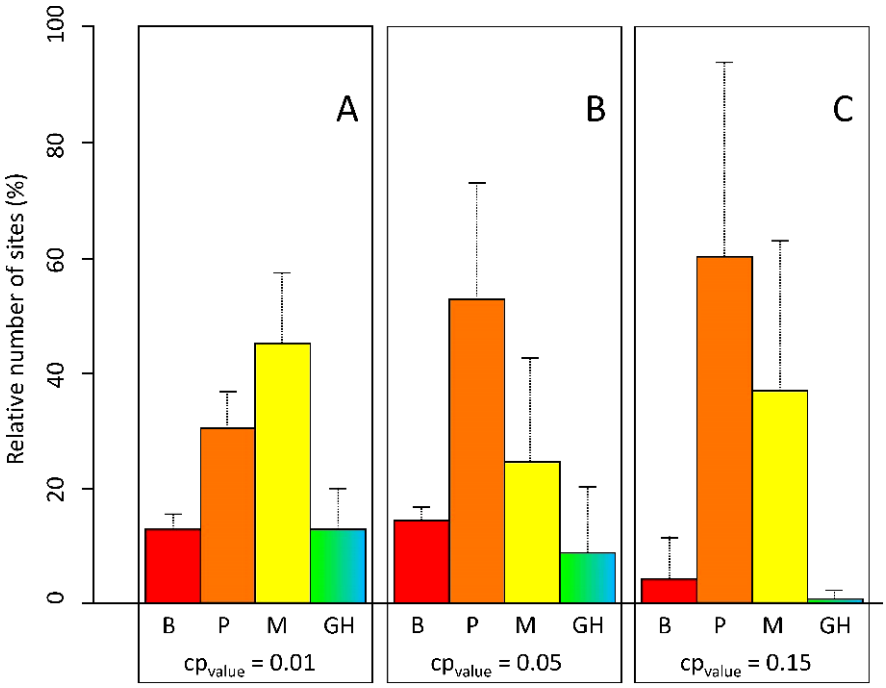

Fig. 5. Visualisation of the predicted ecological water quality in Flanders in 2027 based on regression trees. Four ecological water quality classes are shown, being bad (B, red), poor (P, orange), moderate (M, yellow) and good_high (GH, gradient green to blue). Regression trees were developed with 5-fold cross-validation and three complexity parameters (cp_{value}) were applied; 0.01 (A), 0.05 (B), and 0.15 (C). Each model was built 5 times, average and standard deviation of the predictions per fold were calculated.

sity (BOF15/PDO/061). Elina Bennetsen is supported by the Special Research Fund of Ghent University (B/09924/02).

References

[1] M.E. Andrew and S.L. Ustin, Habitat suitability modelling of an invasive plant with advanced remote sensing data, *Diversity and Distributions* **15** (2009), 627–640. doi:10.1111/j.1472-4642.2009.00568.x.

[2] M.B. Araujo and A. Guisan, Five (or so) challenges for species distribution modelling, *Journal of Biogeography* **33** (2006), 1677–1688. doi:10.1111/j.1365-2699.2006.01584.x.

[3] N.D. Bennett, B.F.W. Croke, G. Guariso, J.H.A. Guillaume, S.H. Hamilton, A.J. Jakeman, S. Marsili-Libelli, L.T.H. Newham, J.P. Norton, C. Perrin, S.A. Pierce, B. Robson, R. Seppelt, A.A. Voinov, B.D. Fath and V. Andreassian, Characterising performance of environmental models, *Environmental Modelling & Software* **40** (2013), 1–20. doi:10.1016/j.envsoft.2012.09.011.

[4] P. Boets, K. Lock, M. Messiaen and P.L.M. Goethals, Combining data-driven methods and lab studies to analyse the ecology of *Dikerogammarus villosus*, *Ecological Informatics* **5** (2010), 133–139. doi:10.1016/j.ecoinf.2009.12.005.

[5] P. Boets, I.S. Pauwels, K. Lock and P.L.M. Goethals, Using an integrated modelling approach for risk assessment of the 'killer shrimp' *Dikerogammarus villosus*, *River Research and Applications* **30** (2014), 403–412. doi:10.1002/rra.2658.

[6] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, 1984.

[7] L.A. Breslow and D.W. Aha, Simplifying decision trees: A survey, *Knowledge Engineering Review* **12** (1997), 1–40. doi:10.1017/S0269888997000015.

[8] C. Cappelli, F. Mola and R. Siciliano, A statistical approach to growing a reliable honest tree, *Computational Statistics & Data Analysis* **38** (2002), 285–299. doi:10.1016/S0167-9473(01)00044-5.

[9] J. Cohen, A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20** (1960), 37–46. doi:10.1177/001316446002000104.

[10] G. De'ath and K.E. Fabricius, Classification and regression trees: A powerful yet simple technique for ecological data analysis, *Ecology* **81** (2000), 3178–3192. doi:10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.

[11] J.F. Deliege, E. Everbecq, P. Magermans, A. Grard, T. Bourouag, C. Blockx and J. Smitz, PEGASE, an integrated river/basin model dedicated to surface water quality assessment: Application to cocaine, *Acta Clinica Belgica* **65** (2010), 42–48. doi:10.1179/acb.2010.108.

[12] L. Dominguez-Granda, K. Lock and P.L.M. Goethals, Using multi-target clustering trees as a tool to predict biological water quality indices based on benthic macroinvertebrates and environmental parameters in the Chaguana watershed (Ecuador), *Ecological Informatics* **6** (2011), 303–308. doi:10.1016/j.ecoinf.2011.05.004.

[13] S. Dzeroski and D. Drumm, Using regression trees to identify the habitat preference of the sea cucumber (*Holothuria leucospilota*) on Rarotonga, Cook Islands, *Ecological Modelling* **170** (2003), 219–226. doi:10.1016/S0304-3800(03)00229-1.

[14] T.C. Edwards, D.R. Cutler, N.E. Zimmermann, L. Geiser and G.G. Moisen, Effects of sample survey design on the accuracy of classification tree models in species distribution models, *Ecological Modelling* **199** (2006), 132–141. doi:10.1016/j.ecolmodel.2006.05.016.

[15] I. El-Baroudy, A. Elshorbagy, S.K. Carey, O. Giustolisi and D. Savic, Comparison of three data-driven techniques in modelling the evapotranspiration process, *Journal of Hydroinformatics* **12** (2010), 365–379. doi:10.2166/hydro.2010.029.

[16] EU, Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community, action in the field of water policy, 2000.

[17] G. Everaert, E. Bennetsen and P.L.M. Goethals, An applicability index for reliable and applicable decision trees in water quality modelling, *Ecological Informatics* **32** (2016), 1–6. doi:10.1016/j.ecoinf.2015.12.004.

[18] G. Everaert, P. Boets, K. Lock, S. Dzeroski and P.L.M. Goethals, Using classification trees to analyze the impact of exotic species on the ecological assessment of polder lakes in Flanders, Belgium, *Ecological Modelling* **222** (2011), 2202–2212. doi:10.1016/j.ecolmodel.2010.08.013.

[19] G. Everaert, I.S. Pauwels, P. Boets, E. Verduin, M.A.A. de la Haye, C. Blom and P.L.M. Goethals, Model-based evaluation of ecological bank design and management in the scope of the European Water Framework Directive, *Ecological Engineering* **53** (2013), 144–152. doi:10.1016/j.ecoleng.2012.12.034.

[20] G. Everaert, I.S. Pauwels and P.L.M. Goethals, Development of data-driven models for the assessment of macroinvertebrates in rivers in Flanders, in: *5th Biennial Meeting of the International Congress on Environmental Modelling and Software (iEMSs 2010): Modelling for Environment's Sake International Environmental Modelling and Software Society (iEMSs)*, D.A. Swayne, W. Yang, A.A. Voinov, A. Rizzoli and T. Filatova, eds, Ottawa, ON, Canada, 2010.

[21] F. Famili, W.-M. Shen, R. Weber and E. Simoudis, Data preprocessing and intelligent data analysis, *International Journal on Intelligent Data Analysis* **1** (1997), 1–28. doi:10.1016/S1088-467X(98)00006-7.

[22] D. Fierens, J. Ramon, H. Blockeel and M. Bruynooghe, A comparison of pruning criteria for probability trees, *Machine Learning* **78** (2010), 251–285. doi:10.1007/s10994-009-5147-1.

[23] M.A.E. Forio, D. Landuyt, E. Bennetsen, K. Lock, T.H.T. Nguyen, M.N.D. Ambarita, P.L.S. Musonge, P. Boets, G. Everaert, L. Dominguez-Granda and P.L.M. Goethals, Bayesian belief network models to analyse and predict ecological water quality in rivers, *Ecological Modelling* **312** (2015), 222–238. doi:10.1016/j.ecolmodel.2015.05.025.

[24] W. Gabriels, P.L.M. Goethals, A.P. Dedecker, S. Lek and N. De Pauw, Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networks, *Aquatic Ecology* **41** (2007), 427–441. doi:10.1007/s10452-007-9081-7.

[25] W. Gabriels, K. Lock, N. De Pauw and P.L.M. Goethals, Multimetric Macroinvertebrate Index Flanders (MMIF) for biological assessment of rivers and lakes in Flanders (Belgium), *Limnologica* **40** (2010), 199–207. doi:10.1016/j.limno.2009.10.001.

[26] P.L.M. Goethals, A.P. Dedecker, W. Gabriels, S. Lek and N. De Pauw, Applications of artificial neural networks predicting macroinvertebrates in freshwaters, *Aquatic Ecology* **41** (2007), 491–508. doi:10.1007/s10452-007-9093-3.

[27] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd edn, Elsevier, San Francisco, 2006.

[28] T.H. Hoang, K. Lock, A. Mouton and P.L.M. Goethals, Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam, *Ecological Informatics* **5** (2010), 140–146. doi:10.1016/j.ecoinf.2009.12.001.

[29] A.J. Jakeman, R.A. Letcher and J.P. Norton, Ten iterative steps in development and evaluation of environmental models, *Environmental Modelling & Software* **21** (2006), 602–614. doi:10.1016/j.envsoft.2006.01.004.

[30] S.E. Jorgensen and G. Bendoricchio, *Fundamentals of Ecological Modelling*, 3rd edn, Elsevier, Amsterdam, 2001.

[31] D. Kocev, S. Dzeroski, M.D. White, G.R. Newell and P. Griffioen, Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition, *Ecological Modelling* **220** (2009), 1159–1168. doi:10.1016/j.ecolmodel.2009.01.037.

[32] A. Kozak and R. Kozak, Does cross validation provide additional information in the evaluation of regression models?, *Canadian Journal of Forest Research – Revue Canadienne De Recherche Forestiere* **33** (2003), 976–987. doi:10.1139/x03-022.

[33] D. Landuyt, S. Broekx, G. Engelen, I. Uljee, M. Van der Meulen and P.L.M. Goethals, The importance of uncertainties in scenario analyses – A study on future ecosystem service delivery in Flanders, *Science of the Total Environment* **553** (2016), 504–518. doi:10.1016/j.scitotenv.2016.02.098.

[34] G.R. Larocque, J.S. Bhatti, J.C. Ascough, J. Liu, N. Luckai, D. Mailly, L. Archambault and A.M. Gordon, An analytical framework to assist decision makers in the use of forest ecosystem model predictions, *Environmental Modelling & Software* **26** (2011), 280–288. doi:10.1016/j.envsoft.2010.03.009.

[35] B.G. Lees and K. Ritman, Decision-tree and rule-induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments, *Environmental Management* **15** (1991), 823–831. doi:10.1007/BF02394820.

[36] J.M. McPherson, W. Jetz and D.J. Rogers, The effects of species' range sizes on the accuracy of distribution models: Ecological phenomenon or statistical artefact?, *Journal of Applied Ecology* **41** (2004), 811–823. doi:10.1111/j.0021-8901.2004.00943.x.

[37] A.M. Mouton, B. De Baets and P.L.M. Goethals, Knowledge-based versus data-driven fuzzy habitat suitability models for river management, *Environmental Modelling & Software* **24** (2009), 982–993. doi:10.1016/j.envsoft.2009.02.005.

[38] A.M. Mouton, B. De Baets and P.L.M. Goethals, Ecological relevance of performance criteria for species distribution models, *Ecological Modelling* **221** (2010), 1995–2002. doi:10.1016/j.ecolmodel.2010.04.017.

[39] K.M. Osei-Bryson, Post-pruning in regression tree induction: An integrated approach, *Expert Systems with Applications* **34** (2008), 1481–1490. doi:10.1016/j.eswa.2007.01.017.

[40] R. Pesch, G. Schmidt, W. Schroeder and I. Weustermann, Application of CART in ecological landscape mapping: Two case studies, *Ecological Indicators* **11** (2011), 115–122. doi:10.1016/j.ecolind.2009.07.003.

[41] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015.

[42] A.E. Rizzoli and W.J. Young, Delivering environmental decision support systems: Software tools and techniques, *Environmental Modelling & Software* **12** (1997), 237–249. doi:10.1016/S1364-8152(97)00016-9.

[43] A. Schneiders, I. Simoens and C. Belpaire, *Waterkwaliteitscriteria Opstellen voor Vissen in Vlaanderen*, INBO, Brussel, 2009 (in Dutch).

[44] K. Soetaert and P.M.J. Herman, *A Practical Guide to Ecological Modelling. Using R as a Simulation Platform*, Springer-Verlag, New York, 2009.

[45] T. Tirelli and D. Pessani, Use of decision tree and artificial neural network approaches to model presence/absence of *Telestes muticellus* in Piedmont (North-western Italy), *River Research and Applications* **25** (2009), 1001–1012. doi:10.1002/rra.1199.

[46] T. Tirelli, L. Pozzi and D. Pessani, Use of different approaches to model presence/absence of *Salmo marmoratus* in Piedmont (Northwestern Italy), *Ecological Informatics* **4** (2009), 234–242. doi:10.1016/j.ecoinf.2009.07.003.

[47] A. Voinov and F. Bousquet, Modelling with stakeholders, *Environmental Modelling & Software* **25** (2010), 1268–1281. doi:10.1016/j.envsoft.2010.03.007.

[48] K. Wilson, A. Newton, C. Echeverria, C. Weston and M. Burgman, A vulnerability analysis of the temperate forests of South central Chile, *Biological Conservation* **122** (2005), 9–21. doi:10.1016/j.biocon.2004.06.015.

[49] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, 2005.

[50] S.C. Zhang, C.Q. Zhang and Q. Yang, Data preparation for data mining, *Applied Artificial Intelligence* **17** (2003), 375–381. doi:10.1080/713827180.

[51] A.F. Zuur, E.N. Ieno, N.J. Walker, A.A. Saveliev and G.M. Smith, *Mixed Effects Models and Extensions in Ecology with R*, Springer Science+Business, Media, New York, 2009, LLC, 2009.