

# Using classification trees to analyze the impact of exotic species on the ecological assessment of polder lakes in Flanders, Belgium

Gert Everaert<sup>a,\*</sup>, Pieter Boets<sup>a</sup>, Koen Lock<sup>a</sup>, Sašo Džeroski<sup>b</sup>, Peter L.M. Goethals<sup>a</sup>

<sup>a</sup> Ghent University, Laboratory of Environmental Toxicology and Aquatic Ecology, J. Plateaustraat 22, B-9000 Ghent, Belgium

<sup>b</sup> Jozef Stefan Institute, Department Knowledge Technologies, Jamova cesta 39, SI-1000 Ljubljana, Slovenia

## ARTICLE INFO

### Article history:

Available online 16 September 2010

### Keywords:

Brackish water  
Classification tree  
Ecological assessment  
*Gammarus tigrinus*  
Exotic species  
*Potamopyrgus antipodarum*

## ABSTRACT

Polder lakes in Flanders are stagnant waters that were flooded by the sea in the past. Several of these systems are colonized by exotic species, but have hardly been studied until present. The aim of the present study was: (1) to assess the influence of exotic macrobenthic species on the outcome of the Multimetric Macroinvertebrate Index Flanders (MMIF) and (2) to use classification trees for evaluating to what extent physical–chemical characteristics affect the presence of exotic species.

In total, 27 mollusc and 10 macro-crustacean species were present in the monitored lakes of which respectively five and four were exotic. The exclusion of the exotic species from the MMIF resulted in a significant decline of this ecological index ( $-0.03 \pm 0.04$ ;  $p = 0.00$ ). This elimination often resulted into a lower ecological water quality class and more samples were classified into the bad and poor ecological water quality classes.

Single-target classification trees for *Gammarus tigrinus* and *Potamopyrgus antipodarum* were constructed, relating environmental parameters and ecological status (MMIF) to the occurrence of both exotic invasive species. The major advantages of using single-target classification trees are the transparency of the rule sets and the possibility to use relatively small datasets. However, this classification technique only predicts a single-target attribute and the trees of the different species are often hard to integrate and use for water managers. As a solution, a multi-target approach was used in the present study. Exotic molluscs and crustaceans communities were modelled based on environmental parameters and the ecological status (MMIF) using multi-target classification trees. Multi-target classification trees can be used in management planning and investment decisions as they can lead to integrated decisions for the whole set of exotic species and avoid the construction of many models for each individual species. These trees provide general insights concerning the occurrence patterns of individual crustaceans and molluscs in an integrated way.

© 2010 Elsevier B.V. All rights reserved.

## 1. Introduction

The European Water Framework Directive (WFD) forces EU member states to determine the ecological status of water bodies (EU, 2000). The goal of this directive is to ensure that the quality of surface water and groundwater in Europe reaches a good ecological status by the year 2015. The Flemish Environment Agency (VMM) used the Belgian Biotic Index (BBI) for more than two decades to monitor the ecological water quality of Flemish rivers (De Pauw and Vanhooren, 1983). Despite the reliability and robustness of this index, a number of technical shortcomings arose about the potential application of the BBI for WFD implementation, in particular the index was not useful for stagnant waters (Gabriels et al., 2010). Therefore, a new type-specific multimetric index, combining the

robustness of the BBI with the versatility of multimetric indices, was developed. This new index, calculated from the taxonomic composition and abundance of the macroinvertebrates, is called the Multimetric Macroinvertebrate Index Flanders (MMIF) and is currently used to assess the ecological water quality in Flanders (Gabriels et al., 2010). The metrics comprised in the MMIF are taxa richness, number of Ephemeroptera, Plecoptera and Trichoptera (EPT), number of other (i.e. non-EPT) sensitive taxa, the Shannon-Wiener diversity index and the mean tolerance score. For each type of river and lake, a set of reference values for all five metrics was determined (Gabriels et al., 2010). Based on the references, a scoring system was developed for each metric consisting of threshold values needed for assigning a score ranging from zero to four (four being nearest to the reference conditions). To obtain the final index, ranging from zero for a very poor ecological quality to one for a very good ecological quality, the five metric scores are summed and subsequently divided by 20. The range of the MMIF index can be considered as an ecological quality ratio (EQR) because the max-

\* Corresponding author. Tel.: +32 092643776; fax: +32 092644199.  
E-mail address: [gert.everaert@ugent.be](mailto:gert.everaert@ugent.be) (G. Everaert).

imum MMIF value of 1 can only be obtained when all metric values are near the type-specific reference value for that metric (Gabriels et al., 2010). In order to meet the target of the WFD in 2015, aquatic systems should have a MMIF-score of 0.6 or 0.7, depending on the water type (Gabriels et al., 2010).

Macroinvertebrates are identified up to genus or family level for the calculation of the MMIF. Because of this coarse taxonomic identification level, shifts in species composition between native and exotic species often remain hidden (Gabriels et al., 2005). The introduction of exotic species might decrease the alpha diversity, which can be masked due the identification level (Gabriels et al., 2005). For example, the invader *Dikerogammarus villosus* might out-compete native gammarids (Bij de Vaate et al., 2002; Boets et al., 2010), but this will not influence the results of the index calculation at family level of a given sample, since Gammaridae are still present and tolerance classes are defined at family or genus level. Additionally, the inclusion of the exotic invasive species, such as *Corbicula*, can lead to an increase of the ecological water quality index depending on the tolerance class assigned to the invader. Previous examples suggest that the use of a standard list of taxa, where tolerance classes are assigned at specific taxonomic levels (e.g. genus or family level), can result in altered assessment scores if exotic species are present (Gabriels et al., 2005). Therefore, it is necessary to examine the influence of exotic species on the ecological assessment of aquatic ecosystems.

Freshwater exotic invasive species are an issue of growing management concern (Vander Zanden and Olden, 2008). Invasive species have one of the most harmful and least reversible impacts on natural ecosystems as they may change the local fauna and flora all around the world (Vitousek et al., 1996; Ricciardi and MacIsaac, 2000). Exotic invasive species may decrease the ecological quality through changes in biological, chemical and physical properties of aquatic ecosystems (Olenin et al., 2007). These changes include: elimination of sensitive or rare species; alteration of native communities; algal blooms; modification of substrate conditions and the shore zones; alterations of oxygen and nutrient content, pH and transparency of the water; accumulation of synthetic pollutants, etc. (Olenin et al., 2007). For instance, Boets et al. (2009) indicated that the exotic macro-crustacean *Procambarus clarkii* predate on native benthic macroinvertebrates, spreads diseases and affects the physical habitat via burrowing activities. Our research focussed on exotic molluscs and macro-crustaceans, because these have probably the highest impact among all aquatic freshwater invaders in Europe (Orendt et al., 2010).

Stimulated by the expansion of the global transport of goods and people, the numbers and costs of exotic species are rising at an alarming rate (Lovell and Stone, 2006). Exotic species may be unintentionally imported by ships discharging their ballast water (Mills et al., 1993; Lovell and Stone, 2006; Colautti et al., 2006). Leung et al. (2006) found that recreational boaters between lakes are an important pathway of overland dispersal of exotic species. Pathogens and parasites have been introduced unintentionally into the USA via infected stock for aquaculture farms (Naylor et al., 2001). Policy makers spend a lot of money trying to control or remove invaders from our environment (Pimentel et al., 2000; Pimentel et al., 2005). Many USA states have recently created exotic species advisory councils that bring together regulators, researchers and other stakeholders to address research, policy and management needs (Lodge et al., 2006). However, managers lack predictive tools to help them prioritise invasion threats and to help them decide where they should allocate the limited resources for prevention and mitigation most effectively (Ricciardi, 2003).

One of the methods applied by managers in the USA is the national Gap Analysis Program (GAP). This method identifies 'gaps' in the network of conservation land and water areas (Scott et al., 1993). The framework documents biogeographic information and

organizational cooperation in ways meaningful to their management and can therefore be useful in the context of exotic species (Jennings, 2000).

Other methods, such as data mining techniques, can be helpful because they allow accurate predictions of species preferences and impacts. Classification trees can give insight in complex, unbalanced, non-linear ecological data where commonly used exploratory and statistical modelling techniques often fail to find meaningful ecological patterns (De'ath and Fabricius, 2000). Classification trees have been applied in numerous ecological studies (Dakou et al., 2007; Boets et al., 2010) and have proven to have a high potential in macroinvertebrate habitat suitability analysis as they combine reliable classifications with a transparent set of rules (Hoang et al., 2009).

Classification trees are decision trees that predict the value of a discrete-valued (nominal) target variable (Breiman et al., 1984). Decision trees are hierarchical structures, where the internal nodes contain tests on the input attributes. Each branch of an internal test corresponds to an outcome of the test and the prediction for the value of the target attribute is stored in a leaf. Each leaf of a decision tree contains a prediction for the target variable. A single-target approach learns a model for each target attribute separately, whereas a multi-target approach builds one model for all target attributes simultaneously (Koccev et al., 2009). Therefore, a single-target approach can be used to predict the possible occurrence of individual exotic species based on physical–chemical parameters, while possible changes in species composition can be highlighted using a multi-target approach.

Polder lakes, situated in the north of Flanders, are brackish, stagnant waters situated on the inland side of the dikes (Delaunoy, 1982). They find their origin in history when, due to flooding by the sea, dikes gave way and land was washed out. The salinity of these lakes, determining the fauna and flora, depends on their age and the possible influence of seepage water (van Puijenbroek et al., 2004). The salinity of polder lakes in Flanders decreases from north to south and from west to east (Delaunoy, 1982). Many of these shallow lakes are hypertrophic and dominated by algal blooms. The eutrophication process became problematic in the 1950s due to run-off from agriculture and discharges from industry and untreated household waste (van Puijenbroek et al., 2004). In the mid-1980s, projects on lake restoration were started in the Netherlands (Van Huet, 1992). In Flanders, apart from a study by Dumont and Gysels (1971), little research has been carried out on these aquatic systems.

The aims of the present paper were (1) to assess the influence of exotic species on the MMIF and (2) to construct single- and multi-target classification trees to predict the presence of exotic species based on physical–chemical parameters and occurrence of other species.

## 2. Materials and methods

### 2.1. Study area and data collection

The dataset contained 108 samples comprising biological and physical–chemical information of 45 polder lakes (Fig. 1). The polder lakes, all located in the Northern part of Flanders, can be divided in three clusters. The first, most westerly oriented cluster of lakes, is situated close to the city of Ostend, whereas the second, most easterly oriented cluster, is located close to the river Scheldt nearby the city of Antwerp. The remaining polder lakes, situated between the first two clusters, are distributed along the Dutch border. Most polder lakes are exploited for recreational purposes: the smaller polder lakes are frequently used for fishing, whereas the bigger lakes are suitable for sailing and windsurfing.

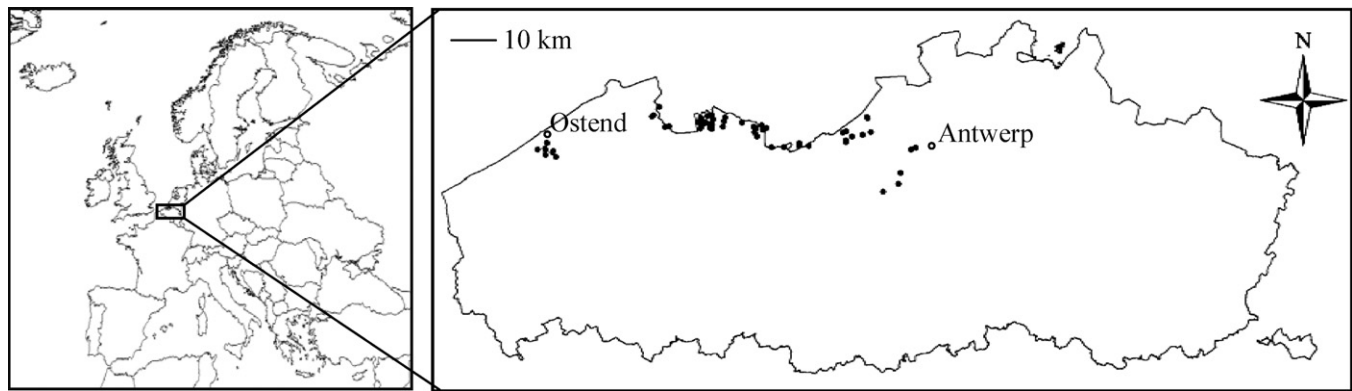


Fig. 1. Location of the monitored polder lakes (black dots) in Flanders.

Physical–chemical and biological data of 28 polder lakes from the period 1992–2006 were provided by the Flemish Environment Agency (VMM). In 2009, 17 additional polder lakes, which were not monitored by the VMM, were sampled to enlarge the dataset. For each location, physical–chemical variables were recorded (Table 1) and a biological sample was taken. Conductivity ( $\mu\text{S}/\text{cm}$ ), pH and water temperature ( $T, ^\circ\text{C}$ ) of the polder lakes were measured in the field using a pH (Metrohm 826 pH mobile) and conductivity meter (WTW Cond. 315i). The dissolved oxygen (DO,  $\text{mg}/\text{L}$ ) concentration was measured in the field using an oxygen electrode (WTW Oxi 330 oximeter). Chloride ( $\text{Cl}^-$ ,  $\text{mg}/\text{L}$ ), nitrate ( $\text{NO}_3^-$ -N,  $\text{mg N}/\text{L}$ ) and orthophosphate ( $\text{oPO}_4^{3-}$ -P,  $\text{mg P}/\text{L}$ ) concentrations were quantified in the laboratory by means of a spectrophotometer using the standard kit Merck spectroquant. In analogy with Costil et al. (2001) and because the VMM provided limited data about the salinity of the polder lakes, the conductivity was used as an indicator for salinity.

All biological samples were taken according to the procedure described by Gabriels et al. (2010). Macroinvertebrates were sampled using a standard handnet. This handnet consisted of a metal frame of approximately 0.2 m by 0.3 m to which a conical net is attached with a mesh size of minimum 300 and maximum 500  $\mu\text{m}$ . The frame was attached to a 2 m long shaft with two handles enabling it to be handled in a similar way as a scythe. With the handnet, all accessible aquatic habitats within a stretch of 10–20 m were sampled. This included the bed substrate (stones, sand or mud), macrophytes (floating, submerged, emerged), immersed roots of overhanging trees and all other natural or artificial substrates, floating or submerged in the water. Each aquatic habitat was explored in order to collect the highest possible diversity of macroinvertebrates. For this purpose, kick sampling was performed by vertically positioning the handnet on the bed and turning over bottom material located immediately upstream by foot or hand. Sampling effort was proportionally distributed over all accessible aquatic habitats during 5 min. Subsequently, the identification of the organisms is carried out up to the taxonomic level as indicated by Gabriels et al. (2010). In the interest of the research, the collected molluscs and macro-crustaceans were identified up

to species level. In this way, it was possible to distinguish exotic species from native ones and to evaluate their impact on the MMIF.

The type-specific MMIF was calculated twice for each sample. During the index calculation, polder lakes were regarded as 'very slightly brackish lakes' (code Bzl). First, the MMIF was calculated for the whole (native + exotic species) biological sample. Second, the MMIF was recalculated exclusively based on the native species found in the sample. Both calculations were included in the final dataset.

Seasonality may not be neglected when monitoring aquatic ecosystems (Gabriels et al., 2010). Using a constraining time frame for sampling may result in missing information on the overall community at a site (Linke et al., 1999), but it can be assumed that a large timeframe is sufficient for water quality assessment purposes (Gabriels et al., 2010). Constraining the sampling period to spring, summer and autumn is recommended to avoid extreme hydrological regimes and temperatures in winter and to minimize the variability of the species detection efficiency among different sampling campaigns. Therefore, all the biological samples were taken in spring.

Biocontamination of sampling sites was assessed using the integrated biocontamination index (IBCI) derived from two metrics: abundance contamination index (ACI) and richness contamination index (RCI) at ordinal rank (Arbačiauskas et al., 2008). The IBCI could be used because multiple calculation of ACI and RCI were available for the same ecosystem (i.e. samples were collected in polder lakes). The IBCI was derived by averaging ACI and RCI per sampling year and ranking IBCI based on the thresholds for the five classes of biocontamination. These classes range from 0 ('no' contamination) to 4 ('severe' contamination). The threshold for the lowest quality limit ('bad' class) is based on the assumption that when exotic species represent more than half the detected orders or when their abundance exceeds 50% of the individuals, the community/assembly has developed as a consequence of the occurrence of exotic species (Arbačiauskas et al., 2008).

The final dataset comprised per sample: (1) the sampling location and year; (2) the presence/absence of different mollusc and

Table 1  
Observed physical–chemical characteristics in the Flemish polder lakes, based on 108 samples.

Variable	Abbreviation	Unit	Minimum	Maximum	Mean	Standard deviation
Chloride	$\text{Cl}^-$	$\text{mg}/\text{L}$	4.5	1330.0	359.7	329.5
Conductivity	–	$\mu\text{S}/\text{cm}$	547	8700	2110	1483
Dissolved oxygen	DO	$\text{mg}/\text{L}$	1.44	25.30	8.33	4.63
Nitrate	$\text{NO}_3^-$ -N	$\text{mg N}/\text{L}$	0.08	9.82	1.64	2.03
Orthophosphate	$\text{oPO}_4^{3-}$ -P	$\text{mg P}/\text{L}$	0.04	4.00	0.70	0.80
pH	–	–	6.36	9.32	8.10	0.54
Water temperature	T	$^\circ\text{C}$	6.4	26.9	16.4	4.1

**Table 2**

Correlation coefficients between physical–chemical properties observed in the polder lakes.

Variable	Cl <sup>−</sup>	Conductivity	DO	NO <sub>3</sub> <sup>−</sup> -N	oPO <sub>4</sub> <sup>3−</sup> -P	pH	T
Cl <sup>−</sup>	–						
Conductivity	0.77 <sup>a</sup>	–					
DO	0.08	0.15	–				
NO <sub>3</sub> <sup>−</sup> -N	−0.13	−0.14	0.12	–			
oPO <sub>4</sub> <sup>3−</sup> -P	0.24	0.26	−0.42 <sup>a</sup>	−0.22	–		
pH	0.22	0.29 <sup>a</sup>	0.66 <sup>a</sup>	−0.16	−0.25	–	
T	−0.33 <sup>a</sup>	−0.19	0.13	−0.33 <sup>a</sup>	−0.39 <sup>a</sup>	0.11	–

<sup>a</sup> Correlation is significant at the level 0.05.

macro-crustacean species; (3) the quantified physical–chemical variables; (4) the IBCI and (5) the MMIF of the sample.

## 2.2. Statistical data processing

The data were first processed using the Statistical Package for the Social Sciences 16.0 (SPSS, 2008). The statistical analyses with SPSS were performed as follows:

- First, the relations between physical–chemical characteristics were explored using the non-parametric Spearman correlation coefficient (Table 2).
- Second, the possible difference between the two ways of calculating the MMIF was evaluated using a non-parametric Kruskal–Wallis test.
- Third, species preferences concerning the physical–chemical conditions of the polder lakes were visualised by Box-and-Whisker plots. These plots were made in SPSS using default settings and comprise upper, median and lower quartiles and upper and lower fences (excluding outliers which were unusually distant from the median).

## 2.3. Modelling field data

### 2.3.1. Construction of single-target classification trees

Classification trees are data driven methods that are particularly useful to develop ecological models based on small datasets (Goethals et al., 2007). The outcome is noteworthy for users, as often relatively reliable models are generated in a very short calculation time and the models are transparent and easy to interpret (Hoang et al., 2009). Classification trees were built through applying the Waikato Environment for Knowledge Analysis (WEKA; Witten and Frank, 2005 version 3.6.1). Rules relating the presence/absence of *Hippeutis complanatus*, *Gammarus tigrinus* and *Potamopyrgus antipodarum* with physical–chemical conditions, sample characteristics and the occurrence of other species were created by means of single-target classification trees using the J48 algorithm (a Java implementation of the C4.5 algorithm) (Witten and Frank, 2005).

### 2.3.2. Construction of multi-target classification trees

The simultaneous occurrence of exotic molluscs and macro-crustaceans was predicted using multi-target classification trees. We used the CLUS system for constructing multi-target decision trees (Blockeel and Struyf, 2002). The resulting trees are an instantiation of the predictive clustering trees (PCTs) framework (Blockeel et al., 1998). In this framework, a tree is viewed as a hierarchy of clusters: a node corresponds to a cluster. PCTs have been used to handle different types of targets: multiple target variables, both discrete and continuous (Struyf and Džeroski, 2006), time series (Džeroski et al., 2007) and hierarchies of classes, with multiple class-labels per example (Vens et al., 2008).

Multi-target classification trees generalize classification trees for the prediction of several discrete-valued target attributes simultaneously (Blockeel et al., 1998; Struyf and Džeroski, 2006). The

leaves of a multi-target classification tree store a vector of class values, instead of storing a single class value like single-target classification trees do. This means that each component of the vector is a prediction for one of the target attributes.

Multi-target classification trees were constructed with a recursive partitioning algorithm from a training set of records. This algorithm is known as TDIDT (top-down induction of decision trees) (Quinlan, 1986). The records include measured values of the descriptive and the target attributes. The tests in the internal nodes of the tree refer to the descriptive, while the predicted values in the leaves refer to the target attributes.

The TDIDT algorithm starts by selecting a test for the root node. Based on this test, the training set is partitioned into subsets according to the test outcome. In the case of binary trees, the training set is split into two subsets: one containing the records for which the test succeeds (typically the left subtree) and the other contains the records for which the test fails (typically the right subtree). This procedure is recursively repeated to construct the subtrees. The partitioning process stops when a stopping criterion is satisfied, then the prediction vector is calculated and stored in a leaf. The *F*-test stopping criterion has been used; a node was split if a statistical *F*-test indicated a significant (at level 0.1) reduction of variance inside the subsets.

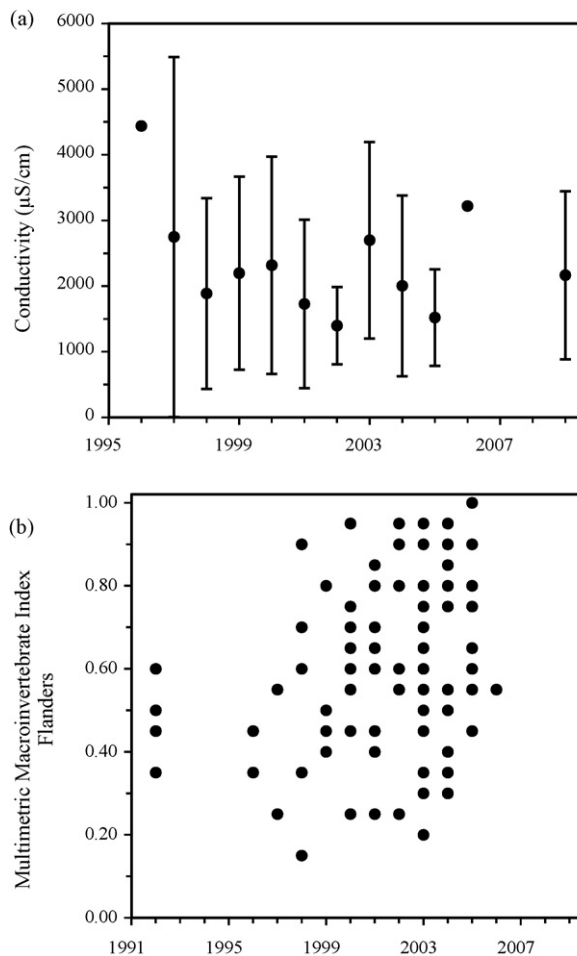
One of the most important steps in the tree induction algorithm is the test selection procedure. For each node a test is selected using a heuristic function computed on the training data. The goal of the heuristic is to guide the algorithm towards smaller trees with good predictive performance. The performance of the produced trees improved based on a heuristic function called SSreduction, which reduced the variance between the observations and the corresponding predictions.

## 2.4. Evaluation of classification trees

Classification trees can be built using a relatively small dataset. In such cases, when all available data should be used for training and validating the model, cross-validation is useful (Goethals et al., 2007). This technique estimates the generalization error of a given model and uses all data to construct and test the model. The stability of both types of classification trees was maximized using a 10-fold cross-validation (Witten and Frank, 2005). In 10-fold cross-validation, the original data are randomly partitioned into 10 subsamples of approximately equal size. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining nine subsamples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data, and the results from the 10-folds are averaged to produce a single estimation.

Several single- and multi-target classification trees were built using multiple combinations of physical–chemical variables (Cl<sup>−</sup>, DO, NO<sub>3</sub><sup>−</sup>-N, oPO<sub>4</sub><sup>3−</sup>-P, pH, T and conductivity), sample characteristics and occurrence of other species. Combinations resulting in sufficiently reliable models were selected based on the Cohen's





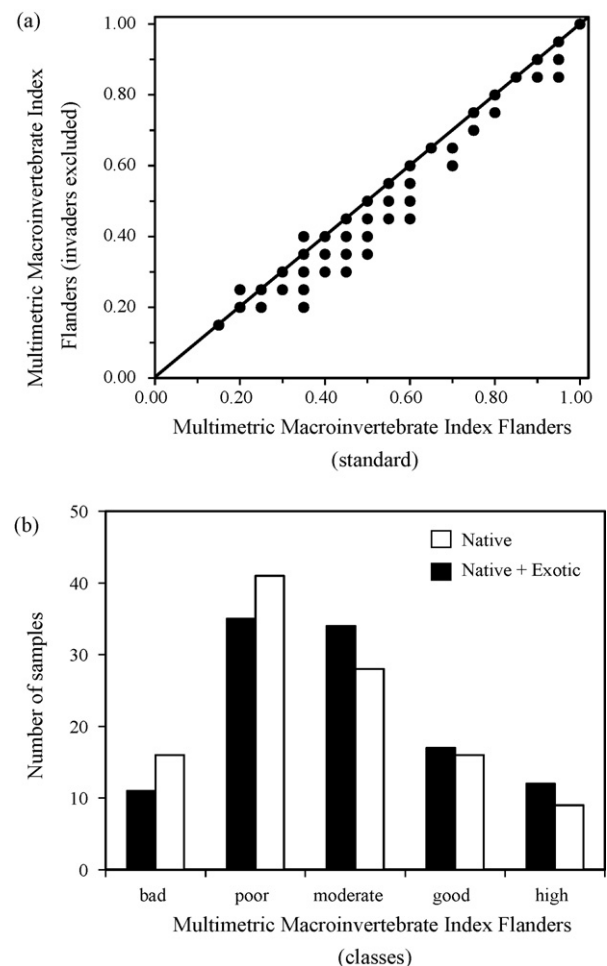
**Fig. 2.** Evolution of the yearly average conductivity (a) and the ecological status quantified by the Multimetric Macroinvertebrate Index Flanders (MMIF) (b) of the polder lakes between 1992 and 2009.

kappa statistic ( $\kappa$ ) (Cohen, 1960) and the percentage of correctly classified instances (CCI). CCI is calculated as the percentage of the true positive and true negative predictions. However, CCI is affected by the prevalence of the taxon being modelled (Fielding and Bell, 1997; Manel et al., 2001). Various authors prefer the use of  $\kappa$  because it is more reliable than CCI.  $\kappa$  measures the proportion of all possible cases of the presence or absence that are predicted correctly by a model after accounting for chance predictions (Hoang et al., 2009). Models with CCI higher than 70% and  $\kappa$  higher than 0.4 were considered reliable (D'heygere et al., 2006; Dakou et al., 2007; Gabriels et al., 2007; Goethals et al., 2007). However, Sim and Wright (2005) suggest that the use of a confidence interval around the sample estimate of  $\kappa$  is better than focussing on the 0.4 threshold. Earlier, Landis and Koch (1977) attempted to indicate the degree of agreement that exists when the Cohen's kappa is found to be in various ranges:  $\leq 0$  (poor); 0–0.2 (slight); 0.2–0.4 (fair); 0.4–0.6 (moderate); 0.6–0.8 (substantial) and 0.8–1 (almost perfect).

### 3. Results

#### 3.1. Environmental variables

The conductivity of the polder lakes, used as an indicator for their salinity, fluctuated between 1996 and 2009. The evolution of the yearly mean conductivity of the polder lakes is illustrated in Fig. 2a. The conductivity was positively cor-

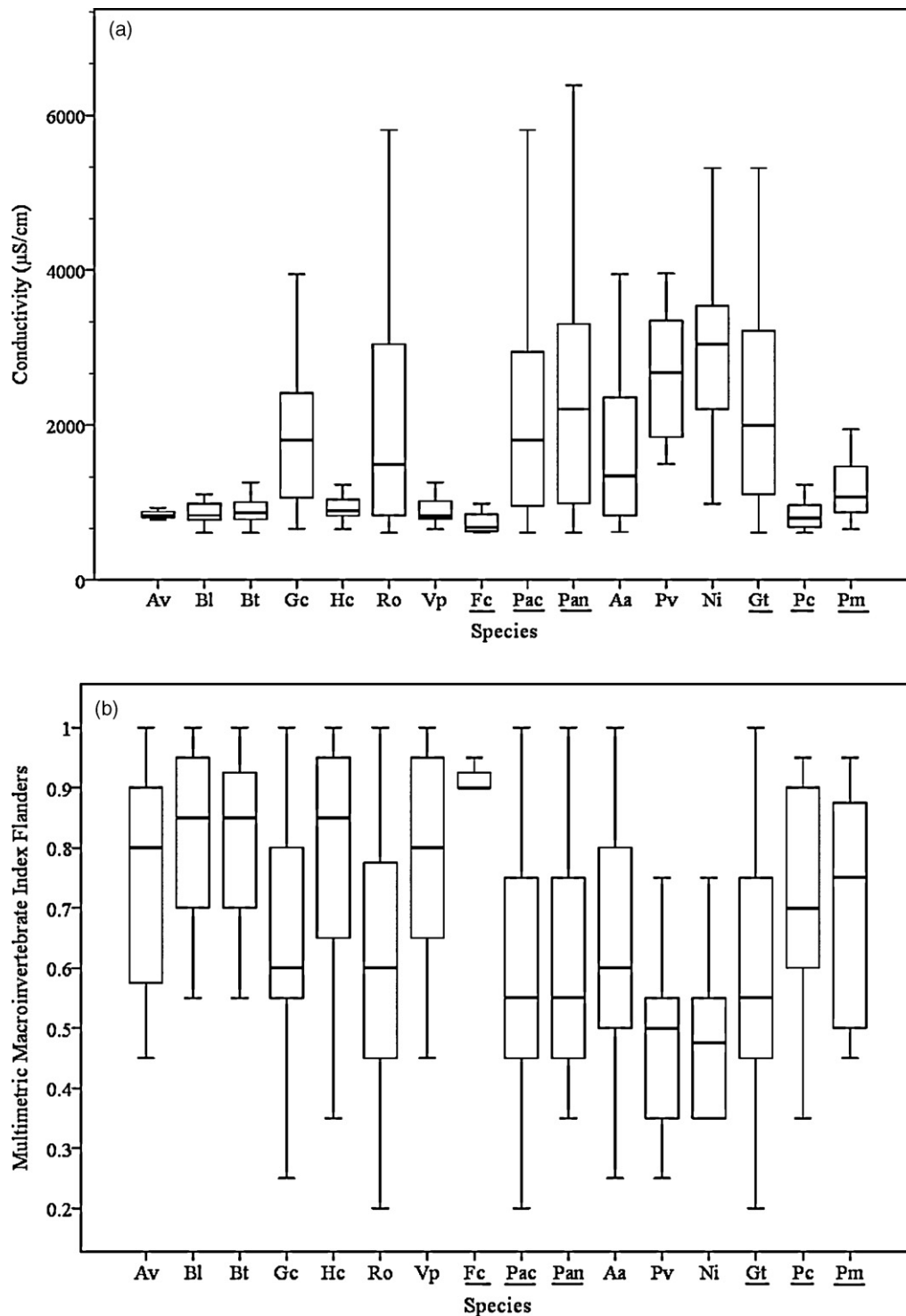


**Fig. 3.** Multimetric Macroinvertebrate Index Flanders (MMIF) (as described by Gabriels et al. (2010)) plotted versus the MMIF calculated after exclusion of exotic species (a) and Multimetric Macroinvertebrate Index Flanders (MMIF) ecological water quality classes including exotic species (black) and excluding exotic species (white) (b).

related with pH ( $r=0.29$ ;  $p=0.004$ ) and chloride concentration ( $r=0.77$ ;  $p=0.000$ ), whereas the oxygen concentration was positively correlated with pH ( $r=0.66$ ;  $p=0.000$ ) and negatively correlated with the orthophosphate concentration ( $r=-0.42$ ;  $p=0.002$ ). The water temperature was negatively correlated with the chloride concentration ( $r=-0.33$ ;  $p=0.011$ ), the nitrate concentration ( $r=-0.33$ ;  $p=0.008$ ) and the orthophosphate concentration ( $r=-0.39$ ;  $p=0.004$ ) (Table 2).

#### 3.2. Identified macroinvertebrates

In total, 27 mollusc species were found, of which five exotic species: *P. antipodarum*, *Physella (Costatella) acuta*, *Ferrissia (Pettancylus) clessiniana*, *Lithoglyphus naticoides* and *Cerastoderma glaucum*. Ten macro-crustaceans were found four of which are exotic species: *G. tigrinus*, *Proasellus coxalis*, *Proasellus meridianus* and *Crangonyx pseudogracilis*. The most frequently found invaders were *P. acuta* (found in 66% of the samples), *G. tigrinus* (59%) and *P. antipodarum* (48%). The most frequently found native mollusc species were *Radix ovata* (55%), *Gyraulus (Armiger) crista* (33%), *Bithynia tentaculata* (31%), *Bithynia leachii* (27%) and *Valvata piscinalis* (27%). The most regularly encountered native macro-crustaceans were *Asellus aquaticus* (50%), *Palaemonetes varians* (24%) and *Neomysis integer* (19%).



**Fig. 4.** Box-and-Whisker plots for conductivity (a) and ecological water quality (b). Exotic species are underlined. Av = *Anisus vortex*, Bl = *Bithynia leachii*, Gc = *Gyraulus crista*, Hc = *Hippeutis complanatus*, Ro = *Radix ovata*, Vp = *Valvata piscinalis*, Fc = *Ferrissia clessiniana*, Pac = *Physella acuta*, Pan = *Potamopyrgus antipodarum*, Aa = *Asellus aquaticus*, Pv = *Palaemonetes varians*, Ni = *Neomysis integer*, Gt = *Gammarus tigrinus*, Pc = *Proasellus coxalis*, Pm = *Proasellus meridianus*.

### 3.3. Influence of exotic species on the ecological status of the polder lakes

In Flanders, the ecological water quality is evaluated based on the MMIF. From the years 1992 to 2006, the mean ecological quality of the polder lakes increased significantly from 0.42 to 0.72 ( $r = 0.31$ ;  $p = 0.00$ ) (Fig. 2b). Between 1992 and 2009, the IBCI of the polder lakes fluctuated between high and severe biocontamination, but no trend could be derived from the data.

Although the exclusion of invaders resulted in a significant drop in MMIF ( $p = 0.02$ ), the index increased for two samples when exotic species were excluded from the calculation (Fig. 3a). The difference between the two ways of calculation varied from  $-0.15$  to  $+0.05$ , with the mean difference being  $-0.03 \pm 0.04$ . The number of samples classified in the moderate, good and high biological water quality categories decreased when exotic species were excluded from the calculation, while more samples were categorized in the bad and poor classes (Fig. 3b).

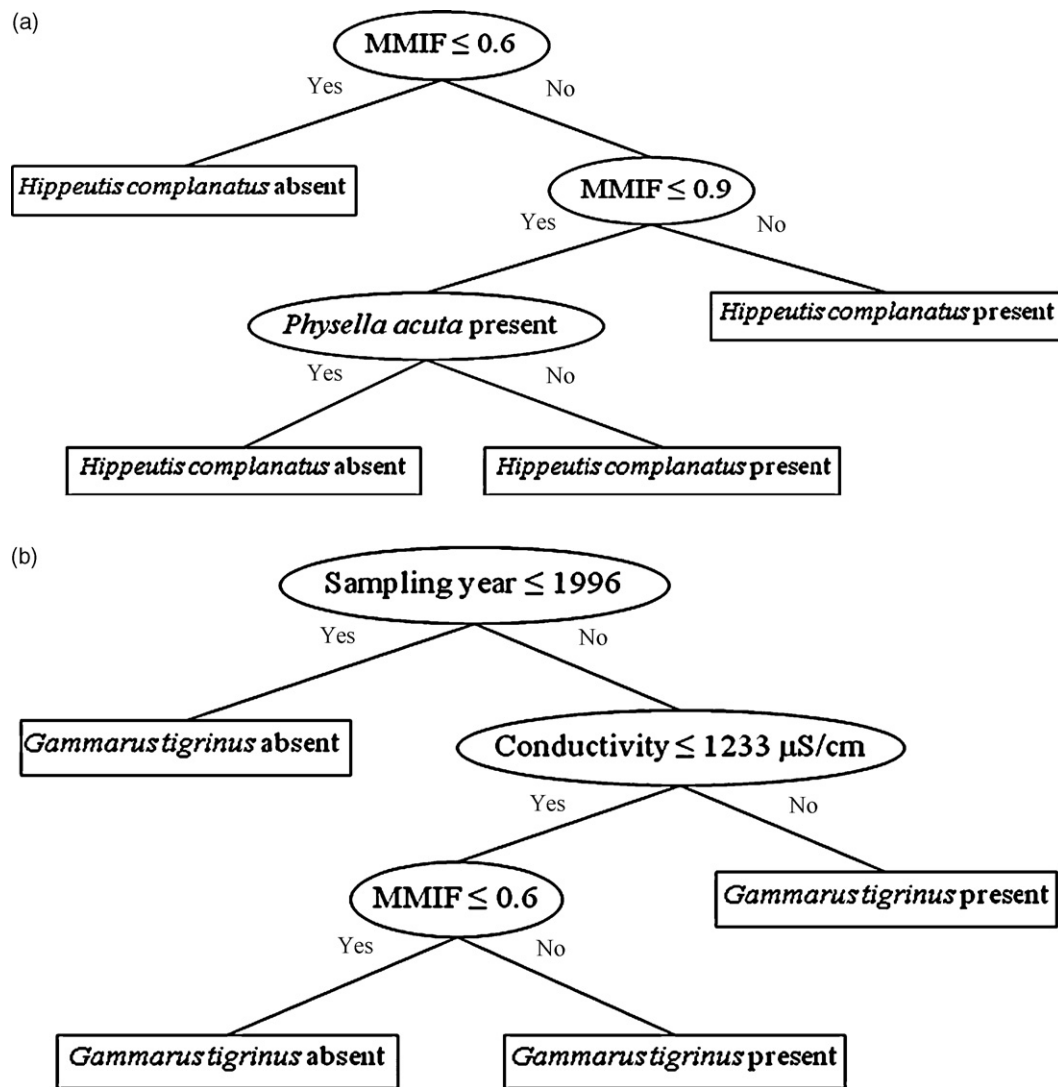


Fig. 5. Single-target classification trees predicting the occurrence of *Hippeutis complanatus* (a) and *Gammarus tigrinus* (b).

### 3.4. Single-species prediction

Classification trees were built using several combinations of input variables. According to this method, conductivity was the only environmental variable containing valuable information for predicting the occurrence of exotic molluscs and macro-crustaceans. The other physical–chemical variables like  $\text{Cl}^-$ , DO,  $\text{NO}_3^-$ -N,  $\text{oPO}_4^{3-}$ -P, pH and T were not selected by the algorithms. Apart from the conductivity, inclusion of information about the MMIF of the sample, sampling year and presence of other species in the sample, resulted in reliable models.

Native species like *B. leachii* and *H. complanatus* typically occur in freshwater (Fig. 4a). Based on a single-target classification tree, the MMIF had to reach at least a value of 0.75 for the presence of *B. leachii* (CCI = 87%;  $\kappa = 0.6$ ), which is also reflected in the Box-and-Whisker plot (Fig. 4b). Similarly, *H. complanatus* was absent as long as the MMIF did not reach 0.6, while in waters with a score above 0.9 it was present in all seven cases (Fig. 5a). The occurrence of this mollusc in the range between 0.6 and 0.9 depended upon the presence of *P. acuta*: if this exotic invasive species lived in the considered polder lakes, *H. complanatus* was absent (CCI = 84%;  $\kappa = 0.5$ ).

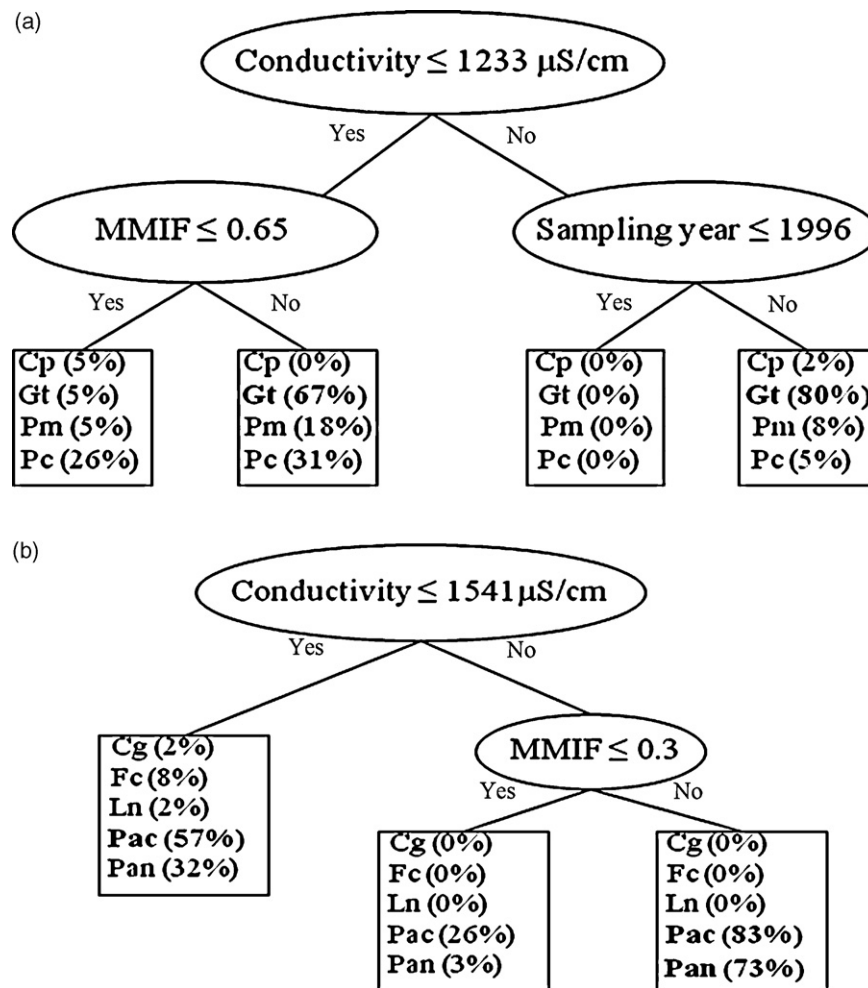
The occurrence of two dominant exotic species, *G. tigrinus* and *P. antipodarum*, was predicted. Predicting the possible occurrence of *G. tigrinus* was interesting due to its invasive behavior. Based on

the year of sampling, the conductivity and the MMIF of the sampling site, it was possible to predict the occurrence of *G. tigrinus* by means of a single-target classification tree (Fig. 5b) with a fair reliability (CCI = 71%;  $\kappa = 0.4$ ). The root of this classification tree confirms that the invader remained absent in the polder lakes until 1997. The second division was based on the conductivity of the polder lakes.

The occurrence of *P. antipodarum* was predicted with moderate reliability using the presence of *P. acuta*. If *P. acuta* was found, *P. antipodarum* was also predicted as present (CCI = 69%;  $\kappa = 0.4$ ).

### 3.5. Multi-target analysis of the exotic subgroup

From our dataset, two multi-target trees predicting the exotic macro-crustacean and mollusc communities were developed (Fig. 6a and b). The leaves of the resulting trees represent the probability of occurrence of the selected species. An exotic species was predicted as present if the probability of occurrence was greater than 50%, these species are indicated in bold. The tree predicting the exotic macro-crustacean community had the following reliabilities per species: *C. pseudogracilis* (CCI = 98%,  $\kappa = 0$ ), *G. tigrinus* (CCI = 64%,  $\kappa = 0.2$ ), *P. meridianus* (CCI = 91%,  $\kappa = 0$ ), *P. coxalis* (CCI = 86%,  $\kappa = 0$ ). The model predicting the mollusc invaders had subsequent performances per species: *C. glaucum* (CCI = 99%,  $\kappa = 0$ ), *F. clessiniana*



**Fig. 6.** Multi-target classification tree predicting the macro-crustaceans community (a) and the mollusc community (b). The probability of occurrence is indicated per species between brackets. Cp = *Crangonyx pseudogracilis*, Gt = *Gammarus tigrinus*, Pm = *Proasellus meridianus*, Pc = *Proasellus coxalis*, Cg = *Cerastoderma glaucum*, Fc = *Ferrissia clessiniana*, Ln = *Lithoglyphus naticoides*, Pac = *Physella acuta*, Pan = *Potamopyrgus antipodarum*.

(CCI = 98%,  $\kappa = 0$ ), *L. naticoides* (CCI = 96%,  $\kappa = 0$ ), *P. acuta* (CCI = 69%,  $\kappa = 0.3$ ), *P. antipodarum* (CCI = 74%,  $\kappa = 0.5$ ).

#### 4. Discussion

Between 1992 and 2006, the overall ecological water quality of the polder lakes changed from moderate ( $0.50 < \text{MMIF} < 0.69$ ) to good ( $0.70 < \text{MMIF} < 0.89$ ). During this period, a new exotic species appeared in the polder lakes. *G. tigrinus* was first found in Flemish waters in the year 1991 (Messiaen et al., 2010), but identification of the samples taken in the polder lakes revealed that *G. tigrinus* occurred in the polder lakes since 1997. Although species belonging to the genus *Gammarus* were originally evaluated as quite sensitive to pollution in our region (De Pauw and Vanhooren, 1983), Koop and Grieshaber (2000), Normant et al. (2007) and Wijnhoven et al. (2003) concluded that *G. tigrinus* was more tolerant to fluctuations in abiotic conditions than native *Gammarus* species. More generally, Devin and Beisel (2007) and Karatayev et al. (2009) found that invaders can generally live and reproduce in a wider range of environmental conditions than native species. Although it is generally accepted that exotic species can have direct and indirect negative effects on ecosystems (Olenin et al., 2007; Boets et al., 2009) and that they are more tolerant towards pollution than native species (Wijnhoven et al., 2003; Devin and Beisel, 2007 and Normant et al., 2007), in the MMIF calculation they are still evaluated in the same way as native species are. Therefore the increase of the MMIF

between 1992 and 2006 can be related to the colonization of the lakes with exotic species.

Most ecological indices have difficulties dealing with exotic species, mainly due to the level of identification and the fixed species list. Excluding exotic species from the ecological water quality assessment (MMIF) resulted in a significant decline of the ecological index, which often resulted in a lower ecological water quality class. One of the reasons for this is that the MMIF does not explicitly attribute a negative role and scoring to exotic species, but assumes that negative impacts are indirectly reflected by changes in the community metrics. Indices like the MMIF that are based on coarse identification levels assume that all species within a given taxon have a similar sensitivity towards pollution, so the tolerance classes are assigned to all species that belong to the corresponding taxa. Gabriels et al. (2005) found a similar drop of the ecological water quality, based on the BBI, after the exclusion of the exotic invasive mollusc *Dreissena polymorpha*: the index reduced in 23.8% of the cases after recalculation. As a solution, tolerance classes should be revised if new exotic species are identified (MacNeil and Briffa, 2009). Similarly, Walley and Hawkes (1996) suggested that the Biological Monitoring Working Party score (BMWP) for Gammaridae should be downgraded from 6 to 4, due to the presence of pollution-tolerant exotic invasive amphipods such as *G. tigrinus* (Karatayev et al., 2009).

It is clear that invaders should be more carefully considered in ecological water quality assessment because of their influence on



the ecological water quality evaluation. In this context, it should be noted that different countries have dissimilar approaches to ecological assessment of surface waters. In the Netherlands, where ecological assessment is based on species-level identifications, exotic species are excluded from the assessment scoring systems (Orendt et al., 2010). However, several researchers consider exotic species as an integral part of the aquatic species community (Gabriels et al., 2005; Cardoso and Free, 2008; Arndt et al., 2009; MacNeil and Briffa, 2009). In Germany, the German Saprobic Index (GSI), used to reflect the organic impact from wastewaters, is also based on species-level data. For a large majority of the sites, exotic species were so dominant that eventual differences in GSI often remained hidden (Arndt et al., 2009). As a solution, an additional metric was defined evaluating the native species composition in relation to the dominance of invaders (Arndt et al., 2009). An example of such kind of metric is the IBCI described by Arbačiauskas et al. (2008) and is used for a similar analysis by MacNeil et al. (2010). Starting from the relative abundance of exotic species within a community and the proportion of exotic species within a community at ordinal taxonomic rank the authors developed an index to measure the biocontamination of aquatic communities. Calculating the IBCI for the different samples originating from the polder lakes revealed that the index remained relatively constant between 1992 and 2006: the polder lakes are always evaluated as highly or severely biocontaminated. With the arrival of *G. tigrinus* in the year 1997, the IBCI did not change significantly. This can be explained by the fact that the dominant exotic molluscs *P. acuta* and *P. antipodarum* have been abundant in the polder lakes for decades, which resulted in a high IBCI since the start of the monitoring.

Most polder lakes are connected through small watercourses (Delaunoy, 1982) so discharges of ballast water from ships in nearby harbours can be an important vector for invaders (Lovell and Stone, 2006). Additionally, due to fish stockings, exotic species are probably unintentionally spread by fisherman (Naylor et al., 2001). Other possible vectors such as recreational activities (water sports or boat trips) should be followed up, because some invaders can disperse through surfboards and boats (Leung et al., 2006). In this context classification trees can be applied to identify polder lakes where the physical–chemical conditions are favourable for invaders. Suitable lakes that remained free of exotic species should be protected.

Classification trees relate species occurrences with environmental variables and/or occurrences of other species. Illustratively, a model was shown predicting the presence of *H. complanatus*. The occurrence of this native species depended upon the occurrence of *P. acuta*. Both species occupy the same type of water and can therefore compete (Gittenberger et al., 1998). This type of single-target classification tree can also be used to predict the occurrence of exotic species. The constructed single-target model predicting the occurrence of *G. tigrinus* confirms what is generally known about this species in Flanders. The first observation of *G. tigrinus* in Flanders was in 1991 (Messiaen et al., 2010), but currently it is spread all over Flanders (Boets et al., unpublished data). The first specimen of *G. tigrinus* in the polder lakes was found in a sample from 1997, but currently, it is a common species in Flemish polder lakes. This information is clearly reflected in the root of the model. Adriaenssens et al. (2006) found that conductivity was an important factor for the distribution of *Gammarus*. Piscart et al. (2005) observed that *G. tigrinus* was more abundant at higher salinity sites in the Meurthe River in north-eastern France, which suggest that rising salinity concentrations affected the species composition and favoured invaders. These findings are reflected in the second division of the constructed classification tree. If the conductivity, which is related to the salinity, was sufficiently high, *G. tigrinus* was predicted as present, if this was not the case, its occurrence depended upon the MMIF. Comparing the second rule of the

single-target classification tree with the Box-and-Whiskers plots confirms that *G. tigrinus* occurs in a wider range of conductivity than the native *Gammarus duebeni*. However, in five samples, at the highest conductivities, both the exotic invasive and the native gammarid were present. Based on our results, it was impossible to predict which equilibrium between both species will be reached. However, in Poland, Grabowski et al. (2006) found that *G. tigrinus* outcompeted *G. duebeni*. In polder lakes with a low to high conductivity (salinity), *G. tigrinus* was often the dominant representative of the gammarids. At even higher conductivities (salinities), *G. duebeni* was also present. At the bottom of the classification tree, the last rule was related to the ecological water quality. This final step indicated that, if conductivity was low, *G. tigrinus* preferred at least moderate water quality. In general, the combination of the classification tree and the Box-and-Whisker plots confirmed that *G. tigrinus* can live in a wide range of physical–chemical conditions.

*P. acuta* and *P. antipodarum*, with first observations in Belgium in 1869 (Adam, 1960) and in 1927 (Keppens and Keppens, 1996), respectively, are widely distributed in Flanders. The presence of the mud snail *P. antipodarum* was related to the presence of *P. acuta*. Similarly, Cope and Winterbourn (2004) found both species together in many streams, ponds and lakes in New Zealand, where *P. antipodarum* is a native and *P. acuta* an exotic invasive species. They concluded that the growth and reproductive output of both snail species were influenced more by the density of conspecifics than by the presence and density of the other species. According to Gittenberger et al. (1998), both exotic invasive molluscs are able to live in water with salinities up to 8‰ and are tolerant to pollution. Only their feeding habits differ: whereas *P. antipodarum* only needs detritus to grow and reproduce, *P. acuta* also feeds on carrion. The constructed Box-and-Whisker plots confirm that both species prefer water with a relatively high conductivity and a poor or moderate ecological status. Leppäkoski and Olenin (2000), Gérard et al. (2003) and Alonso and Castro-Diez (2008) obtained similar conclusions: *P. antipodarum* tolerates a wide range of environmental conditions. These findings are also reflected in the multi-target classification tree predicting the exotic mollusc community: at higher conductivities and if the ecological water quality was poor, both species were predicted as present.

Predicting the occurrence of one species can be relevant in certain cases, but river managers are, inspired by the WFD, interested in the evolution of the whole macroinvertebrate community. In our research, multi-target classification trees were built to predict the presence of multiple exotic species at once based on environmental conditions. This promising technique replaces a set of different single-target classification trees with a single tree that is easier to interpret for decision makers. However, the multi-target classification tree predicting the exotic macro-crustacean community resembled the single-target tree for *G. tigrinus*, which was the most abundant species. Species like *P. coxalis* and *P. meridanus* occurred in few samples and consequently, they were predicted as absent (probability <50%). For the exotic mollusc community, similar results were found: only the most dominant species (*P. acuta* and *P. antipodarum*) were predicted as present. Models with knowledge rules relevant for all species can be obtained if all species have the similar prevalences (e.g. 50%). However, the construction of such a dataset was not convenient as it would lead to a small number of records. Alternatively, the predictive ability of the model can be optimized, manipulating the threshold of probability of occurrence from 50% to 10% for example, so that less frequently encountered species can also be predicted as present. This manipulation would lead to better model performances for less frequently encountered species and would not affect the predictive ability of the model towards widespread invaders.

The reliability of the trees could be possibly further improved by the application of optimization techniques like genetic algo-

rhythms (D'heygere et al., 2003, 2006) as well as boosting and bagging methods (Dakou et al., 2007). However this was in this stage not relevant due to the relatively small dataset. Additionally, models and their derived analyses could be improved by the inclusion of other environmental variables. Dedecker et al. (2005) illustrated for instance that the ammonium concentration and the chemical oxygen demand were important variables to predict *Gammarus pulex* in rivers. Other methods for the improvement of the models are related to the integration of knowledge-based methods (Mouton et al., 2009), information from laboratory experiments (Boets et al., 2010) or data about extra environmental variables obtained via water quality and other models (Merckx et al., 2009). Additionally, more data points and eventually the application of modelling techniques that can deal with a lot of zero points, such as the zero-inflated count models, can be beneficial in terms of modelling performances (Lambert, 1992). Lee and Jin (2006) proposed a decision tree for zero-inflated count data, using a maximum of zero-inflated Poisson likelihood as the split criterion and found this tree more efficient than a classically grown tree. Implementing these optimizations will lead to useful multi-target classification trees as they integrate decisions for the whole set of exotic species. Multi-target classification trees are an interesting tool to create a clear juridical framework for dealing with aquatic invaders. They can help to convince stakeholders by showing the potential risks of several activities and the related impacts and they can be used for management planning and investment decisions.

## 5. Conclusions

Ecological indices based on coarse identification levels (such as the MMIF) assume that all species within the identification level have a similar sensitivity towards pollution. Since the introduction of pollution-tolerant invaders, this hypothesis is often violated and one has to be aware that this can lead to an overestimation of the ecological quality. As a solution, tolerance classes should be revised in case new invaders are identified in Flemish waters.

Based on a relatively small amount of data, data driven modelling techniques, such as classification trees, can be constructed with reasonable model performances. Predicting the occurrence of individual exotic species was possible by means of single-target classification trees. These models gave insight in the particular preferences of an exotic invasive species, whereas multi-target approaches gave an integrated insight in the potential exotic macroinvertebrate community subgroup, what is in particular relevant for water managers to protect and restore surface waters.

## Acknowledgements

We would like to thank the Flemish Environment Agency for the opportunity to study their samples. Koen Lock is currently supported by a post-doctoral fellowship from the Fund for Scientific Research (FWO-Vlaanderen, Belgium). We would like to thank Darko Aleksovski (from the Jozef Stefan Institute, Department Knowledge Technologies, Ljubljana, Slovenia) for the construction of the multi-target classification trees.

## References

Adam, W., 1960. Mollusques Terrestres et Dulcicoles. Royal Belgian Institute of Natural Sciences, Brussels, Belgium, 402 pp.

Adriaenssens, V., Goethals, P.L.M., De Pauw, N., 2006. Fuzzy knowledge-based models for prediction of *Asellus* and *Gammarus* in watercourses in Flanders (Belgium). *Ecol. Model.* 195, 3–10.

Alonso, A., Castro-Diez, P., 2008. What explains the invading success of the aquatic mud snail *Potamopyrgus antipodarum* (Hydrobiidae, Mollusca)? *Hydrobiologia* 614, 107–116.

Arbačiauskas, K., Semenchenko, V., Grabowski, M., Leuven, R.S.E.W., Paunović, M., Son, M.O., Csányi, B., Gumuliauskaitė, S., Konopacka, A., Nehring, S., van der

Velde, G., Vezhnovetz, V., Panov, V.E., 2008. Assessment of biocontamination of benthic macroinvertebrate communities in European inland waterways. *Aquat. Invasion* 3, 211–230.

Arndt, E., Fiedler, S., Böhme, D., 2009. Effects of invasive benthic macroinvertebrates on assessment methods of the EU Water Framework Directive. *Hydrobiologia* 635, 309–320.

Bij de Vaate, A., Jazdzewski, K., Ketelaars, H.A.M., Gollash, S., Van der Velde, G., 2002. Geographical patterns in range extension of Ponto-Caspian macroinvertebrate species in Europe. *Can. J. Fish. Aquat. Sci.* 59, 1159–1174.

Blockeel, H., De Raedt, L., Ramon, J., 1998. Top-down induction of clustering trees. In: Shavlik, J. (Ed.), *Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann, San Francisco (Madison July 24–27), pp. 55–63.

Blockeel, H., Struyf, J., 2002. Efficient algorithms for decision tree cross-validation. *J. Mach. Learn. Res.* 3, 621–650.

Boets, P., Lock, K., Cammaerts, R., Plu, D., Goethals, P.L.M., 2009. Occurrence of the invasive crayfish *Procambarus clarkii* (Girard, 1852) in Belgium (Crustacea: Cambaridae). *Belg. J. Zool.* 139, 173–176.

Boets, P., Lock, K., Messiaen, M., Goethals, P.L.M., 2010. Combining datadriven methods and lab studies to analyse the ecology of *Dikerogammarus villosus*. *Ecol. Inf.* 5, 133–139.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. *Classification and Regression Trees*. Wadsworth, Belmont, 358 pp.

Cardoso, A.C., Free, G., 2008. Incorporating invasive alien species into ecological assessment in the context of the Water Framework Directive. *Aquat. Invasion* 3, 361–366.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46.

Colautti, R.I., Bailey, S.A., van Overdijk, C.D.A., Amundsen, K., MacIsaac, H.J., 2006. Characterised and projected costs of nonindigenous species in Canada. *Biol. Invasion* 8, 45–59.

Cope, N.J., Winterbourn, M.J., 2004. Competitive interactions between two successful molluscan invaders of freshwaters: an experimental study. *Aquat. Ecol.* 38, 83–91.

Costil, K., Dussart, G.B.J., Daguzan, J., 2001. Biodiversity of aquatic gastropods in the Mont St-Michel basin (France) in relation to salinity and drying of habitats. *Biodivers. Conserv.* 10, 1–18.

Dakou, E., D'heygere, T., Dedecker, A., Goethals, P.L.M., De Pauw, N., Lazaridou-Dimitriadou, M., 2007. Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). *Aquat. Ecol.* 41, 399–411.

De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81, 3178–3192.

Dedecker, A., Goethals, P.L.M., D'heygere, T., Gevrey, M., Lek, S., De Pauw, N., 2005. Application of artificial neural network models to analyse the relationship between *Gammarus pulex* L. (Crustacea: Amphipoda) and river characteristics. *Environ. Monit. Assess.* 111, 223–241.

De Pauw, N., Vanhooren, G., 1983. Method for biological quality assessment of watercourses in Belgium. *Hydrobiologia* 100, 153–168.

Delaunois, H., 1982. *Landschapspark Krekengebied*. Bond Beter Leefmilieu, Brussels, Belgium, 700 pp.

Devin, S., Beisel, J.N., 2007. Biological and ecological characteristics of invasive species: a gammarid study. *Biol. Invasion* 9, 13–24.

D'heygere, T., Goethals, P.L.M., De Pauw, N., 2003. Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. *Ecol. Model.* 160, 291–300.

D'heygere, T., Goethals, P.L.M., De Pauw, N., 2006. Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. *Ecol. Model.* 195, 20–29.

Dumont, H.J., Gysels, H., 1971. Étude faunistique et écologique sur les criques de la Flandre orientale et le long de l'Escaut. Considérations sur leur chimisme, leur faune planktonique, entomologique et malacologique et discussion de leur état biologique actuel. *Ann. Soc. R. Zool. Belg.* 101, 157–181.

Džeroski, S., Gjorgioski, V., Slavkov, I., Struyf, J., 2007. Analysis of time series data with predictive clustering trees. In: *Proceedings of the Fifth International Workshop on Knowledge Discovery in Inductive Databases, LNCS 4747*, pp. 63–80.

EU, 2000. EC Water Framework Directive (2000/60/EC). Official Journal of the European Communities, European Commission, Brussels, Belgium.

Gabriels, W., Goethals, P.L.M., De Pauw, N., 2005. Implications of taxonomic modifications and alien species on biological water quality assessment as exemplified by the Belgian Biotic Index method. *Hydrobiologia* 542, 137–150.

Gabriels, W., Goethals, P.L.M., Dedecker, A.P., Lek, S., De Pauw, N., 2007. Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networks. *Aquat. Ecol.* 41, 427–441.

Gabriels, W., Lock, K., De Pauw, N., Goethals, P.L.M., 2010. Multimetric macroinvertebrate index flanders (MMIF) for biological assessment of rivers and lakes in Flanders (Belgium). *Limnologia* 40, 199–207.

Gérard, C., Blanc, A., Costil, K., 2003. *Potamopyrgus antipodarum* (Mollusca: Hydrobiidae) in continental aquatic gastropod communities: impact of salinity and trematode parasitism. *Hydrobiologia* 493, 167–172.

Goethals, P.L.M., Dedecker, A.P., Gabriels, W., Lek, S., De Pauw, N., 2007. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. *Aquat. Ecol.* 41, 491–508.

Grabowski, M., Konopacka, A., Jazdzewski, K., Janowska, E., 2006. Invasions of alien gammarid species and retreat of natives in the Vistula Lagoon (Baltic Sea Poland). *Helgol. Mar. Res.* 60, 90–97.

Gittenberger, E., Janssen, A.W., Kuiper, W.J., Kuiper, J.G.J., Meijer, T., Van der Velde, G., Peeters, G.A., 1998. *De Nederlandse Zoetwatermollusken. Recente en Fossiele*

- Weekdieren uit Zoet en Brak Water. KNNV Uitgeverij & EIS-Nederland, Leiden, Nederland, 288 pp.
- Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.* 24, 38–49.
- Jennings, M.D., 2000. Gap analysis: concepts, methods, and recent results. *Landscape Ecol.* 15, 5–20.
- Karatayev, A.Y., Burlakova, L.E., Padilla, D.K., Mastitsky, S.E., Olenin, S., 2009. Invaders are not a random selection of species. *Biol. Invasion* 11, 2009–2019.
- Keppens, M., Keppens, D., 1996. Verspreiding van de Land-En Zoetwatermollusken van Dendermonde. Royal Belgian Institute of Natural Sciences, Brussels, Belgium, 36 pp.
- Kocev, D., Džeroski, S., White, M.D., Newell, G.R., Griffioen, P., 2009. Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. *Ecol. Model.* 220, 1159–1168.
- Koop, J.H.E., Grieshaber, M.K., 2000. The role of ion regulation in the control of the distribution of *Gammarus tigrinus* (Sexton) in salt-polluted rivers. *J. Comp. Physiol. B* 170, 75–83.
- Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34, 1–14.
- Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.
- Lee, S., Jin, S., 2006. Decision tree approaches for zero-inflated count data. *J. Appl. Stat.* 33, 853–865.
- Leppäkoski, E., Olenin, S., 2000. Non-native species and rates of spread: lessons from the brackish Baltic Sea. *Biol. Invasion* 2, 151–163.
- Leung, B., Bossenbroek, J.M., Lodge, D.M., 2006. Boats, pathways, and aquatic biological invasions: estimating dispersal potential with gravity models. *Biol. Invasion* 8, 241–254.
- Linke, S., Bailey, R.C., Schwindt, J., 1999. Temporal variability of stream bioassessments using benthic macroinvertebrates. *Freshw. Biol.* 42, 575–584.
- Lodge, D.M., Williams, S., MacIsaac, H.J., Hayes, K.R., Leung, B., Reichard, S., Mack, R.N., Moyle, P.B., Smith, M., Andow, D.A., Carlton, J.T., McMichael, A., 2006. Biological invasions: recommendations for U.S. policy and management. *Ecol. Appl.* 16, 2035–2054.
- Lovell, S.J., Stone, S.F., 2006. The economic impacts of aquatic invasive species: a review of the literature. *Agric. Resour. Econ. Rev.* 35, 195–208.
- MacNeil, C., Briffa, M., 2009. Replacement of a native freshwater macroinvertebrate species by an invader: implications for biological water quality monitoring. *Hydrobiologia* 635, 321–327.
- MacNeil, C., Briffa, M., Leuven, R.S.E.W., Gell, F.R., Selman, R., 2010. An appraisal of a biocontamination assessment method for freshwater macroinvertebrate assemblages; a practical way to measure a significant biological pressure? *Hydrobiologia* 638, 151–159.
- Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. *J. Appl. Ecol.* 38, 921–931.
- Merckx, B., Goethals, P., Steyaert, M., Vanreusel, A., Vincx, M., Vanaverbeke, J., 2009. Predictability of marine nematode biodiversity. *Ecol. Model.* 220, 1449–1458.
- Messiaen, M., Lock, K., Gabriels, W., Vercauteren, T., Wouters, K., Boets, P., Goethals, P.L.M., 2010. Alien macrocrustaceans in freshwater ecosystems in the eastern part of Flanders (Belgium). *Belg. J. Zool.* 140, 30–39.
- Mills, E.L., Leach, J.H., Carlton, J.T., Secor, C.L., 1993. Exotic species in the Great Lakes: a history of biotic crises and anthropogenic introductions. *J. Great Lakes Res.* 19, 1–54.
- Mouton, A.M., De Baets, B., Van Broekhoven, E., Goethals, P.L.M., 2009. Prevalence-adjusted optimisation of fuzzy models for species distribution. *Ecol. Model.* 220, 1776–1786.
- Naylor, R.L., Williams, S.L., Strong, D.R., 2001. Aquaculture—a gateway for exotic species. *Science* 294, 1655–1656.
- Normant, M., Feike, M., Szaniawska, A., Graf, G., 2007. Adaptation of *Gammarus tigrinus* Sexton 1939 to new environments: some metabolic investigations. *Thermochim. Acta* 458, 107–111.
- Olenin, S., Minchin, D., Daunys, D., 2007. Assessment of biopollution in aquatic ecosystems. *Mar. Pollut. Bull.* 55, 379–394.
- Orendt, C., Schmitt, C., van Liering, C., Wolfram, G., de Deckere, E., 2010. Include or exclude? A review on the role and suitability of aquatic invertebrate neozoa as indicators in biological assessment with special respect to fresh and brackish European waters. *Biol. Invasion* 12, 265–283.
- Pimentel, D., Lach, L., Zuniga, R., Morrison, D., 2000. Environmental and economic costs of nonindigenous species in the United States. *Bioscience* 50, 53–65.
- Pimentel, D., Zuniga, R., Morrison, D., 2005. Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecol. Econ.* 52, 273–288.
- Piscart, C., Moreteau, J.C., Beisel, J.N., 2005. Biodiversity and structure of macroinvertebrate communities along a small permanent salinity gradient (Meurthe River, France). *Hydrobiologia* 551, 227–236.
- Quinlan, J.R., 1986. Induction of decision trees. *Mach. Learn.* 1, 81–106.
- Ricciardi, A., MacIsaac, H.J., 2000. Recent mass invasion of the North American Great Lakes by Ponto-Caspian species. *Trends Ecol. Evol.* 15, 62–65.
- Ricciardi, A., 2003. Predicting the impacts of an introduced species from its invasion history: an empirical approach applied to zebra mussel invasions. *Freshw. Biol.* 48, 972–981.
- Scott, J.M., Davis, F., Csuti, B., Noss, R., Butterfield, B., Groves, C., Anderson, H., Caicco, S., Derchia, F., Edwards, T.C., Ulliman, J., Wright, R.G., 1993. Gap analysis: a geographic approach to protection of biological diversity. *Wildl. Monogr.* 123, 1–41.
- Sim, J., Wright, C.C., 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys. Ther.* 85, 257–268.
- SPSS, 2008. SPSS 16.0 for Windows, Rel. 16.0.2. Chicago, SPSS Inc.
- Struyf, J., Džeroski, S., 2006. Constraint based induction of multi-objective regression trees. In: *Proceedings of the Fourth International Workshop on Knowledge Discovery in Inductive Databases, LNCS 3933*, pp. 222–233.
- Hoang, T.H., Lock, L., Mouton, A., Goethals, P.L.M., 2009. Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. *Ecol. Inform.* 5, 140–146.
- Van Huet, H.J.W.J., 1992. Phosphorus eutrophication in the SW Frisian lake district. Phosphorus balances and simulation of reduction scenarios. *Hydrobiologia* 233, 271–281.
- van Puijenbroek, P.J.T.M., Janse, J.H., Knoop, J.M., 2004. Integrated modelling for nutrient loading and ecology of lakes in The Netherlands. *Ecol. Model.* 174, 127–141.
- Vander Zanden, M.J., Olden, J.D., 2008. A management framework for preventing the secondary spread of aquatic invasive species. *Can. J. Fish. Aquat. Sci.* 65, 1512–1522.
- Vens, C., J. Struyf, J., Schietgat, L., Džeroski, S., Blockeel, H., 2008. Decision trees for hierarchical multi-label classification. *Mach. Learn.* 73, 185–214.
- Vitousek, P.M., D'Antonio, C.M., Loope, L.L., Westbrooks, R., 1996. Biological invasions as global environmental change. *Am. Sci.* 84, 468–478.
- Walley, W.J., Hawkes, H.A., 1996. A computer-based reappraisal of the biological monitoring working party scores using data from 1990 river quality survey of England and Wales. *Water Res.* 9, 2086–2094.
- Wijnhoven, S., van Riel, M.C., Van der Velde, G., 2003. Invasive and indigenous freshwater gammarid species: physiological tolerance to water temperature in relation to ionic content of water. *Aquat. Ecol.* 37, 151–158.
- Witten, I.H., Frank, E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 560 pp.