Available online at www.sciencedirect.com
ScienceDirect
International Journal of Approximate Reasoning
44 (2007) 65–90
INTERNATIONAL JOURNAL OF APPROXIMATE REASONING
www.elsevier.com/locate/ijar

Interpretability-preserving genetic optimization of linguistic terms in fuzzy models for fuzzy ordered classification: An ecological case study

Ester Van Broekhoven [a,*], Veronique Adriaenssens [b,c], Bernard De Baets [a]

[a] *Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, B-9000 Gent, Belgium*

[b] *Department of Applied Ecology and Environmental Biology, Ghent University, Jozef Plateaustraat 22, B-9000 Gent, Belgium*

[c] *Environment Agency, Evenlode House, Howbery Park, Wallingford, Oxon OX10 8BD, United Kingdom*

Received 18 July 2005; received in revised form 22 February 2006; accepted 16 March 2006
Available online 14 July 2006

Abstract

Fuzzy ordered classifiers were used to assign fuzzy labels to river sites expressing their suitability as a habitat for a certain macroinvertebrate taxon, given up to three abiotic properties of the considered river site. The models were built using expert knowledge and evaluated on data collected in the Province of Overijssel in the Netherlands. Apart from a performance measure for crisp classifiers common in the aquatic ecology domain, the percentage of correctly classified instances (% CCI), two performance measures for fuzzy (ordered) classifiers are introduced in this paper: the percentage of correctly fuzzy classified instances (% CFCI) and the average deviation (AD). Furthermore, results of an interpretability-preserving genetic optimization of the linguistic terms, applying once binary encoding and once real encoding, are presented.

Keywords: Habitat suitability models; Fuzzy ordered classifiers; Linguistic fuzzy models; Genetic algorithms; Interpretability

* Corresponding author. Tel.: +32 9 264 6018; fax: +32 9 264 6220.
E-mail addresses: Ester.VanBroekhoven@UGent.be (E. Van Broekhoven), Veronique.Adriaenssens@environment-agency.gov.uk (V. Adriaenssens), Bernard.DeBaets@UGent.be (B. De Baets).

0888-613X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2006.03.003

1. Introduction

New requirements at the EU level, mainly covered by the Water Framework Directive [1], urge the member states to extend their assessment methodologies to implement the desired river management. A methodology of interest in this context is the modelling of habitat suitability. Habitat suitability models describe which abiotic conditions are appropriate for a certain taxon or species to establish a population [2]. In this study benthic macroinvertebrates living in aquatic ecosystems are considered. Benthic macroinvertebrates are invertebrate organisms that inhabit mainly bottom substrates of freshwater habitats [3]. The term 'macro' assumes that they are large enough to be seen without magnification and that they are retained in a net with mesh size of 500 μm. Because of their central role in aquatic ecosystems, macroinvertebrates are widely used as indicators for assessing freshwater quality [4].

The development of habitat suitability models is not an easy task. The available knowledge is usually only verbally described, with terminology and meaning differing from source to source. On the other hand, data available is not only scarce, but insufficiently representative for all river conditions, and can therefore play at most a role in model optimization, but not in model identification [5,6]. Taking into account these limitations and the ultimate use of these models in decision support, requiring understandability to the end user, we opted for linguistic fuzzy models and a knowledge-based design approach followed by an interpretability-preserving data-driven optimization of the membership functions.

As will be explained further on, this modelling problem asks for a model that gives a shaded indication of a certain river site's suitability as habitat for a certain macroinvertebrate taxon. Therefore, fuzzy classifiers were applied in this study, instead of classical models with crisp outputs or crisp classifiers. A more detailed description of the habitat suitability models, built using expert knowledge described in literature, is given in Section 2. In Section 3, the data collected in the Province of Overijssel in the Netherlands [7] on which the models were evaluated, referred to in this work as the EKOO data set, is discussed. The three measures used to evaluate the models, percentage of correctly classified instances (% CCI), percentage of correctly fuzzy classified instances (% CFCI) and average deviation (AD) are presented in Section 4. The membership functions of the input variables of the models of 12 taxa were optimized using a classic genetic algorithm with binary chromosomes, as well as a real-coded genetic algorithm. During the model optimization the accuracy of the models was tried to be improved, while preserving their interpretability. Section 5 deals with the different aspects of the optimization of the linguistic terms: the selection of the models to be optimized, the properties of the genetic algorithms and the obtained results. Finally, conclusions are summarized in Section 6.

2. Habitat suitability models

The models considered in this study describe the suitability of river sites along source brooks up to small rivers in the Central and Western Plains of Europe, a region defined in [8], as a habitat for the 86 macroinvertebrate taxa listed in Appendix A. The model development was based on eight knowledge sources (references are given in [9]) summarizing ecological studies carried out in the Netherlands, France, Germany and Belgium, describing which river conditions are preferred and which situations are tolerated by

different macroinvertebrate taxa. Hereafter, different aspects of the model identification process are discussed: the selection of variables, the assignment of linguistic values and corresponding membership functions to all variables and the construction of rule bases.

As described in detail in [9], the selected input variables should be of high ecological importance to the macroinvertebrate taxa under study as well as to the whole macroinvertebrate community and should be of importance to river management. Furthermore, knowledge about their preferences for certain environmental conditions needs to be available and the variables need to be included in the EKOO data set. Physical variables do provide effective assessment criteria when rivers are not affected by physical–chemical degradation [10]. However, in the Central and Western Plains of Europe, the main threats for biological communities in rivers are the deteriorated physical–chemical water quality conditions. This is mainly due to increased nutrient and organic loading mainly caused by agricultural activities and pollution originating from households. Therefore, apart from stream width and stream velocity, two variables determining the river type and reflecting the water quantity conditions, an additional input variable is used, expressing the physical–chemical conditions at a river site. Physical–chemical conditions and their effect on the macroinvertebrate population at a certain river site can be expressed by the saprobic status (measured by the ammonium concentration), the trophic status (measured by the nitrate and phosphate concentration) or the ionic status (measured by the electrical conductivity). As, in the region considered in this habitat suitability modelling problem, high (resp. low) nitrate concentrations generally coincide with high (resp. low) phosphate, ammonium and overall nutrient concentrations, all four variables can be used individually as a measure of one of the factors influencing the abundance of macroinvertebrates, i.e., the nutrient and organic load in the river. For each macroinvertebrate taxon, four different models were constructed, an A-model, an N-model, a P-model and a C-model, containing stream width, stream velocity and either ammonium concentration (A), nitrate concentration (N), phosphate concentration (P) or electrical conductivity (C) as input variables. The occurrence of some of the 86 considered macroinvertebrate taxa is independent of the stream width. In these models stream width is not included and only two input variables are used.

Due to the different context of the studies described in the eight publications used as a source of expert knowledge, meanings given to the used linguistic terms are not identical in all eight publications. However, in all considered studies, a similar number of linguistic values is assigned to variables as stream width, stream velocity and nutrient and organic loading and in most cases similar expressions are applied to refer to the different situations distinguished. To all variables three to five linguistic values are assigned. An overview of the linguistic values is given in Table 1. All values are defined by trapezial membership functions forming a Ruspini partition [11], as illustrated in Fig. 1(a) for the five linguistic values for ammonium concentration (in order of increasing organic load): *oligosaprobic*, β,α-*oligosaprobic*, β-*mesosaprobic*, α-*mesosaprobic* and *polysaprobic* conditions. The values of the membership function parameters of all variables summarized in Table 2, are based on crisp boundaries found in literature. The kernel of each of the membership functions is the intersection of the crisp intervals used in the different literature sources to define the corresponding linguistic term. As we have opted for fuzzy partitions, the supports of the membership functions are determined by the kernels of the membership functions of the adjacent linguistic values and the lower and upper bounds of the underlying domain.

Table 1
Linguistic values assigned to the input and output variables

Variable	Linguistic values
Stream width	{spring/small stream, upper course stream, middle course stream, lower course stream/small river}
Stream velocity	{low, moderate, high}
Ammonium concentration	{oligosaprobic, β,α-oligosaprobic, β-mesosaprobic, α-mesosaprobic, polysaprobic}
Nitrate concentration	{oligotrophic, β-mesotrophic, α-mesotrophic, eutrophic, hypertrophic}
Phosphate concentration	{oligotrophic, β-mesotrophic, α-mesotrophic, eutrophic, hypertrophic}
Electrical conductivity	{oligoionic, β-mesoionic, mesoionic, α-mesoionic, polyionic}
Abundance	{absent, low, moderate, high}

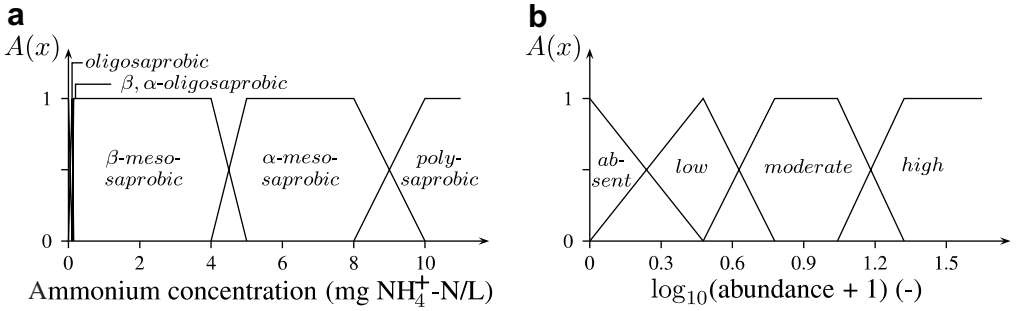

Fig. 1. Definition of the five linguistic values assigned to ammonium concentration and the four fuzzy abundance classes through membership functions.

Table 2
Parameters of the membership functions defining the linguistic values in Table 1

Variable	Membership function parameters
Width (m)	{0, 0, 2, 2, 4, 4, 6, 201}
Velocity (m/s)	{0, 0, 0.25, 0.25, 0.5, 1.2}
Ammonium conc. (mg NH_4^+-N/L)	{0, 0, 0.10, 0.10, 0.15, 4, 5, 8, 10, 30}
Nitrate conc. (mg NO_3^--N/L)	{0, 0, 0.15, 0.15, 0.3, 0.3, 0.4, 0.4, 0.45, 112}
Phosphate conc. (mg PO_4^{3-}-P/L)	{0, 0, 0.008, 0.008, 0.015, 0.015, 0.025, 0.025, 0.045, 5.45}
Conductivity (μS/cm)	{0, 150, 250, 450, 550, 750, 850, 1050, 1150, 2880}
$\log_{10}$(abundance + 1) (–)	{0, 0, 0.477121, 0.477121, 0.778151, 1.041393, 1.322219, 3.602169}

Fig. 2 shows how the parameters should be interpreted.

A site's suitability as a habitat for macroinvertebrates cannot be measured directly. As output variable of the developed habitat suitability models, the abundance of a macroinvertebrate taxon at a river site is used. The abundance is a measure for habitat suitability: the higher the abundance of a taxon, the higher the site's suitability as a habitat. Furthermore the EKOO data set contains the number of sampled individuals of the 86 taxa considered at all investigated river sites. It cannot be the purpose of a habitat suitability model to predict a precise numerical value for the occurrence of a given taxon. No ecologist is interested in or would even trust a model stating an occurrence of, e.g., 37 individuals. It is rather the magnitude of the abundance which is of interest. In this paper four linguistic values were assigned to the variable: *absent*, *low*, *moderate* and *high*. They are defined

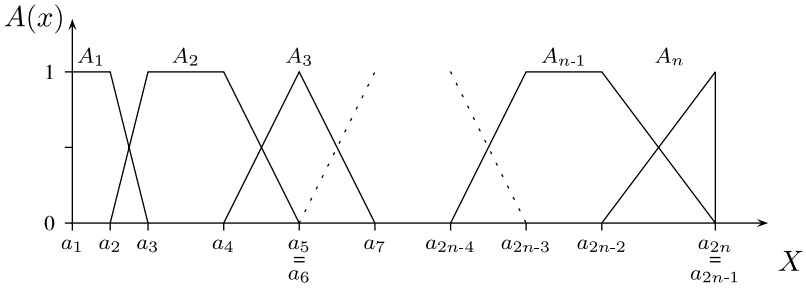

Fig. 2. To characterize n trapezial membership functions forming a fuzzy partition, $2n$ parameters were used.

by the membership functions shown in Fig. 1(b) with the help of the same experts assigning the membership functions of the input variables. In order to take into account the non-linear response of macroinvertebrate taxa to environmental conditions [12], the abundance values were log-transformed. When comparing abundance values relative differences rather than absolute differences should be considered, since the difference between 1 and 2 individuals found at a river site is more significant than the difference between 101 and 102 recorded individuals. We also want to stress that these abundance values are not equal to the exact number of individuals present at a site, but are proportional to the number of individuals present at a site (see the sampling procedures in Section 3).

The four linguistic values of stream width, the three linguistic values of stream velocity and the five linguistic values of the variables describing the nutrient and organic concentration, define 60 environmental situations. For the procedure followed during the rule base development, i.e., the assignment of a linguistic abundance value to this 60 environmental situations, we refer to [9]. In the A-, N-, P- and C-models of the 86 macroinvertebrate taxa, including respectively, ammonium concentration, nitrate concentration, phosphate concentration and electrical conductivity as input variables, the same membership functions were used. The rule bases of the models of the different taxa differed, but were identical for the four models of a certain taxon [13]. All constructed rule bases were complete and contained 60 rules of the following type:

IF *width* IS *upper course stream*
AND *velocity* IS *low*
AND *nitrate concentration* IS *eutrophic*
THEN *abundance* IS *moderate*

In Fig. 3 the rule base of *Proasellus meridianus* is shown. *Proasellus meridianus* is an example of a taxon whose occurrence is independent of the stream width, as one can see from the rule base. Furthermore, according to the rules derived from the eight consulted knowledge sources, its abundance is the same in oligosaprobic (resp. oligotrophic and oligoionic) conditions as in β,α-oligosaprobic (resp. β-mesotrophic and β-mesoionic) conditions. If two consecutive linguistic values of a variable yield the same model output for all combinations of linguistic values of the other input variables, then the corresponding rules are merged and a new linguistic value is introduced defined as the convex hull of the membership functions of the original linguistic values. Therefore, in the reduced model the variables ammonium, nitrate and phosphate concentration and conductivity, take four

Proasellus meridianus		stream width			
		spring / small stream	upper course stream	middle course stream	lower course stream / small river
stream velocity	oligosaprobic / oligotrophic / oligoionic				
	low	Absent	Absent	Absent	Absent
	moderate	Low	Low	Low	Low
	high	Absent	Absent	Absent	Absent
	β,α-oligosaprobic / β-mesotrophic / β-mesoionic				
	low	Absent	Absent	Absent	Absent
	moderate	Low	Low	Low	Low
	high	Absent	Absent	Absent	Absent
	mesosaprobic / α-mesotrophic / mesoionic				
	low	Low	Low	Low	Low
	moderate	Moderate	Moderate	Moderate	Moderate
	high	Low	Low	Low	Low
	α-mesosaprobic / eutrophic / α-mesoionic				
	low	Moderate	Moderate	Moderate	Moderate
	moderate	High	High	High	High
	high	Moderate	Moderate	Moderate	Moderate
	polysaprobic / hypertrophic / polyionic				
	low	Absent	Absent	Absent	Absent
	moderate	Low	Low	Low	Low
	high	Low	Low	Low	Low

Fig. 3. Rule base of the four models describing the habitat suitability for *Proasellus meridianus*.

values instead of five, for ammonium concentration these linguistic values are 'oligosaprobic to β,α-oligosaprobic', 'β-mesosaprobic', 'α-mesosaprobic' and 'polysaprobic' conditions. The linguistic value 'oligosaprobic to 'β,α-oligosaprobic' conditions is defined as the convex hull of the membership function of 'oligosaprobic' conditions and the membership function of 'β,α-oligosaprobic' conditions. As a result of the reduction of input variables and linguistic values, the number of rules in the rule base decreases. The rule base of the resulting, fully reduced model for *Proasellus meridianus* is shown in Fig. 4. This model reduction procedure is carried out for the models of all 86 taxa, resulting in models with different numbers of input variables, membership functions and number of rules.

Given the available qualitative expert knowledge and uncertainty in the definitions of the used linguistic expressions, linguistic fuzzy models are the most appropriate model types for the modelling problem. As no crisp abundance value, but a shaded indication of a site's habitat suitability is desired in river management, we opted for a fuzzy classification. The model output $\mathbf{y}_{\text{model}}$ is a set of four values ranging between 0 and 1

Proasellus meridianus		stream velocity		
		low	moderate	high
saprobic / trophic / ionic status	oligosaprobic to β,α-oligosaprobic / oligotrophic to β-mesotrophic / oligoionic to β-mesoionic	Absent	Low	Absent
	mesosaprobic / α-mesotrophic / mesoionic	Low	Moderate	Low
	α-mesosaprobic / eutrophic / α-mesoionic	Moderate	High	Moderate
	polysaprobic / hypertrophic / polyionic	Absent	Low	Low

Fig. 4. Reduced rule base of the four models describing the habitat suitability for *Proasellus meridianus*.

and summing up to 1: {(*absent*, $A_1(\mathbf{y}_{\text{model}})$), (*low*, $A_2(\mathbf{y}_{\text{model}})$), (*moderate*, $A_3(\mathbf{y}_{\text{model}})$), (*high*, $A_4(\mathbf{y}_{\text{model}})$)}. Due to the inherent order on the terms 'absent', 'low', 'moderate' and 'high', this is a clear example of fuzzy ordered classification. When calculating the fulfilment degrees of the rules, the minimum t-norm was applied for the conjunction. For each linguistic abundance value, the maximum fulfilment degree of the rules containing this linguistic abundance value in their consequent is determined. Finally, the model output is obtained by normalizing these maximum fulfilment degrees. Note that the membership functions in the output domain are not used to determine the model output.

3. The EKOO data set

The data used in this study to evaluate and optimize the habitat suitability models were collected in running waters in the Province of Overijssel in the Netherlands. They are part of a larger data set described by Verdonschot [7], which apart from the 445 data points collected along running waters and used in this study, also includes data collected in pools and lakes, canals and large standing waters.

At each site, 70 abiotic variables were measured, as stream width, depth, temperature, transparency of the water column, bank shape, substratum, dissolved oxygen concentration, pH, nitrate concentration and phosphate concentration, and samples were taken of the major habitats, the water body and the bottom habitat to collect macroinvertebrates. In shallow sites, habitats with vegetation were sampled by sweeping a hand net (20×30 cm, mesh size 500 μm) through each vegetation type, several times over a length of 0.5–1 m. Bottom habitats were sampled by vigorously pushing the hand net through the upper few centimeters of each type of substratum over a length of 0.5–1 m. Next, the habitat samples were combined for the site to give a single sample with a standard area of 1.5 m^2 (1.2 m^2 of vegetation and 0.3 m^2 of bottom). At sites lacking vegetation, the standard sampling was confined to the bottom habitats. In deeper sites, five samples from the bottom habitats were taken with an Ekman-Birge sampler. These five grab-samples were equivalent to one 0.5 hand net bottom sample. The macroinvertebrate samples were taken to the laboratory, sorted by eye, counted and identified to species level, except for some chironomid taxa.

In this work the term 'EKOO data set' does not refer to the complete data set described in [7], but only to those data used in this study: the values of the six abiotic variables, stream width, stream velocity, ammonium concentration, nitrate concentration, phosphate concentration and electrical conductivity, and the number of sampled individuals of the 86 macroinvertebrate taxa listed in Appendix A at 445 sites along running waters.

Hours of field work and meticulous determination in the lab of the sampled animals were needed to obtain this data set, which makes it a large data set in its domain, but unfortunately still rather small for model evaluation and certainly for model identification purposes. Apart from being sparse, the data hold another awkward property typical to their origin: due to seasonal variations, weather differences at sampling moment and different sediments, data holding similar values for the considered environmental variables show highly variable registered abundances. This is illustrated in Fig. 5 for the A-model of *Proasellus meridianus*. Therefore, this data cannot be expected to reveal an unambiguous relationship between the selected abiotic variables and macroinvertebrate abundance. At a vast majority of the sites no individuals were recorded for all 86 taxa considered in this study as illustrated for *Proasellus meridianus* and *Plectronemia conspersa* in Fig. 6 and discussed in more detail in Section 5.2.

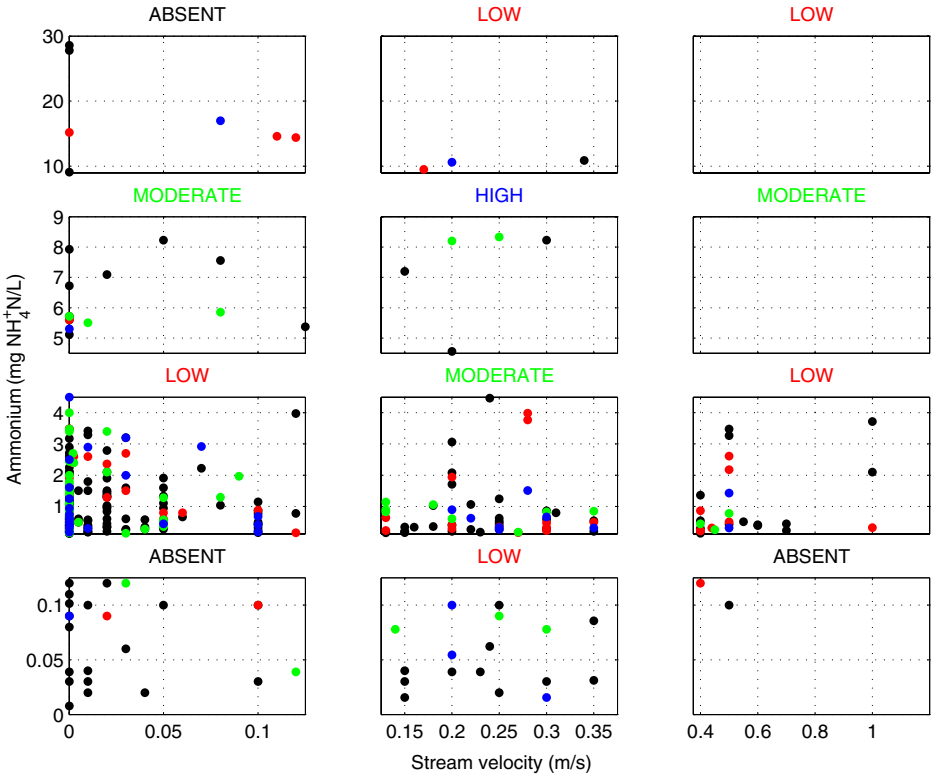

Fig. 5. Data points in the different parts of the input space defined by the 0.5-cuts of the membership functions of velocity and ammonium concentration. The points are coloured according to the crisp (see Eq. (2) for the defuzzification procedure) abundance classes to which the measured abundance of *Proasellus meridianus* belongs.

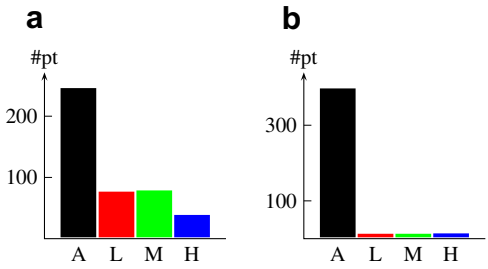

Fig. 6. Distribution of the data points over the crisp abundance classes *absent*, *low*, *moderate* and *high* for (a) *Proasellus meridianus* and (b) *Plectronemia conspersa* (see Eq. (2) for the defuzzification procedure).

4. Evaluation of fuzzy ordered classifiers

4.1. Format of the reference output

In order to compare the output obtained with the fuzzy ordered classifiers to the information in the EKOO data set, model and reference output should have the same format.

In this study the membership degrees of the crisp abundance values in the data set to the linguistic abundance values, defined by membership functions shown in Fig. 1(b), are used as reference output.

4.2. Three performance measures

In this section three performance measures applied in this study are introduced. In the formulae below, N is the number of data points, n the number of fuzzy classes, $A_i(\mathbf{y}_{\text{data},j})$ the membership degree of the jth output to the ith linguistic output value and $A_i(\mathbf{y}_{\text{model},j})$ the membership degree to the ith linguistic output value obtained as model output for the jth input of the data set.

4.2.1. Percentage of correctly classified instances

In ecology the percentage of correctly classified instances (% CCI) is frequently used to compare the performance of crisp classifiers [14]. Correctly classified data points have a contribution of 1 to the global performance, while data points assigned to a wrong class have a contribution of 0. In order to be able to compare our fuzzy classifiers to crisp classifiers in literature, the outputs were defuzzified and the % CCI was calculated as follows:

$$\%\ \text{CCI} = \frac{100}{N}\sum_{j=1}^{N}\left(1-\frac{1}{2}\sum_{i=1}^{n}\left|A_{\text{crisp},i}(\mathbf{y}_{\text{data},j})-A_{\text{crisp},i}(\mathbf{y}_{\text{model},j})\right|\right) \tag{1}$$

with

$$A_{\text{crisp},i}(\mathbf{y}) = \begin{cases} 1 & \text{if } i = \min\{k|A_k(\mathbf{y}) = \max_{l=1}^{n} A_l(\mathbf{y})\}, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

4.2.2. Percentage of correctly fuzzy classified instances

As we are dealing with fuzzy classifiers, we defined a new performance measure inspired by the % CCI and similar to the measure presented in [15]: the percentage of correctly fuzzy classified instances (% CFCI). If the model output is identical to the reference output, the data point has a contribution of 1 to the global performance. As long as there are classes to which both model output and reference output have a non-zero membership degree, the corresponding data point has a positive contribution. Only if no class exists to which both model output and reference output have a non-zero membership degree, the corresponding data point has a contribution of 0 to the global performance:

$$\%\ \text{CFCI} = \frac{100}{N}\sum_{j=1}^{N}\left(1-\frac{1}{2}\sum_{i=1}^{n}\left|A_i(\mathbf{y}_{\text{data},j})-A_i(\mathbf{y}_{\text{model},j})\right|\right). \tag{3}$$

4.2.3. Average deviation

The % CFCI has the advantage that it can be understood intuitively. However, it is not an appropriate objective function for the optimization of a fuzzy ordered classifier, as % CFCI is not sensitive to the position of the classes where the wrong classification occurs. When visually comparing the reference output in Table 3 with the model outputs b and d and given the fact that the output classes are ordered from A_1 to A_4, one would certainly

Table 3
Four fuzzy classification examples and their corresponding performances expressed by % CCI, % CFCI and AD

	$\mathbf{y}_{\text{data}}$				$\mathbf{y}_{\text{model}}$				% CCI	% CFCI	AD
	A_1	A_2	A_3	A_4	A_1	A_2	A_3	A_4			
a	0	0.2	0.8	0	0.8	0.2	0	0	0	20	1.6
b	0	0.2	0.8	0	0	0.4	0.6	0	100	80	0.2
c	0	0.2	0.8	0	0	0.1	0.8	0.1	100	90	0.2
d	0	0.2	0.8	0	0	0	0.8	0.2	100	80	0.4

say that model output b approximates the reference output better than model output d. However, that same % CFCI is assigned to examples b and d, as the sum of the absolute differences in membership degree in the reference and model output to the four individual classes is identical, as shown in Fig. 7.

Therefore, another performance measure for fuzzy classifiers with an ordered set of classes is introduced, returning the average deviation (AD) between the position of the class obtained with the model and the position of the class stored in the reference data set. The AD varies from 0 to $n-1$ and is calculated as follows:

$$\text{AD} = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n-1}\left|\sum_{k=1}^{i}A_k(\mathbf{y}_{\text{data},j}) - \sum_{k=1}^{i}A_k(\mathbf{y}_{\text{model},j})\right|. \tag{4}$$

The measure AD is illustrated in Table 3 on the same examples as the two other performance measures. At first sight it seems hard to get insight in AD. When considering the cumulative membership degrees, i.e., the sum of the membership degrees to a class and

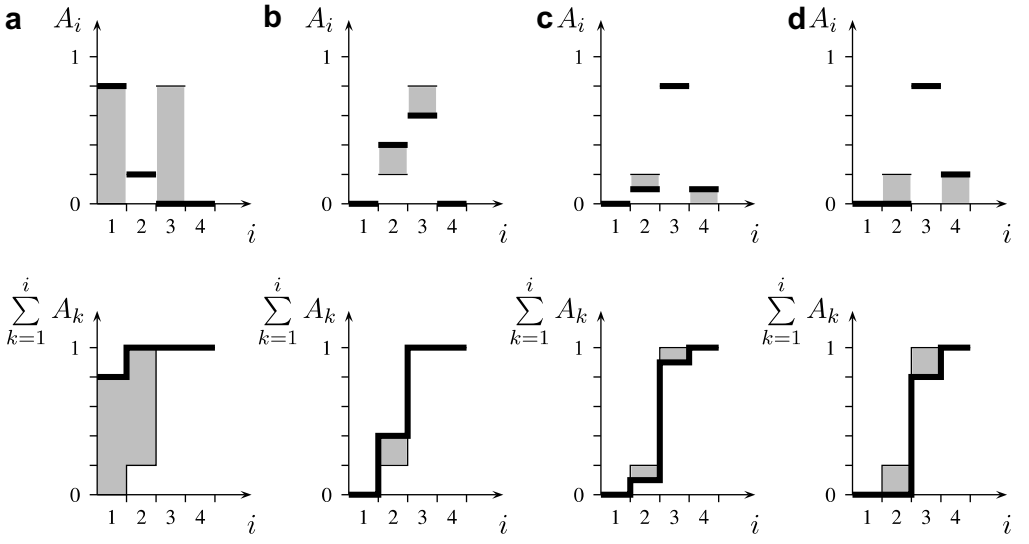

Fig. 7. Illustration of the performance measures % CFCI and AD for the fuzzy classification examples in Table 3. In the figures in the top row, illustrating % CFCI, the thin and thick lines indicate the reference and model output, respectively. In the figures in the second row, illustrating AD, the thin and thick lines are the cumulative functions of the reference and model output, respectively.

its lower classes as in Fig. 7, instead of the membership degrees, one sees that the AD is nothing else but the area between the cumulative functions of model and reference output.

The AD is zero if the model output equals the reference output and increases with increasing distance between the reference output and the model output. The AD distinguishes between examples b and d, whereas the % CFCI does not. On the other hand, the same AD, but a different % CFCI, is obtained for examples b and c. In example b the membership degree assigned to class A_2 is 0.2 too high. This surplus of membership degree should in fact be assigned to the adjacent class A_3. In example c the membership degree assigned to class A_4 is 0.1 too high and this surplus of membership degree should in fact have been assigned to class A_2, i.e., two classes lower. The distance between the reference output is therefore 1×0.2 for example b and 2×0.1 for example c. The % CFCI however is a measure of the sum of the errors made for each individual class. For example b the error in membership degree is 0.2 for the two classes A_2 and A_3, whereas in example d the errors are 0.1 for the two classes A_2 and A_4.

Note that the AD is insensitive to the direction of the wrong classification as the absolute values of the differences are taken. If classifying an instance in a too high class is worse (or better) than classifying it in a too low class, the AD should be computed using the same formula as Eq. (4) but without taking absolute values of the differences.

4.3. Model performance

In Fig. 8 the three performance values obtained for the four models of the 86 macroinvertebrate taxa are plotted. One sees that similar values are obtained for % CCI as for its fuzzy alternative, % CFCI, and that AD tends to decrease with increasing % CFCI. The % CFCI of the A-, N-, P- and C-models of all taxa are shown in Fig. 9. For almost all taxa, higher % CFCI-values are obtained for models including nitrate or phosphate concentration as input variable than for those including ammonium concentration or conductivity. The obtained model performances are discussed in more detail in [16].

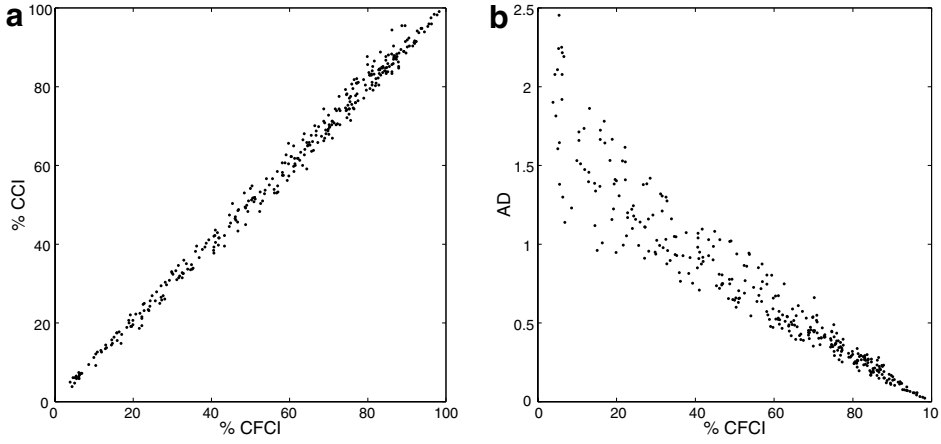

Fig. 8. Comparison of the % CFCI-values to (a) the % CCI- and (b) AD-values.

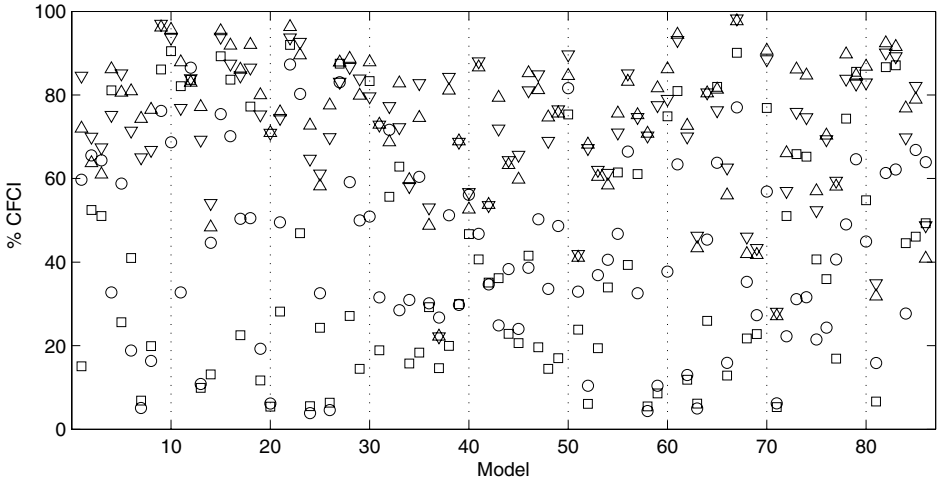

Fig. 9. Percentage correctly fuzzy classified instances for the A-model (□), N-model (▽), P-model (△) and C-model (○) for the 86 macroinvertebrate taxa.

5. Optimization of the linguistic terms

5.1. Introduction

In this section we want to improve the accuracy of the habitat suitability models for the region where the EKOO data set was collected, while maintaining the interpretability, i.e., the descriptive power of the models [5,6,17]. In the framework of this study, interpretability means that the river manager consulting the models is familiar with all components of the designed models and is able to get insight in the models just by looking at the different components. Given the uniformity of the qualitative information in the eight consulted knowledge sources, the rules in the rule bases of the developed models can be considered generally applicable to the Central and Western Plains of Europe. The knowledge sources also clearly reveal that the definition of linguistic values of environmental variables slightly differ from one river basin to another. Therefore the rule bases were kept unchanged, yet only the membership functions of the input variables were optimized in such a way that after optimization all fuzzy sets still represent the meaning assigned by experts to the corresponding linguistic values. As no straightforward relation exists between the membership functions and the output of a linguistic fuzzy model, a genetic algorithm [18–20] was used as optimization method as it works on the complete solution of the optimization problem, in this case being the whole set of membership function parameters.

5.2. Model selection

As mentioned in Section 3, the EKOO data set is characterized by ambiguous data as well as by a highly non-uniform distribution of the data over the four abundance classes *absent*, *low*, *moderate* and *high*.

The more different phenomena described by the model are included in a data set and the more uniform the distribution of the examples in the data set over the different phenomena is, the more appropriate the data set is for optimization. The input values of an ideal training data set are distributed uniformly over the different regions of the input space described by the antecedents of the rules. The different regions of the three-dimensional input space are hereby described by the 0.5-cuts of the membership functions of the linguistic values of the three input variables. Therefore, one could opt to select those taxa with the most uniformly distributed input values for model optimization. As the same input values and the same membership functions are used in respectively all A-, N-, P- and C-models, the distribution of the input values over the different regions of the input space is the same for all models of a given type. Therefore, in this case, the uniformity of the distribution of the data points over the input space is an inappropriate selection criterion.

Clearly, the distribution over the abundance classes needs to be taken into account to establish a decisive selection criterion. Therefore, the taxa whose data sets reveal the most uniform distribution over the crisp abundance classes, defined by the 0.5-cuts of the membership functions of the fuzzy abundance classes, were selected for optimization. As a measure for the uniformity of the distribution, entropy was used (convention $0 \cdot \log_2 0 = 0$):

$$\text{entropy} = -\frac{1}{\log_2 n} \sum_{i=1}^{n} p_i \cdot \log_2 p_i \tag{5}$$

with

$$p_i = \frac{1}{N} \sum_{j=1}^{N} A_{\text{crisp},i}(\mathbf{y}_{\text{data},j}).$$

The entropy is 1 for a uniform distribution and 0 if all data points are assigned to the same abundance class as is the case for *Odontomesa fulva*. For *Proasellus meridianus* and *Plectronemia conspersa*, of which the data point distributions are shown in Fig. 6, an entropy of respectively, 0.835 and 0.322 is obtained. In Table 4 entropy values for some other taxa are given. Note that entropy is a non-linear concept. When a distribution is highly non-uniform, as for *Agabus affinis*, the shift of 1 data point from the most frequent class to a less frequent class results in an entropy increase of at least 0.009. Given a more uniform

Table 4
Distributions of data points over four crisp classes and the corresponding entropy

Taxon name	Number of data points classified as				
	absent	*low*	*moderate*	*high*	Entropy
Odontomesa fulva	445	0	0	0	0.000
Agabus affinis	444	1	0	0	0.012
Elmis aenea	443	2	0	0	0.021
Proasellus meridianus	247	78	80	40	0.835
Erpobdella octoculata	237	106	64	38	0.841

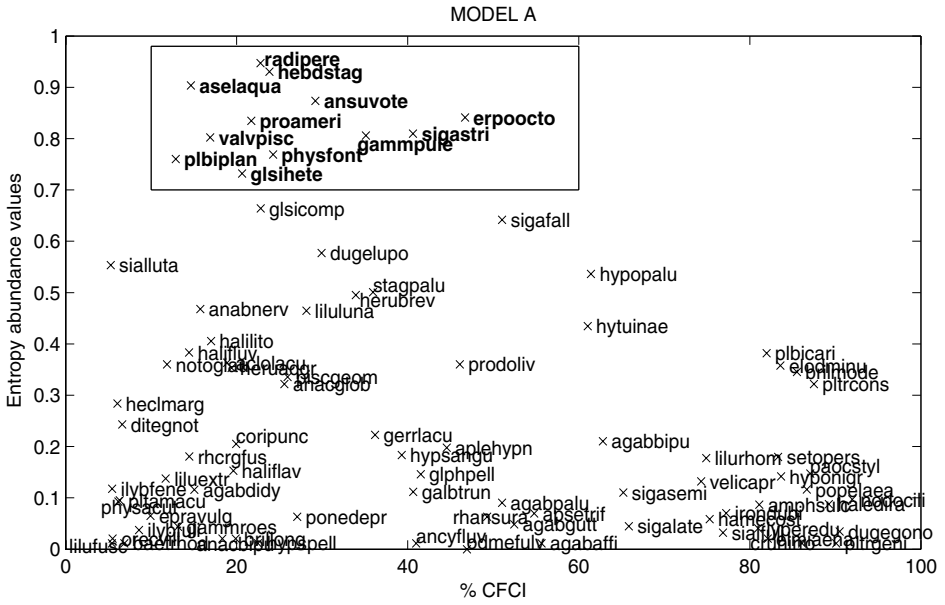

Fig. 10. Entropy and % CFCI of the 86 models including the ammonium concentration as input variable. The 12 models selected for optimization are indicated with a box.

initial distribution, a larger shift towards a more uniform distribution, gives a smaller entropy increase, for instance an entropy increase with 0.006 for *Erpobdella octoculata* compared to the entropy for *Proasellus meridianus*.

In Fig. 10 the entropy of the data distribution over the abundance classes for the 86 macroinvertebrate taxa is plotted as a function of the % CFCI of the A-model of the corresponding taxon. The figure gives an insight into the obtained values for the performance measures. One can see that a *good* performance according to the values of the performance measure often coincides with a low entropy. These *good* performing models are all models of macroinvertebrate taxa of which no individuals were collected at almost all 445 sampled sites and which are therefore not really evaluated by the data set. The 12 models selected for optimization have an entropy larger than 0.7 and are indicated with a box in Fig. 10. The threshold 0.7 was chosen arbitrarily: it separates 12, more or less clustered taxa from taxa with lower entropies. The selected taxa are: *Physa fontinalis*, *Anisus vortex*, *Asellus aquaticus*, *Erpobdella octoculata*, *Gammarus pulex*, *Glossiphonia heteroclita*, *Helobdella stagnalis*, *Planorbis planorbis*, *Proasellus meridianus*, *Radix peregra*, *Sigara striata* and *Valvate piscinalis*.

5.3. Properties of the genetic algorithm

The n_i membership functions of an input variable of the considered models are characterized by a vector of $2n_i$ reals, $\mathbf{a}_i = [a_{1,i}, a_{2,i}, \ldots, a_{2n_i,i}]$, satisfying the following two constraints:

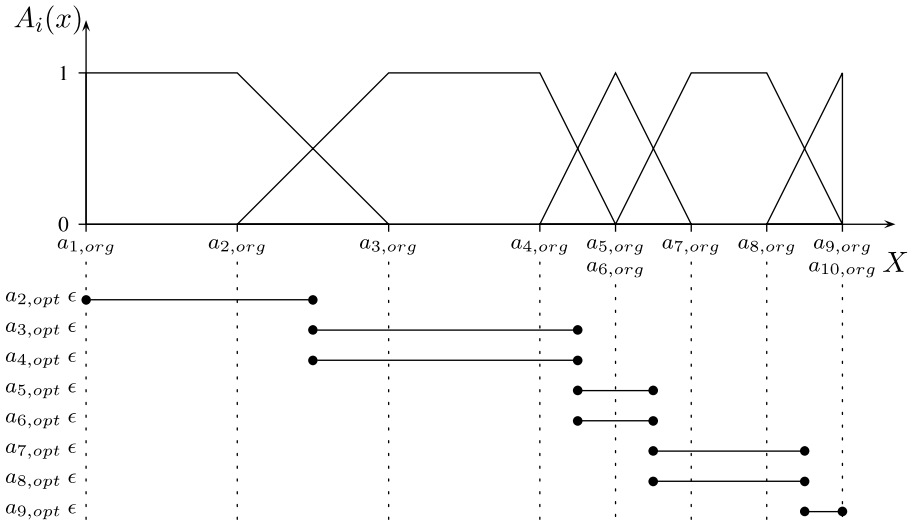

Fig. 11. Illustration of the optimization intervals used for the membership function parameters during the bounded simulation.

$$\forall j \in \mathbb{N}, \quad 1 \leqslant j \leqslant n_i : a_{2j-1,i} \leqslant a_{2j,i}, \tag{6}$$

$$\forall j \in \mathbb{N}, \quad 1 \leqslant j < n_i : a_{2j,i} < a_{2j+1,i}. \tag{7}$$

In this study both a binary-coded as well as a real-coded genetic algorithm are applied. The representation of the membership function parameters by a binary vector (using Gray encoding), restricts the values the parameters can take to a limited set of values defined by the upper and lower bound of the optimization interval and the length of the binary string, but has the advantage that it allows the use of very straightforward crossover and mutation strategies. The real-coded genetic algorithm is directly applied to a vector containing the real values of the optimized parameters, which allows for a finer tuning of the parameters. Two optimizations were carried out: a bounded and a free optimization. During the bounded optimization the kernels of the optimized membership functions are always subsets of the 0.5-cuts of the corresponding original membership functions (as illustrated in Fig. 11), whereas during the free optimization only the number of membership functions of the fuzzy partition is fixed for each input variable. The free optimization was carried out to investigate how the optimization process evolves if no constraints are set. The membership function parameters were coded as binary strings of 7 and 10 bits per parameter, respectively for the bounded and free optimization, respectively.

The structure of the genetic algorithm used to optimize the parameters of the trape-zial membership functions of the input variables of the A-, N-, P- and C-models of the 12 selected taxa is shown in Algorithm 1. A thorough investigation of the influence on the genetic algorithm performance of different mutation, crossover and selection procedures and the optimization of their parameters was outside the scope of this study. We carried out some fragmentary investigation of the parameter settings of the selected mutation an crossover procedures with some of the 48 models and applied the best setting obtained to optimize the membership functions of all 48 models.

Algorithm 1. Genetic algorithm

```
t ← 0
Initialize Population P_t at random
foreach Individual of P_t do
    Decode chromosome
    if chromosome represents unfeasible solution then
        Try to restore chromosome
    end
    If chromosome represents feasible solution then
        Calculate fitness of the individual
    else
        Assign very bad fitness value to the individual
    end
end
while stop criterion not reached do
    Select individuals by tournament selection
    Recombine individuals by crossover and mutation
    foreach Child of P_t do
        Decode chromosome
        if chromosome represents unfeasible solution then
            Try to restore chromosome
        end
        if chromosome represents feasible solution then
            Calculate fitness of the individual
        else
            Assign very bad fitness value to the individual
        end
    end
    Replace worst individual of P_{t+1} by best individual of P_t
    P_t ← P_{t+1}
    t ← t + 1
end
```

The same procedure was followed by the binary-coded and real-coded algorithm, except for the recombination and mutation. Each optimization starts with a population of 100 randomly generated strings, which, in case they do not represent a feasible solution, are tried to be restored by replacing them by (the binary representation of) a vector consisting of substrings of sorted real values of the unfeasible string for each variable. Note that this restoration procedure does not always result in a string satisfying Eq. (7).

During the search, each model was evaluated on each of the 445 data points, using a weighted average deviation (wAD) in which the weights guarantee that each region of

the input space defined by the 0.5-cuts of the membership functions of the non-optimized models has the same contribution to the fitness:

$$\mathrm{wAD} = \sum_{j=1}^{N} w_j \cdot \sum_{i=1}^{n-1} \left| \sum_{k=1}^{i} A_k(\mathbf{y}_{\mathrm{data},j}) - \sum_{k=1}^{i} A_k(\mathbf{y}_{\mathrm{model},j}) \right| \tag{8}$$

with

$$w_j = \frac{1}{N_j \cdot n_{\mathrm{regions}}}.$$

In the definition of the weights w_j, N_j is the number of data points in the same region of the input space as the jth input of the data set and $n_{regions}$ is the number of regions in which the input space is divided.

At each generation step, 100 parents were selected by tournament selection. Two by two the parents were recombined and mutated, resulting in two children. In the binary-coded algorithm, uniform crossover is applied (crossover probability = 0.95). Each bit of the strings obtained after recombination, or, in case no crossover was carried out, the strings of the parents, were changed with a mutation probability being the reverse of the length of the binary string. In the real-coded algorithm, one child is created with heuristic crossover and one with arithmetical crossover (crossover probability = 0.95). The procedure of the heuristic crossover described in [20] was slightly adapted to guarantee that each real value $a_{\mathrm{child}_1,l}$ in the string of the child derived from the corresponding values $a_{\mathrm{parent}_b,l}$ and $a_{\mathrm{parent}_w,l}$ of the best and, respectively, the worst performing parent of the two parents, is an element of the optimization interval $[b_l,B_l]$. In Eq. (10), r_1 is a random number between 0 and 1 and identical for all values of a string during a recombination:

$$a_{\mathrm{interval},l} = \max(b_l, \min(B_l, 2a_{\mathrm{parent}_b,l} - a_{\mathrm{parent}_w,l})), \tag{9}$$

$$a_{\mathrm{child}_1,l} = \min(a_{\mathrm{parent}_b,l}, a_{\mathrm{interval},l}) + r_1(\max(a_{\mathrm{parent}_b,l}, a_{\mathrm{interval},l}) - \min(a_{\mathrm{parent}_b,l}, a_{\mathrm{interval},l})), \tag{10}$$

$$a_{\mathrm{child}_2,l} = \frac{1}{2}(a_{\mathrm{parent}_b,l} + a_{\mathrm{parent}_w,l}). \tag{11}$$

The real strings of the children, or, in case no recombination was carried out, the strings of the parents, were mutated as described in Eq. (12). Each value a_l is replaced by a randomly selected (uniform probability distribution) value a'_l from an interval around a_l being at most as large as $p_{\mathrm{mut}}\%$ of the interval $[b_l,B_l]$ ($p_{\mathrm{mut}} = 3$ and $p_{\mathrm{mut}} = 0.4$ for the bounded and, respectively, the free optimization). In Eq. (12), r_2 is a random number between 0 and 1 and r_3 a random binary digit, both being identical for all values of a string during a recombination:

$$a'_l = \begin{cases} \min\left(a_l + \frac{1}{2} r_2 p_{\mathrm{mut}}(B_l - b_l), B_l\right) & \text{if } r_3 \text{ is } 0, \\ \max\left(a_l - \frac{1}{2} r_2 p_{\mathrm{mut}}(B_l - b_l), b_l\right) & \text{if } r_3 \text{ is } 1. \end{cases} \tag{12}$$

Children not satisfying Eqs. (6), (7) are tried to be restored, following the same procedure as during the initialization of the population. Furthermore, elitism is applied in the algorithm. The genetic algorithm was stopped if only small improvements of the fitness of the best individual ($\triangle$ fitness <0.001) were obtained during the last 50 consecutive generations or if the 1000th generation was reached. Hundred repetitions were carried out for each

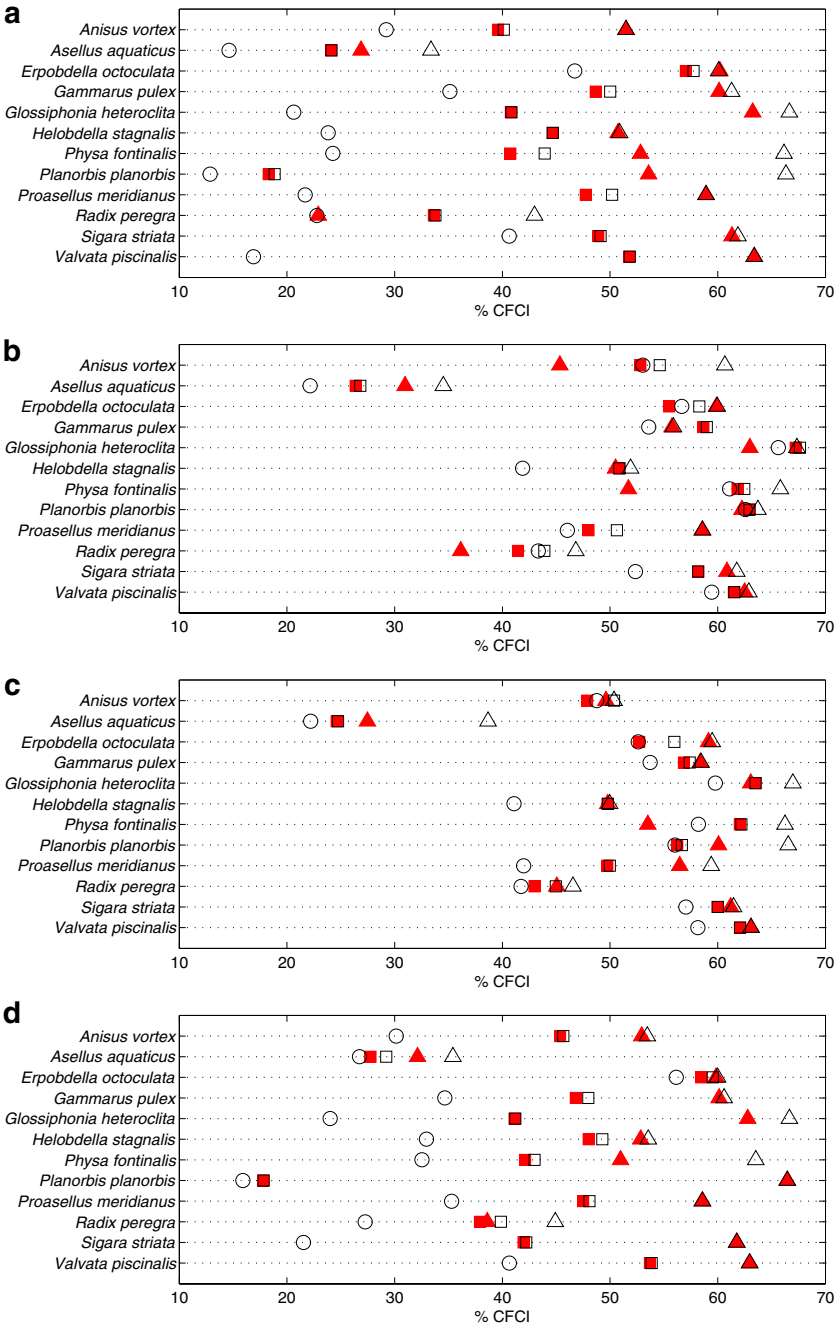

Fig. 12. Percentage of correctly fuzzy classified instances for the original models (○) and the models obtained through bounded optimization with the binary-coded GA (■), free optimization with the binary-coded GA (▲), bounded optimization with the real-coded GA (□) and free optimization with the real-coded GA (△) for the 12 selected taxa: (a) A-models, (b) N-models, (c) P-models and (d) C-models.

optimization and the model with the highest % CFCI among the 100 candidate models was retained as result of the optimization.

5.4. Optimization results

The results obtained for the four models of the 12 selected taxa are summarized in Fig. 12. One expects the models obtained with the real-coded genetic algorithm to perform at least as good as the corresponding models obtained with the binary-coded genetic algorithm as the search space of the binary-coded genetic algorithm is a subset of the search space of the real-coded genetic algorithm. Furthermore, the model obtained through free optimization is expected to outperform the corresponding model obtained through bounded optimization, which on its turn is expected to score better than the original model. Strictly speaking, the performance of the genetic algorithm can only be compared based on the performance of the original and optimized models according to the performance measure wAD, used as fitness function. In Fig. 12, however, the % CFCI of the original and optimized models are given, as % CFCI can be understood intuitively and resembles the performance measure % CCI commonly used in ecology. When analyzing the results in Fig. 12, one should always keep in mind the variability of the relationship, illustrated in Fig. 8(b), between the two performance measures non-weighted AD and % CFCI.

The models obtained with the real-coded GAs do not perform worse than those obtained with the binary-coded GAs, except for the A-model for *Erpobdella octoculata* obtained through free optimization. For this model, the optimized model obtained with the real-coded genetic algorithm shows a negligible worse performance of 0.1% compared to the model obtained with the binary-coded genetic algorithm (Fig. 12(a)). When considering the wAD as performance measure, 8 of the 96 real-coded GAs do not return a better solution than their binary-coded counterpart, which indicates that the implemented control structures were maladjusted to these eight membership function optimization problems.

For the models obtained with the binary-coded genetic algorithm, the expected order of the % CFCI-values of, respectively, the original model and the models obtained through bounded and free optimization, is not respected by the results recorded for the A-model of *Radix peregra*, the N-models of *Anisus vortex*, *Erpobdella octoculata*, *Gammarus pulex*, *Glossiphonia heteroclita*, *Helobdella stagnalis*, *Physa fontinalis*, *Planorbis planorbis* and *Radix peregra*, nor for the P-models of *Anisus vortex*, *Glossiphonia heteroclita* and *Physa fontinalis*. When applying the real-coded genetic algorithm only the % CFCI-values of the original, bounded and freely optimized N-models of *Gammarus pulex* and *Glossiphonia heteroclita* do not respect the expected order. When considering the wAD, all optimized models perform better than the corresponding original models and the expected performance order was recorded for all optimizations, except for the N-models of *Anisus vortex*, *Asellus aquaticus*, *Physa fontinalis* and *Radix peregra*. For these four models, a smaller wAD is obtained for the models returned by bounded optimization with the binary-coded GA than for the models obtained by free optimization with the binary-coded GA. The reversed order of the performances might be caused by the binary coding, restricting the values taken by the membership function parameters in the optimized models to a limited set of values. Thus, when using binary encoding the search space of the binary-coded genetic algorithm applied during the free optimization might simply not contain a solution outperforming the solution returned by the bounded optimization. The fact that all wAD-values obtained by the real-coded GAs respect the expected order, supports the above argument.

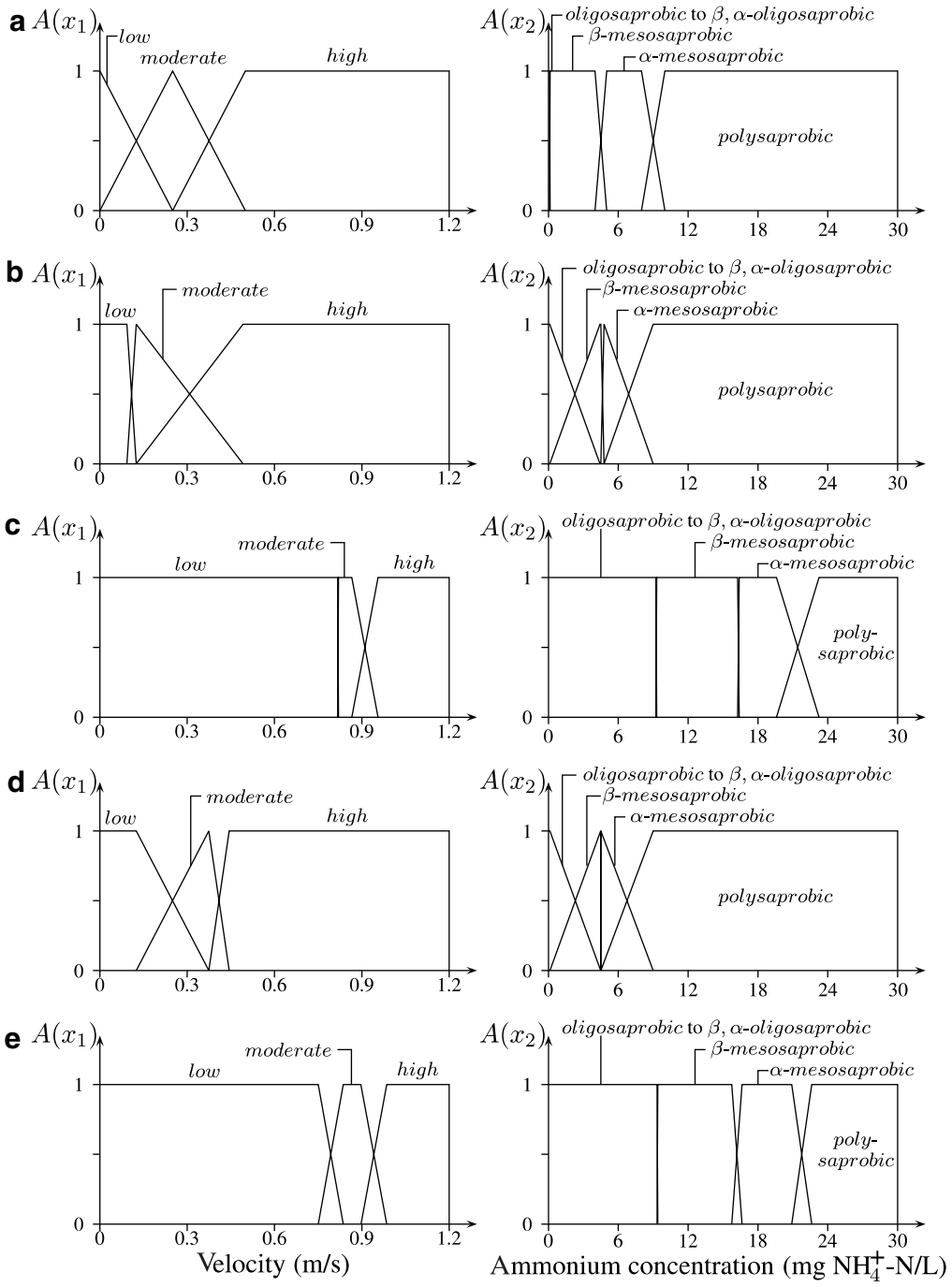

Fig. 13. Membership functions of the A-model of *Proasellus meridianus*: (a) original model and models obtained through (b) bounded optimization with the binary-coded GA, (c) free optimization with the binary-coded GA, (d) bounded optimization with the real-coded GA and (e) free optimization with the real-coded GA.

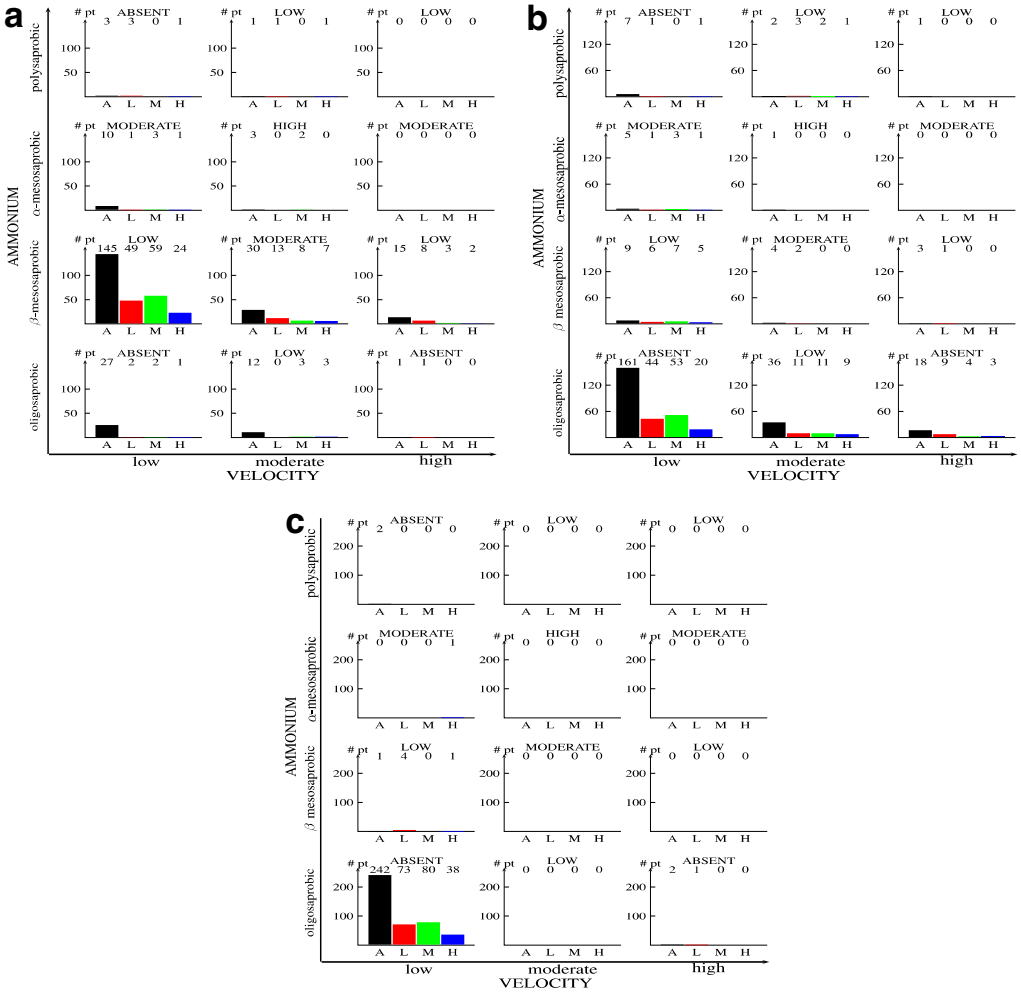

Fig. 14. Distribution of the data points over the abundance classes in the different regions of the input space defined by 0.5-cuts of the membership functions of (a) the original model, (b) the model obtained through bounded optimization with the binary-coded GA and (c) free optimization with the binary-coded GA of the A-model of *Proasellus meridianus*.

In Figs. 13 and 14 the results obtained for the A-model of *Proasellus meridianus* are shown. Note that the membership function describing the *oligosaprobic* to β,α-*oligosaprobic* conditions (hereafter called *oligosaprobic*) in the original model has such a small support that it can hardly be noticed in Fig. 13(a). For the A-model of *Proasellus meridianus*, as for most models of the other selected taxa, the results obtained with the real-coded genetic algorithm are very similar to the results obtained with the binary-coded genetic algorithm. This is especially true in case of the bounded optimization where the membership function parameters of the optimized models obtained with both algorithms are often equal to the lower or upper bound, or the second or next-to-last value of the corresponding optimization interval.

In Fig. 13 one sees that the membership functions of the velocity value *low* and the *oligosaprobic* conditions are extended towards higher velocities and ammonium concentrations, respectively. The membership functions in Figs. 13(c) and (e) no longer reflect the meaning given by the experts to the linguistic values. During the bounded optimization the extension is however limited by the constraints described in Section 5.3, which guarantees the interpretability of the fuzzy partitions of the optimized models. In Fig. 14 the number of data points belonging to the four defuzzified abundance classes $A_{\text{crisp},i}$ (see Eq. (2) for the defuzzification procedure) in the different regions of the input space are given and visualized by means of histograms for the original models and the two models obtained with the binary-coded genetic algorithm. No histograms are shown for the models obtained with the real-coded genetic algorithm, as similar membership functions were obtained with the binary-coded and real-coded genetic algorithm. One sees that, by extension of the support of the velocity value *low* and the *oligosaprobic* conditions, more data points and in particular more data points belonging to the abundance class *absent*, fire the rule

IF *vel* **IS** *low* **AND** *ammon* **IS** *oligotrophic* **THEN** *abundance* **IS** *absent*,

instead of the rules

IF *vel* **IS** *low* **AND** *ammon* **IS** β-*mesotrophic* **THEN** *abundance* **IS** *low*,
IF *vel* **IS** *moderate* **AND** *ammon* **IS** *oligotrophic* **THEN** *abundance* **IS** *low*,
IF *vel* **IS** *moderate* **AND** *ammon* **IS** β-*mesotrophic* **THEN** *abundance* **IS** *moderate*,

which results in a better score for the used fitness wAD as well as for the other performance measures % CCI, % CFCI and AD.

The differences between the results obtained with the bounded and free optimizations illustrate that one should not only focus on the accuracy of a model when evaluating its performance, but that the global performance of a model implies a balance between its interpretability and its accuracy. In the framework of this study, interpretability means that the river manager consulting the models is familiar with all components of the designed models and is able to get insight in the models just by looking at the different components. In order to guarantee interpretability, the definition of the linguistic values, i.e., the membership functions, should correspond to those used in the domain of biological water quality assessment. Therefore, the models obtained with bounded optimization are considered to have a better performance than those obtained with free optimization, even if higher accuracies are obtained for the latter.

6. Conclusions

In this study fuzzy ordered classifiers were used to classify river sites according to their suitability as a habitat for macroinvertebrates. The classifiers were evaluated using data collected in the Province of Overijssel in the Netherlands. Two performance measures were introduced in this paper: the percentage of correctly fuzzy classified instances, % CFCI, for fuzzy (ordered) classification, and the average deviation, AD, for fuzzy ordered classification.

Furthermore, one type of interpretability-preserving data-driven optimization, as well as an accuracy-oriented optimization, were applied using both a binary-coded and a real-coded genetic algorithm. For four models the binary-coded genetic algorithms

returned less accurate solutions for the accuracy-oriented optimization than for the constrained optimization, due to the fact that the optimized membership function parameters only take values from a limited set of values. A shortcoming which, as shown by the experiments, can be remedied by applying real encoding instead of binary encoding. The real-coded GAs applied in this study, however, showed maladjusted to eight of the 96 addressed membership function optimization problems, as an exhaustive investigation of the control structures of the genetic algorithms was outside the scope of this study.

A purely accuracy-oriented optimization is no option when one wants to preserve the interpretability of the habitat suitability models under study with the EKOO data set. In this case, expert knowledge was a prerequisite to build interpretable models in order to define the rule bases and determine the optimization intervals of the membership function parameters. The accuracy-oriented optimization, however, gives a better insight in the driving force during the bounded optimization, i.e., the tendency to classify as much data points as possible in the abundance class *absent* by increasing the regions were the input is mapped to *absent*, and stresses the importance of uniformly distributed and unambiguous training data for model optimization.

Acknowledgements

The first and second author, respectively, received funding from BOF-UGent, the Special Research Fund of Ghent University, under contract no. B/03843 and from IWT-Vlaanderen, the Institute for the Promotion of Innovation by Science and Technology in Flanders. The authors would like to express their gratitude to Piet Verdonschot of Alterra Green World Research, Wageningen, the Netherlands, for providing them with the EKOO data set.

Appendix A. List of macroinvertebrate taxa

In Table A.1 all 86 macroinvertebrate taxa considered in this study are listed. In the first column the index is given as used in this manuscript, followed by the full taxon name and the abbreviation used in this study in the second and third column. The 12 taxa selected for optimization of the membership functions are indicated in bold.

Table A.1
Macroinvertebrate taxa

	Taxon name	Taxon code
1	*Agabus didymus*	agabdidy
2	*Agabus guttatus*	agabgutt
3	*Agabus paludosus*	agabpalu
4	*Amphinemura sulcicolis*	amphsulc
5	*Anacaena globulus*	anacglob
6	*Ancyclus fluviatilus*	ancyfluv
7	*Baetis rhodani*	baetrhod
8	*Brillia longifurca*	brillong
9	*Crunoecia irrorata*	crunirro

(*continued on next page*)

Table A.1 (*continued*)

	Taxon name	Taxon code
10	*Dugesia gonocephala*	dugegono
11	*Elmis aenea*	elmiaena
12	*Elodes minuta*	elodminu
13	*Ephemera vulgata*	epravulg
14	*Gammarus roesellii*	gammroes
15	*Halesus radiatus*	haledira
16	*Hydroporus nigrita*	hyponigr
17	*Hydropsyche pellucidula*	hypspell
18	*Ironoquia dubia*	irondubi
19	*Limnephilus extricates*	liluextr
20	*Limnephilus fuscifornis*	lilufusc
21	*Limnephilus lunatus*	lililuna
22	*Notidobia ciliaris*	nodocili
23	*Odontomesa fulva*	odmefulv
24	*Orectochillus villosus*	orecvill
25	***Physa fontinalis***	physfont
26	*Platambus maculatus*	pltamacu
27	*Plectrocnemia conspersa*	pltrcons
28	*Nebrioporus depressus*	ponedepr
29	*Rheocricotopus group fuscipes*	rhcrgfus
30	*Sericostoma personatum*	setopers
31	*Acroloxus lacustris*	aclolacu
32	*Agabus affinis*	agabaffi
33	*Agabus bipustulatus*	agabbipu
34	*Anabolia nervosa*	anabnerv
35	*Anacaena bipustulatus*	anacbipu
36	***Anisus vortex***	ansuvote
37	***Asellus aquaticus***	aselaqua
38	*Corixa punctata*	coripunc
39	*Dugesia lugubris/polychroa*	dugelupo
40	***Erpobdella octoculata***	erpoocto
41	*Galba trunculata*	galbtrun
42	***Gammarus pulex***	gammpule
43	*Gerris lacustris*	gerrlacu
44	*Glossiphonia complanata*	glsicomp
45	***Glossiphonia heteroclita***	glsihete
46	*Glyphotaelius pellucidus*	glphpell
47	*Haliplus flavicollis*	haliflav
48	*Haliplus fluviatilis*	halifluv
49	*Haliplus lineatocollis*	halilito
50	*Haementaria costata*	hamecost
51	***Helobdella stagnalis***	hebdstag
52	*Hemiclepsis marginata*	heclmarg
53	*Helophorus aquaticus/grandis*	heruaqgr
54	*Helophorus brevipalpis*	herubrev
55	*Hydroporus palustris*	hypopalu
56	*Hydropsyche angustipennis*	hypsangu
57	*Hygrotus inaequalis*	hytuinae
58	*Ilybius fenestratus*	ilybfene
59	*Ilybius fuliginosus*	ilybfuli
60	*Limnephilus rhombicus*	lilurhom
61	*Lype reducta*	lyperedu
62	*Notonecta glauca*	notoglau

Table A.1 (*continued*)

	Taxon name	Taxon code
63	*Physa acuta*	physacut
64	*Piscicola geometra*	piscgeom
65	*Planorbis carinatus*	plbicari
66	***Planorbis planorbis***	plbiplan
67	*Plectrocnemia geniculata*	pltrgeni
68	***Proasellus meridianus***	proameri
69	***Radix peregra***	radipere
70	*Sialis fuliginosa*	sialfuli
71	*Sialis lutaria*	sialluta
72	*Sigara falleni*	sigafall
73	*Sigara lateralis*	sigalate
74	*Sigara semistriata*	sigasemi
75	***Sigara striata***	sigastri
76	*Stagnicola palustris*	stagpalu
77	***Valvata piscinalis***	valvpisc
78	*Velia caprai*	velicapr
79	*Brillia modesta*	brilmode
80	*Aspectrotanypus trifascipennis*	apsetrif
81	*Dicrotendipes group notatus*	ditegnot
82	*Polypedilum laetum agg.*	popelaea
83	*Parametriocnemus stylatus*	paocstyl
84	*Aplexa hypnorum*	aplehypn
85	*Prodiamesa olivacea*	prodoliv
86	*Rhantus suturalis*	rhansura

References

[1] EU, Directive of the European Parliament and of the Council 2000/60/EC establishing a framework for community action in the field of water policy, European Union, The European Parliament, The Council, PE-CONS 3639/1/00 REV 1 EN, 2000, 62 p. + annexes.

[2] A. Guisan, N.E. Zimmerman, Predictive habitat distribution models in ecology, Ecological Modelling 135 (2000) 147–168.

[3] R.H. Rosenberg, V.H. Resh (Eds.), Freshwater Biomonitoring and Benthic Macroinvertebrates, Chapman and Hall, New York, NY, USA, 1993.

[4] N. De Pauw, G. Vanhooren, Method for biological quality assessment of watercourses in Belgium, Hydrobiologia 100 (1983) 153–168.

[5] J. Casillas, O. Cordón, F. Herrera, L. Magdalena (Eds.), Interpretability Issues in Fuzzy Modeling, Studies in Fuzziness and Soft Computing, vol. 128, Springer Verlag, Heidelberg, 2003.

[6] J. Casillas, O. Cordón, F. Herrera, L. Magdalena (Eds.), Accuracy Improvements in Fuzzy Modeling, Studies in Fuzziness and Soft Computing, vol. 129, Springer Verlag, Heidelberg, 2003.

[7] P.F.M. Verdonschot, Ecological characterization of surface waters in the Province of Overijssel (the Netherlands), Ph.D. dissertation, Landbouwuniversiteit Wageningen, Wageningen, The Netherlands, 1990.

[8] J. Illies, Limnofauna Europaea, second ed., Fischer, Stuttgart, 1978.

[9] V. Adriaenssens, E. Van Broekhoven, P.F.M. Verdonschot, B. De Baets, N. De Pauw, Knowledge and rule base development for macroinvertebrate habitat suitability modelling in river management, submitted for publication.

[10] J.R. Karr, K.D. Fausch, P.L. Angermeier, P.R. Yant, L.J. Schlosser, Assessing biological integrity in running waters: a method and its rationale, Special publication 5, Illinois Natural History Survey, Champaign, Illinois, USA, 1986.

[11] E. Ruspini, A new approach to clustering, Information and Control 15 (1969) 22–32.

[12] B. Statzner, J. Gore, V. Resh, Hydraulic stream ecology: observed patterns and potential applications, Journal of the North American Benthological Society 7 (1988) 307–360.
[13] E. Van Broekhoven, V. Adriaenssens, B. De Baets, P.F.M. Verdonschot, Rule bases of habitat suitability models for macroinvertebrate taxa. Available from URL: <http://users.ugent.be/~bdebaets/habitatsuitabilitymodels/rulebases.pdf>.
[14] S. Manel, H.C. Williams, S.J. Ormerod, Evaluating presence–absence models in ecology: the need to account for prevalence, Journal of Applied Ecology 38 (2001) 921–931.
[15] U. Bodenhofer, E.P. Klement, Genetic optimization of fuzzy classification systems – a case study, in: B. Reusch, K.-H. Temme (Eds.), Computational Intelligence in Theory and Practice, Advance in Soft Computing, Physica Verlag, Heidelberg, 2001, pp. 183–200.
[16] E. Van Broekhoven, V. Adriaenssens, B. De Baets, P.F.M. Verdonschot, Fuzzy rule-based macroinvertebrate habitat suitability models for running waters, Ecological Modelling, in press, doi:10.1016/j.ecolmodel.2006.04.006.
[17] M. Drobics, U. Bodenhofer, E.P. Klement, FS-FOIL: an inductive learning method for extracting interpretable fuzzy descriptions, International Journal of Approximate Reasoning 32 (2003) 131–152.
[18] O. Cordón, F. Gomide, F. Herrera, F. Hoffmann, L. Magdalena, Ten years of genetic fuzzy systems: current framework and new trends, Fuzzy Sets and Systems 141 (2004) 5–31.
[19] D.E. Goldberg, Genetic Algorithms in Search Optimization and Machine Learning, Addison-Wesley Longman, 1989.
[20] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, third ed., Springer, Berlin, 1996.