Entropy-based fuzzy set optimisation for reducing ecological model complexity

A. M. Mouton
Bernard De Baets
A. Peter
G. Holzer
R. Müller

See next page for additional authors

Follow this and additional works at: https://scholarsarchive.byu.edu/iemssconference
Entropy-based fuzzy set optimisation for reducing ecological model complexity

A. M. MOUTON\textsuperscript{a}, B. De Baets\textsuperscript{b}, A. Peter\textsuperscript{c}, G. Holzer\textsuperscript{c}, R. Müller\textsuperscript{c} and P.L.M. Goethals\textsuperscript{a}

\textsuperscript{a} Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, J. Plateaustraat 22, B-9000 Ghent, Belgium

\textsuperscript{b} KERMIT: Research Unit Knowledge-based Systems, Ghent University, Coupure links 653, B-9000 Ghent, Belgium

\textsuperscript{c} Swiss Federal Institute for Environmental Science and Technology (EAWAG), Forschungszentrum für Limnologie, 6047 Kastanienbaum, Switzerland

Abstract: In recent years, fuzzy logic has been acknowledged as a suitable approach for species distribution modelling due to its transparency and its ability to incorporate the ecological gradient theory. Specifically, the overlapping class boundaries of a fuzzy model are similar to the transitions between different environmental conditions. However, the need for ecological expert knowledge is an important constraint when applying fuzzy species distribution models. Recent research has shown that data-driven fuzzy models may solve this ‘knowledge acquisition bottleneck’ and this paper is a further contribution. The aim was to reduce the complexity of a data-driven fuzzy habitat suitability model for European grayling (Thymallus thymallus) in the Aare River (Thun, Switzerland). Therefore, we applied an entropy-based fuzzy set selection algorithm, which allowed minimisation of the number of fuzzy sets needed for data-driven fuzzy model development. Comparison of the presented model with a previously developed model revealed that the entropy-based algorithm reduced model complexity substantially without a significant decrease in predictive accuracy. The results of this study could minimise monitoring costs and efforts, and enhance communication between water managers and stakeholders due to increased model transparency.

Keywords: Fuzzy logic, grayling, species distribution modelling, model complexity, data driven.

1. INTRODUCTION

Ecological data often is imprecise and characterized by high uncertainty. Therefore, instead of using crisp values for quantification of variables, many experts apply linguistic descriptions such as “low”, “high” or “moderate”. Fuzzy systems allow transforming these linguistic descriptions into a mathematical framework in which suitable data processing can be performed [Kampichler, et al., 2000]. The main advantages of fuzzy classifiers are their simplicity and linguistic interpretability, which are important factors for the usability and acceptance of a model [Adriaenssens, et al., 2004, Van Broekhoven, et al., 2006]. This turns fuzzy systems into a popular technique for ecological modelling, resulting in numerous applications. Yet, a purely knowledge-driven approach, aiming at formalizing problem-relevant human expert knowledge, is difficult and tedious. Recent research has shown that complementing fuzzy systems by data-driven techniques can solve this “knowledge acquisition bottleneck” [Žnidaršic, et al., 2006]. For example, the induction of fuzzy rule-based models by heuristic search algorithms is often used in the field of fuzzy rule learning [Hüllermeier, 2005].
To preserve the interpretability of the fuzzy model, however, the number of parameters applied in the model should be limited. Therefore, the aim of this paper was to reduce the complexity of a data-driven fuzzy habitat suitability model for European grayling (Thymallus thymallus) in the Aare River (Thun, Switzerland). We applied and evaluated an entropy-based fuzzy set selection algorithm, which allowed minimisation of the number of fuzzy sets needed for data-driven fuzzy model development. Specifically, the hypothesis was tested that the entropy-based algorithm reduced model complexity substantially without a significant decrease in predictive accuracy. The results of this study could minimise monitoring costs and efforts, and enhance communication between water managers and stakeholders due to increased model transparency.

2. MATERIAL AND METHODS

2.1 Study area

The studied site is a 1300 m stretch of the Aare river in the Bern department, Switzerland, and is situated along the city of Thun. Up to this point, the Aare river is draining an area of about 2490 km² and is classified as a 7th order stream [Strahler, 1957]. The average flow is 111 m³.s⁻¹, with respective base and peak flows of 23 and 570 m³.s⁻¹. The Aare river at the studied site was originally a braided river with large gravel banks. However, since the beginning of the 18th century anthropogenic disturbances were introduced for flood control and hydropower generation [EAWAG, 2002]. Hence, the flow regime is altered and controlled by flood control weirs. Nevertheless, the studied site contains some of the major spawning habitats for European grayling in Switzerland.

To allow for the development of the habitat suitability model, an intensive monitoring campaign was set up. In the studied stretch, 50 cross-sections were defined and water depth was measured along each cross-section at equal distances of about 1 m using a Raytheon 760 depth measuring device (Raytheon, MA, USA). Flow velocity was measured with a Flo-Mate 2000 flow meter (Marsh-McBirney Inc., MD, USA) at 40 % of the water column height in 14 of these cross-sections at equal distances of about 25 m, resulting in 63 measurements. The substrate composition was assessed by underwater photography and visual assessment with DIN A4 frames. Hence, the dominating substrate of the different patches in the studied stretch could be defined. If a patch was covered by macrophytes, both substrate percentages were set to 99 %. This substrate combination can not be observed in the river stretch and hence the definition of this specific situation will not affect optimisation results. All data were collected at a flow of ca. 100 m³.s⁻¹ and no significant flow changes were observed during the measurements.

A finite element grid of the studied stretch with 5625 elements and 22500 nodes was generated using SMS (surface water modelling system, Brigham Young University) software, while the size of the grid cells was adjusted depending on river geometry. Flow velocity and depth values were calculated at each node by a 2-dimensional hydraulic model which was generated using FESWMS (Finite Element Surface Water Modelling System, U.S. Geological Survey). Additional measurements of depth and flow velocity were performed in the whole stretch and more specific in the spawning areas to validate the hydraulic model. The hydraulic modelling was conducted by Schneider & Jorde Ecological Engineering in cooperation with the Swiss Federal Institute of Aquatic Science and Technology (EAWAG) and is described in detail in [EAWAG, 2002].

European grayling spawns in faster flowing patches (0.1 – 0.4 m.s⁻¹) with fine to medium-sized gravel substrate. During egg deposition, the trembling female grayling is pushing its abdomen in the gravel substrate, hereby creating small grooves [Fabricus and Gustafson, 1955]. These light-coloured grooves can easily be distinguished from the substrate which is mostly covered with dark brown algae. Hence, the spawning grounds of grayling were visually identified and localised using GPS (Garmin 12X). Each grid cell in the studied stretch was defined as suitable for spawning or not by combining the results of the hydraulic simulations with the observations of the spawning grounds. The resulting dataset
contained 22510 grid cells, 1 output variable indicating whether the habitat was suitable for spawning or not and 4 input variables characterising each grid cell (Table 1).

2.2. Fuzzy set parameter optimisation

The values assigned to the four input variables depth, flow velocity, percentage of fine gravel and percentage of medium-sized gravel, were defined by fuzzy sets [Zadeh, 1965] and not by conventional sets with crisp boundaries (hereafter called crisp sets). When using these crisp sets, for instance depths below 1.5 m would be considered 'low', depths between 1.5 and 3 m 'moderate' and depths higher than 3 m 'high'.

A given depth either belongs to a set (it has a membership degree of 1 to this set) or it does not. A fuzzy set is described by its membership function, indicating the membership degree for each variable value to this set. As the boundaries between these sets are overlapping, an element can partially belong to a fuzzy set and thus have a membership degree to this set ranging from zero to one. Hence, the linguistic statement ‘the depth is quite low but tending to be moderate’ can be translated into a depth which has a membership degree of 0.4 to the 'low' fuzzy set and of 0.6 to the 'moderate' set. In this study, all membership functions had trapezoidal shapes and were defined by four parameters ($a_m$, $b_m$, $c_m$ and $d_m$): the membership degree linearly increases between $d_m$ and $b_m$ from 0 to 1, is equal to 1 between $b_m$ and $c_m$ and linearly decreases from 1 to 0 between $c_m$ and $d_m$. A triangular membership function is obtained when $b_m$ equals $c_m$.

The parameters of the membership functions corresponding to the fuzzy sets of the input variables have often been derived from an expert knowledge. However, if a fuzzy set of an input variable contains very few training instances, rules which apply to this fuzzy set will be trained inadequately. Therefore, a uniform distribution of the input variables over the fuzzy sets was needed to generate reliable rule bases. The Shannon–Weaver entropy [Shannon and Weaver, 1963] indicated this uniformity and was applied to optimise the parameters of the membership functions of the input variables. The fuzzy sets were converted into crisp ones whose boundaries were the points having a membership degree of 0.5 to the corresponding fuzzy set. The entropy $E$ is given by (convention $0 \log_2 0 = 0$):

$$E = -\frac{1}{\log_2 n} \sum_{i=1}^{n} p_i \log_2 p_i$$

where $n$ is the number of classes and $p_i$ is the proportion of data points belonging to class $i$.

The algorithm for parameter optimisation starts with $n$ equal to 2, and then the parameters of the fuzzy sets are adjusted in steps of $0.5/tr$ with $r$ the range of the variable of which the fuzzy sets are optimised, and $s$ the fixed stepsize. For each variable, parameter optimisation started by creating two crisp sets with boundary at $r/n$ with $n$ the number of sets, which is equal to two. The crisp sets were transformed into fuzzy sets by setting the parameters of each set $m$, $a_m$, $b_m$, $c_m$ and $d_m$ (Fig. 1) as follows:

$$c_{m+i} = a_{m+i} = (m+1) \frac{r}{n} \cdot i \frac{r}{2s} ,$$

$$d_{m+i} = b_{m+i} = (m+1) \frac{r}{n} \cdot i \frac{r}{2s} ,$$

Fig. 1. The parameters of the fuzzy set $m$ of variable $X_i$.
with \( t \) indicating the iteration of the sets optimisation and thus \( t = 1 \) in this situation. Next, the entropy \( E_n \) of this fuzzy set configuration was calculated as described in Eq. 1 and this entropy was set to \( E_{\text{best}} \). This entropy \( E_{\text{best}} \) is compared with the entropy fixed threshold \( E_{\text{thres}} \) and the algorithm was terminated if \( E_{\text{best}} > E_{\text{thres}} \).

If \( E_{\text{best}} < E_{\text{thres}} \), the algorithm searched for the fuzzy set of which an expansion of the set boundaries could lead to the greatest increase in entropy. For each set \( m \), the upper boundary was expanded as follows:

\[
c_{m_{t+1}} = a_{m_{t+1}} = (m + 1) \cdot \frac{r}{n} - (t + 1) \cdot \frac{r}{2s},
\]

Eq. 4

\[
d_{m_{t+1}} = b_{m_{t+1}} = (m + 1) \cdot \frac{r}{n} + (t + 1) \cdot \frac{r}{2s}.
\]

Eq. 5

The entropy of this new fuzzy set configuration in iteration \( t+1 \), \( E_{m_{t+1}} \), was calculated and compared to \( E_{\text{best}} \). If \( E_{m_{t+1}} > E_{\text{best}} \), the algorithm continued with this new configuration and the upper boundary of the set \( m \) was further adjusted according to Eq. 4 and Eq. 5. If \( E_{m_{t+1}} < E_{\text{best}} \), the last adjustment of the fuzzy set was aborted and the algorithm continued with adjusting the boundaries of the next fuzzy set.

To avoid that the algorithm would generate erroneous fuzzy sets in which the total membership degree of a variable value would exceed one, the boundary adjustment of the fuzzy sets in each iteration \( t \) was limited as follows:

\[
(t + 1) \cdot \frac{r}{2s} \leq \min(\min(c_{m_{t+1}} - b_{m_{t+1}}, c_{m_{t+1}} - b_{m_{t+1}}), \min(c_{m_{t+1}} - b_{m_{t+1}}, c_{m_{t+1}} - b_{m_{t+1}}))
\]

Eq. 6

If the boundary adjustment of fuzzy set \( m \) reached this limit and the entropy was still lower than \( E_{\text{thres}} \), the set was split into two symmetric sets. This split was created in the same way as the first split of the total variable range into two fuzzy sets, but the variable range was now replaced by the range of the fuzzy set \( m \). The entropy of this new set configuration, \( E_{m_{t+1}} \), was calculated and compared to \( E_{\text{best}} \). If \( E_{m_{t+1}} < E_{\text{best}} \), the algorithm did not split the sets and continued with the next fuzzy set. If \( E_{m_{t+1}} > E_{\text{best}} \), the algorithm restarted from this new configuration. If an entropy which exceeded \( E_{\text{thres}} \) was obtained during sets optimisation, the boundaries of the remaining sets were still optimised, but without splitting the sets.

The applied fuzzy sets optimisation method ensures that the distribution of the training data instances over the fuzzy sets is optimal. In a situation with \( n \) fuzzy sets, sets which contain less than \( 100/n \% \) of the data will be expanded as far as possible, whereas sets which contain more than \( 100/n \% \) of the data will be reduced to or split into smaller sets. This method avoids that empty or poorly represented sets are included in the model and increases model efficiency by deleting redundant sets. However, a more uniform distribution of the input data over the fuzzy sets does not guarantee that each fuzzy rule is represented uniformly in the input data. Water depth values, for instance, can be uniformly distributed over a ‘low’ and a ‘high’ fuzzy set, but this does not imply that many sampling sites with a high flow velocity and a high depth are represented in the training data set.

In this paper, the impact of the fuzzy sets on the model training result was analysed by comparing two training scenarios. In the first scenario, models were trained based on fuzzy sets which were derived from expert knowledge and were described in an ecological study of spawning grayling in the Aare river [EAWAG, 2002]. In the second scenario, models were trained based on the fuzzy sets which were derived from the fuzzy set parameter optimisation approach described in this paper.

2.3. Fuzzy rule-based modelling and rule base optimisation

The fuzzy rule base combines the input variables into the habitat suitability for spawning grayling and consisted of if-then rules, such as ‘IF depth IS moderate AND flow velocity IS high AND percentage of fine gravel IS high AND percentage of medium-sized gravel IS moderate THEN habitat IS suitable’. The if-part of the rule, the antecedent, describes in which situation this rule applies, while the then-part, the consequent, indicates whether the
habitat in this situation is suitable or not for spawning grayling. Given crisp values of the four input variables, the output of the fuzzy model is calculated as described by Van Broekhoven et al. [2006]. For each instance, its membership degrees to the membership functions (Fig. 1) of each input variable is calculated. The degree of fulfilment of each rule is then calculated as the minimum of the membership degrees in its antecedent. Finally, to each linguistic output value a fulfilment degree is assigned equal to the maximum of the fulfilment degrees of all rules with the output value under consideration in their consequent. The approach is similar to the Mamdani-Assilian procedure [Assilian, 1974, Mamdani, 1974] in which the fuzzy output is defuzzified in a crisp one based on the fuzzy sets of the output variables. However, the output variable in this work was already defined by two crisp sets: present and absent. Therefore, a different type of model was applied: a fuzzy classifier. The model output was assigned to the fuzzy set with the highest fulfilment degree, which allowed comparison of the modelled output with the observed output and calculation of performance measures. If the output variable consists of two sets, this approach is very similar to the defuzzification procedures used in Mamdani-Assilian models.

To generate a reliable habitat suitability model, the consequents of the fuzzy rules were optimised using a nearest ascent hill-climbing algorithm. Starting from fixed fuzzy sets and a randomly selected rule base, the consequent of one rule is changed into its neighbouring fuzzy set and the impact on model performance is calculated. If model performance increases, the algorithm continues with the adjusted rule base, if not, it continues with the original one. Ten-fold cross-validation was applied to indicate the robustness of the optimisation results. The folds were constructed by randomising the original data set and assigning each data point to one fold without replacement. The species prevalence (i.e. the frequency of occurrence) was constant for all ten folds and equal to the prevalence of the original dataset (0.203 = 4579/22510).

Models were trained based on Cohen’s Kappa [Cohen, 1960], which is derived from the confusion matrix [Fielding and Bell, 1997] and ranges from -1 to 1. This measure was selected as a training performance criterion because model training based on Kappa showed good results in a previous study [Mouton, et al., accepted]. Each training iteration was stopped when no further increase of the performance measure on the test fold was observed. Each training iteration was repeated and the obtained rule base was compared to each rule base obtained in previous iteration steps. The resulting rule base similarity indicates the percentage of rule consequents that were different between two rule bases. If the rule base with the highest performance on the test fold was obtained 3 times, training continued on another fold. The training algorithm is described in detail in Mouton et al. (accepted).
3. RESULTS AND DISCUSSION

The results show that entropy-based fuzzy set optimisation leads to fewer fuzzy sets than fuzzy set construction based on expert knowledge (Fig. 2). Moreover, the entropy of the optimised fuzzy sets is higher than the entropy of the expert knowledge-based sets.

Consequently, entropy-based fuzzy set optimisation leads to sets over which the input data is distributed more uniformly (Table 1). Specifically, the entropy values show that the distribution of the input data over the expert-based fuzzy sets is less uniform than the distribution over the entropy-based sets. Due to this lack of uniformity, the likelihood increases that some rules of the model do not represent an environmental condition which is present in the studied data set or even in the natural environment (Mouton et al.,).

Such "ghost rules" increase model complexity without contributing to the model performance and should thus be avoided. Although there is no universal relation between the number of ghost rules and the uniformity of the input data distribution over the fuzzy sets, the results in this paper show that the fuzzy set optimisation approach leads to fewer ghost rules in the model (Table 2).

Table 2 shows that for the expert-knowledge based model, 50 % of the rules are ghost rules which are not represented in the dataset, whereas for the model with the optimised fuzzy sets, 75 % of the rules represents at least 0.5 % of the sampling points. River managers may thus get the false impression that the model with 108 fuzzy rules is more accurate than the model with 24 rules.

<table>
<thead>
<tr>
<th>Variables and their units</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow velocity (m.s⁻¹)</td>
<td>0.818</td>
</tr>
<tr>
<td>Depth (m)</td>
<td>0.808</td>
</tr>
<tr>
<td>Percentage of fine gravel (%)</td>
<td>0.822</td>
</tr>
<tr>
<td>Percentage of medium-sized gravel (%)</td>
<td>0.940</td>
</tr>
</tbody>
</table>

Table 1. The entropy of the different input variables of the fuzzy model. This entropy reflects the uniformity of the distribution of the input data over the fuzzy sets which were derived from expert knowledge (EK) and from entropy-based optimisation.

![Fig. 2. The fuzzy sets which were derived from expert knowledge (a) and from entropy-based fuzzy set optimisation (b).](image)
The fuzzy set optimisation approach also significantly reduces the model complexity. Specifically, the expert knowledge-based model contained 108 different fuzzy sets and thus the rule base of this model consisted of 108 different rules, whereas the rule base of the entropy-based optimised model contained 24 different rules.

Table 2. Distribution of the samples in the data set over the environmental situations considered in the fuzzy habitat suitability models. The model with the expert knowledge (EK)-based sets and the model with the optimised sets considered respectively 108 and 24 different environmental conditions.

<table>
<thead>
<tr>
<th>Percentage of samples represented by rule</th>
<th>Number of rules EK-based</th>
<th>%</th>
<th>Number of rules Optimised</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 %</td>
<td>54</td>
<td>50</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>0 - 0.5 %</td>
<td>22</td>
<td>20</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>0.5 - 5 %</td>
<td>26</td>
<td>24</td>
<td>12</td>
<td>50</td>
</tr>
<tr>
<td>&gt; 5 %</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>25</td>
</tr>
</tbody>
</table>

This model simplification significantly reduced the computational time of the consequent rule base optimisation process. However, a more important advantage of this approach is the improvement of model transparency and thus user friendliness. Although a fuzzy model with 108 rules may still be more transparent than some black box data-driven approaches such as Artificial Neural Networks, the high number of different environmental situations which are represented by the fuzzy rules may confuse river managers. Moreover, Table 3 illustrates that the significant reduction of model complexity does not result in a substantial decrease in model performance. Although the Kappa and TSS values decrease slightly, the percentage of correctly classified instances (CCI) is higher for the model with the optimised fuzzy sets.

Table 3. The performance of the fuzzy models obtained after rule base optimisation based on fuzzy sets which were derived from expert knowledge (EK) and from entropy-based fuzzy set optimisation. The model performance was calculated for the complete dataset and is quantified by the percentage of correctly classified instances (CCI), Cohen's Kappa and the true skill statistic (TSS).

<table>
<thead>
<tr>
<th>Performance criterion</th>
<th>EK-based</th>
<th>Optimised</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCI</td>
<td>0.71</td>
<td>0.79</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.34</td>
<td>0.32</td>
</tr>
<tr>
<td>TSS</td>
<td>0.33</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Not only do these results show that different performance criteria focus on different aspects of model performance, but they also confirm that reducing the model complexity by fuzzy set optimisation does not necessarily lead to a substantial reduction of model performance. Moreover, the predictions of both models agree for 82 % of the data points. Further analysis of the differences between the predictions of both models revealed that most of the prediction differences occurred at the edge of habitat patches. Specifically, in some cases one model predicted a smaller patch to be suitable for spawning grayling than the other model. This difference in model predictions was caused by the different fuzzy sets and consequently the complex model could predict the suitability of some habitat patches more accurately than the simplified model could. However, analysis of the model performance showed that in general, the predictions of the simplified model were even more accurate than those of the complex model.

Finally, sampling costs and efforts increase substantially if a higher number of different environmental conditions is represented in the model because several authors suggest that the distribution of the input data over the different environmental conditions should be as uniform as possible [Guisan and Zimmermann, 2000, Hirzel and Guisan, 2002].

4. CONCLUSION

In this paper, an entropy-based fuzzy set optimisation algorithm was applied in order to reduce the complexity of a fuzzy rule-based species distribution model. The results showed that model complexity could be reduced significantly without a substantial loss in model
performance. Although more complex models may be appropriate for specific situations, application of more simplified model may reduce sampling costs and efforts. The reduction in complexity also enhances the transparency of the model and may thus be a valuable contribution to integrated river management by improving communication between modellers, river managers and different stakeholders.

ACKNOWLEDGEMENTS

This research was sponsored by the Special Research Fund (BOF) of Ghent University. The habitat demand data for the presented case study were developed by Schneider & Jorde Ecological Engineering and the Swiss Federal Institute for Environmental Science and Technology (EAWAG).

REFERENCES


EAWAG, Fishereiliches Gutachten über die Aarebaggerung in Thun, Forschungsanstalt des ETH-Bereichs, EAWAG (Fish Federal Institute for Environmental Science and Technology), Kastanienbaum, Switzerland, 2002.


