A variable length chromosome genetic algorithm approach to identify species distribution models useful for freshwater ecosystem management.

Sacha Gobeyn, and Peter L.M. Goethals

Ghent University, Laboratory of Environmental Toxicology and Aquatic Ecology, Coupure Links 653, B-9000 Ghent, Belgium
Sacha.Gobeyn@ugent.be

Abstract. Increasing pressure on freshwater ecosystems requires river managers and policy makers to take actions to protect ecosystem health. Species distribution models (SDMs) are identified as appropriate tools to assess the effect of pressures on ecosystems. A number of methods are available to model species distributions, however, it remains a challenge to identify well-performing models from a large set of candidate models. Metaheuristic search algorithms can aid to identify appropriate models by scanning possible combinations of explanatory model variables, model parameters and interaction functions. This large search space can be efficiently scanned with simple genetic algorithms (SGAs). In this paper, we test the potential of a variable length chromosome SGA to perform parameter estimation (PE) and input variable selection (IVS) for a macroinvertebrate SDM. We show that the SGA is an appropriate tool to identify fair to satisfying performing SDMs. In addition, we show that SGA performance and the uncertainty varies as a function of the chosen hyper parameters. The results can aid to further optimise the algorithm so models explaining species distributions can be identified and used for analysis in river management.

Keywords: species distribution models; model identification; genetic algorithms; freshwater management; macroinvertebrate species; input variable selection; parameter estimation

1 Introduction

Freshwater ecologist and river managers are in need for system analysis techniques to investigate a wide range of ecological questions and support decision making. Species distribution models (SDMs) aiming to describe the species response to driving processes, have shown to be valuable tools in ecosystem health management. Many approaches to identify SDMs are available, however, the challenge remains to test a large set of candidate explanatory models.

Genetic algorithms (GAs) classified under evolutionary algorithms and inspired by various mechanisms observed in evolution (i.e. reproduction, mutation, selection) are promising approaches to evaluate a large search space [12,15,18]. Consequently, GAs are used to select input variables (input variable selection, IVS) for SDMs by using them as a wrapper for data-driven approach [3]. They are also used to estimate parameter values (parameter estimation, PE) for fuzzy logic SDMs [5,25]. PE and IVS are important aspects in SDM identification and it can hypothesized whether a joint approach can be encoded in GAs.

In this paper, we present the use of a simple genetic algorithm (SGA) for PE and IVS for an SDM. To do so, we encode the optimisation problem in a variable length chromosome. The approach is tested for a freshwater species, *cloeon dipterum*, with the Limnodata of the Netherlands. The acquired SDM performance, parameters and input variables are analysed. In addition, a sensitivity analysis is done to test the effect of the algorithm hyper parameters on the SGA performance (Section 3). The results of this approach are discussed in section 4.

2 SDM development

The SDM is developed by following a four step approach (Figure 1). First, a number of ecological concepts are used to define the model. Second, the data are gathered and processed to construct the model (step 3). In a final step, a search algorithms is implemented and used to identify well-performing models.

1.1 Model concept

Filter theory is used as basic concept for the SDM. In this theory, the realized species assemblage is explained by a number of hierarchical filters, i.e. dispersal, abiotic and biotic filters [14,21]. Here, it is used because of its structural nature dividing the explanatory processes of species presence/absence in several filters. Only abiotic filtering is considered because the effect of pollutants on the species assemblage is assumed as the most relevant source of information for ecosystem health management.

Species response curves (or habitat preference curves) defining the biological response to abiotic gradients are used as to reflect the abiotic filters. The biological response can be expressed by many measures, i.e. species presence, abundance, density, usable area or volume. Species presence is used as a measure for biological response because it is assumed to be a robust measure for biological response [6].

In this paper, fine-scale and large-scale abiotic filters are considered. Fine-scale filters are filters acting at a local scale filtering species due to point specific pollution. In addition, the river typology characteristics (e.g. geology, river/catchment slope, ..) are considered to be large-scale filters which act on a river or catchment scale [21].

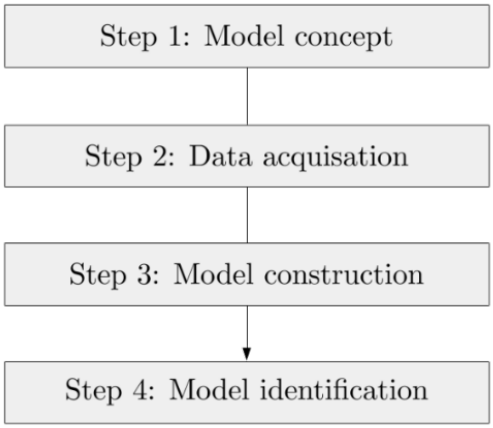

Fig. 1. Overview of methodology to develop SDMs (adapted after [2]).

1.2 Data

The Limnology Neerlandica database (http://www.stowa.nl/) and information on the river typology [4] are processed and compiled to a coupled database. The Limnodata is a database containing observations of the biology (macroinvertebrates, fish and macroflora) and physical-chemical state over 20 years in the Netherlands. The river typology is defined as a function of river catchment characteristics, i.e. average river slope, water source, average river width, catchment area, tidal influence, catchment geology [22].

The observations of the macroinvertebrate species *cloeon dipterum* are extracted from the Limnodata. The records are transformed from abundance to presence/absence in order to get an insight in the spatial and temporal distribution patterns of the species. Outliers in the physico-chemical data are investigated by inspecting summary statistics **(**mean, minimum, maximum and percentile values) and visually analysing box plots, histograms and dot plots. A number of variables are tested to physical boundaries. For instance, the width and depth of rivers are assessed as a function of the river type. In addition, the mass balance for nitrogen and phosphorus is inspected. In total 133 values are inspected in-depth leading to the omission of 102 records from the data. Finally, the correlation between variables is calculated so to exclude highly correlated variables and reduce dimensionality of the problem (Table 1).

Table 1. Overview of physico-chemical variables. #n: not included because of insufficient samples after coupling with biological data. ex. = excluded, corr. = correlated to, r = spearman rank correlation, min. = minimum, $\widetilde{X}$ = median, $\acute{X}$ = mean**,** max. = maximum, Chlor. a = chlorophyll a, Cond. = conductivity, Transp. = transparency, Kjel. = kjeldahl N, R. = river, Temp. = temperature.

variable	ex.	reason	min.	$\widetilde{X}$	$\acute{X}$	max.
%DO	x	Corr. DO (r = 0.89)	0.00	80.00	78.07	277.00
BOD_2	x	#n	10.00	10.00	92.87	2000.00
BOD_5			0.05	2.00	3.55	360.00
Chloride			1.00	40.00	56.99	1250.00
Chlor. a	x	#n	0.10	9.00	19.51	1170.00
COD	x	#n	2.00	26.00	32.57	200.00
Cond.	x	Corr. Chloride (r = 0.79)	0.50	50.00	52.48	542.00
DO			0.00	8.80	8.64	29.00
Transp.			0.00	0.50	0.50	3.00
Flow	x	#n	0.00	0.10	0.69	33.34
Kjel. N	x	Corr. NH_4 (r = 0.91)	0.00	1.70	2.60	70.00
NH_3-N			0.00	0.01	0.03	6.10
NH_4-N			0.00	0.40	1.13	80.00
NO_2-N			0.00	0.06	0.10	6.30
NO_3-N			0.00	3.50	4.89	64.00
PO_4-P			0.00	0.07	0.28	26.00
pH			3.60	7.40	7.33	10.40
R. depth	x	#n	0.00	0.40	0.66	5.00
R. width	x	#n	0.02	3.00	5.94	135.00
SO_4			1.00	62.00	68.15	6200.00
Temp.			-1.00	11.50	11.73	32.00
Total N	x	Corr. NO_3-N (r = 0.93)	0.05	5.56	7.07	66.30
Total P	x	Corr. to PO_4-P (r = 0.92)	0.00	0.20	0.47	29.00
Velocity	x	#n	0.00	20.00	24.15	300.00

1.3 Model construction

Species response curves are defined for the fine-scale filters (continuous variables). The species response curves are assumed to have a non-symmetric unimodal trapezoid shape chosen as a simplification of a bell-shaped curve [1,13]. The curves are allowed to be asymmetric so they can skew from extreme (heavy polluted) conditions [1,16]. Four parameters (a_1 , a_2 , a_3 and a_4) are used to define the trapezoid curve:

$$SI_f\left(x_i^j\right)=\begin{cases} 0 & \text{if } x_i^j < a_1 \\ \dfrac{\left(x_i^j-a_1\right)}{\left(a_2-a_1\right)} & \text{if } x_i^j \in [a_1, a_2[\\ 1 & \text{if } x_i^j \in [a_2, a_3] \\ \dfrac{\left(a_4-x_i^j\right)}{\left(a_4-a_3\right)} & \text{if } x_i^j \in]a_3, a_4] \\ 0 & \text{if } a_4 < x_i^j \end{cases} \quad (1)$$

With SI_f , the suitability index for the fine-scale filters, x_i^j , the input value i ($\in \{0,1,..,N\}$, n data points) for variable j ($\in \{0,1,..,M\}$, m variables). The parameters a_1 and a_4 describe the range of the conditions in which a species is able to survive. The parameters a_2 and a_3 describe the preferable range of conditions for the species (i.e. SI =1). The values of a_1 and a_4 are set by the minimum and maximum values of the observations for which the species is observed. For the large-scale abiotic filters, suitability indices are defined based on a set of parameters (a_1 , a_2 , .., and a_r) and the class (categorical):

$$SI_l\left(x_i^k\right)=\begin{cases} a_1 & if \quad x_i^k = C_1^k \\ a_2 & if \quad x_i^k = C_2^k \\ \cdots & \\ a_r & if \quad x_i^k = C_r^k \end{cases} \quad (1)$$

With SI_l , the suitability index for the large-scale filter, x_i^k , the input value i ($\in \{0,1,..,N\}$, n data points) for categorical variable k ($\in \{0,1,..,O\}$, o variables). The habitat suitability index (HSI) for a point i is calculated by multiplying the geometric mean for the fine and large-scale filters:

$$HSI_i = \left(\prod_{j=1}^{m} SI\left(x_i^j\right) \right)^{\frac{1}{m}} * \left(\prod_{k=1}^{o} SI\left(x_i^k\right) \right)^{\frac{1}{o}} \quad (1)$$

2.1 Model identification with simple genetic algorithms

The aim of the model identification tool is to identify a number of input variables and coupled species response curve parameters with an optimisation algorithm. This algorithm has to be able to efficiently search a large unconstrained space since it is difficult to a priori define the shape of a species response (skewed, Gaussian, …). In addition, a number of solutions is possibly more informative than one solution. Therefore, it is preferred to obtain an ensemble. The tool should be compatible with high performance computing to facilitate repeated runs for uncertainty analysis. Even more, it is required to be an open source package, available freely online, so to increase code and approach transparency (Figure 2).

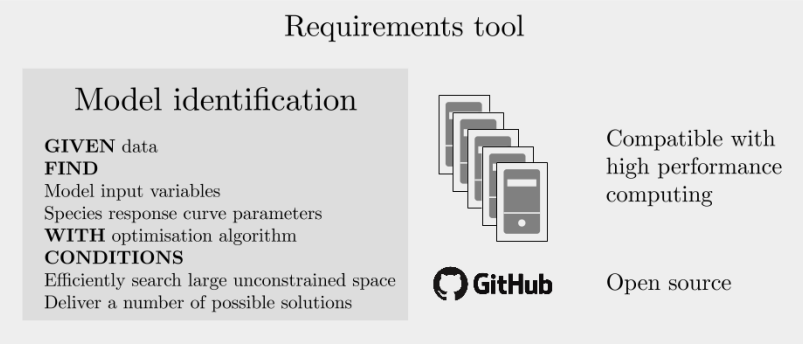

Fig. 1. Requirements for model identification tool for species distribution models.

An SGA with three operators, i.e. selection, crossover and mutation is implemented and used as optimisation algorithm. An SGA requires the encoding of the phenotype, i.e. the model, in a genotype. This genotype is typically coded as a binary string. This string is then translated to a model in a genotype-phenotype mapper. A list of lists is programmed to implement a variable length chromosome (Figure 3). The genome is defined by a second order binary string when a bit in the first order binary string has a value of one. The first order binary string is translated in a mapper by either in- or excluding the variable (one = present, zero =absent). The second order binary string is translated to parameter values of a_2 and a_3 in the mapper function by transforming every three bit sequence to an integer representation which is used to define the values for a_2 and a_3 (equation 1) and a_r (equation 2).

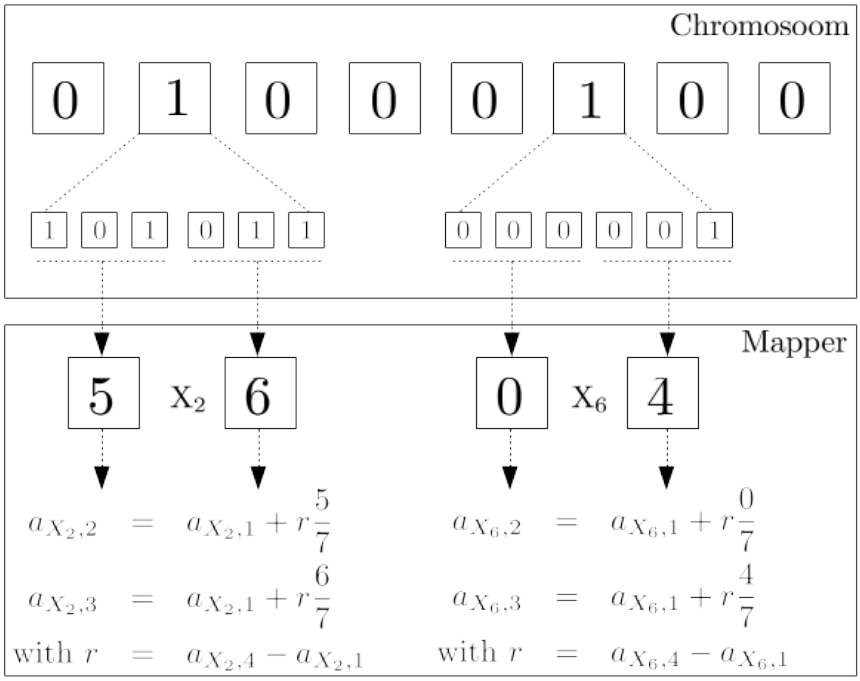

Fig. 1. Definition of chromosome and mapper function. The genome is programmed as a list of lists, where a second order binary string is defined when a bit of the first order string has the value of one. Every three bits of this second order string are translated to an integer which is used to define the values of the parameters a_2 , a_3 and a_r . In this example, the second and sixth variable are considered in the model. The parameters for the species response curves are defined by second order binary strings (six bits). The first three bits for variable X_2 are used to define $a_{X2,2}$ and the last three bits to define $a_{X2,3.}$ A binary coding is used to define a fraction (i.e. 5/7 and 6/7) of the total range r (a_4 - a_1) which is added to the parameter a_1 to obtain values for a_2 and a_3. Parameters a_2 and a_3 are respectively bounded by the range [a_1, a_3] and [a_2,a_4]. For the categorical variables, the parameters a_r are bounded by zero and one.

The tournament selection method is used to select the fittest individuals from a population as parents [11]. The selection rate defined as the fraction of the population that survives for the next step of mating is multiplied with the population size to obtain a number of parents. In the crossover operator, the parents are randomly paired to mate and produce offspring with a certain rate, i.e. crossover rate. If mating does not occur, the parents are replaced in the population. The last operator, mutation, is defined as the probability that a random gene is assigned a new value (0 → 1 or 1 → 0). The fitness of the chromosomes is the sum of squared errors (SSE) calculated with the HSI_i values and the observed presence or absence Pr_i :

$$Pr_i - HSI_i$$
$$¿¿$$
$$¿$$
$$¿$$
$$SSE = \sum_{i=1}^{n} ¿ \quad (1)$$

3 Results

The SGA is implemented and used to identify near-optimal models for the species *cloeon dipterum*. In the first part of this section, a set of hyper parameters (mutation and crossover rate) for the SGA are tested so to estimate the effect of hyper parameter choice on the algorithm performance. In the second part, the results found with the SGA and near-optimal hyper parameters are used to analyse the acquired model structure and performance.

3.1 Sensitivity of SGA

The SGA sensitivity as a function of the hyper parameter values are shown in Figure 4. For this experiment, an initial near-optimal set of parameters is determined by following the guidelines of [12]. The required number of chromosomes P are estimated by applying equation 5:

$$\frac{FE}{P}\log_{10}\left(1-\frac{1}{P}\right) = -M - \log_{10}\left(\sqrt{\frac{l}{12}}\right) \quad (1)$$

With M equal to three, FE, the number function evaluations determined by dividing the computational time available by the average runtime of one simulation and l, the chromosome length. For l, the maximum possible length of the chromosome is used (= 111 = three bits * (two parameters * 12 continues variables + 13 parameters for categorical variables)). With equation 5, 100 is found as a value for P. The mutation rate is calculated by dividing five by P (pm = 0.05*100 %) and the crossover rate (pc) is set to 100 % [12]. It is assumed that the performance of the SGA is near-optimal with these values. In order to verify the choice of the values, the sensitivity of the SGA performance to the values is checked by assessing the effect of the surrounding values of the found near-optimal values for the crossover and mutation rate (nine point grid with pc = {60, 85, 100} and pm = {1, 5, 10}).

The best found solution follows a similar evolution for the nine sets of hyper parameters with a varying convergence and performance. When inspecting the effect

of the mutation rate (pm) on the performance of the algorithm, one observes that the SGA analysis with a mutation rate of 1 % gives on average the best solutions (Figure 4, left panel). The evolution of the best solution found with a mutation rate of 5 % is similar whereas a higher mutation rate (10 %) leads to less optimal solutions. The initial speed with which these solutions are found is highest for a mutation rate of 5 %, however, the population converges - on average - earlier. For the crossover rate (Figure 4, right panel), one observes that the sensitivity of the performance is lower than for the mutation rate.

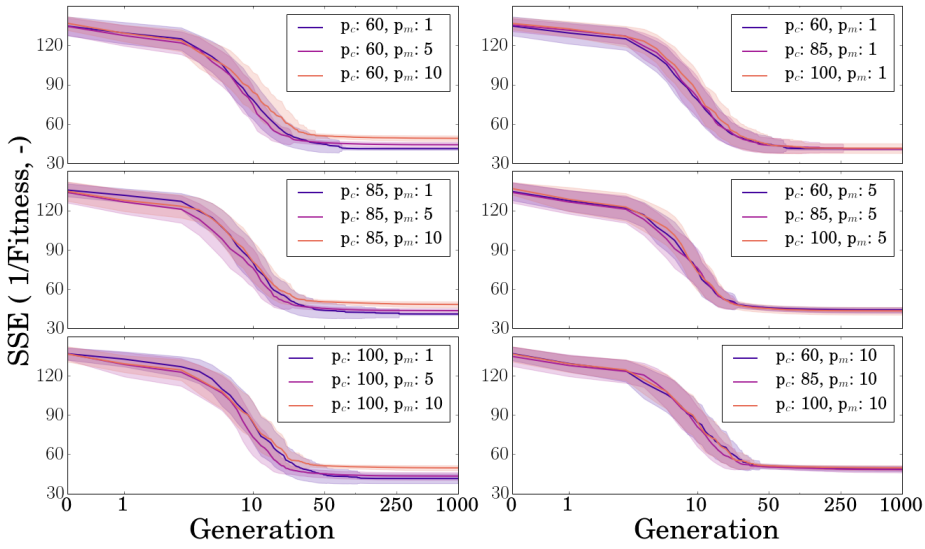

Fig. 1. Evolution of SSE (inverse of fitness) as a function of the number of generations. On the left, the results are shown for varying mutation rates and constant crossover rates. On the right, the results are shown for constant mutation rates and varying crossover rates. The uncertainty on the analysis is acquired by repeating the SGA a number of times with different initial conditions and preserving the best solution every generation.

One observes that a varying degree of uncertainty is observed for different hyper parameter values (Figure 4 and 5). The uncertainty is estimated by repeatedly running the SGA with a number of initial conditions and preserving the best solution over the generations for every SGA run. The variation of this uncertainty follows a hyperbole as a function of the mean SSE (and thus the generation) (Figure 5, left panel). At the point of convergence (low SSE, Figure 5, right panel), the uncertainty on the found near-optimal solutions for a crossover rate of 85 % is lower than for a crossover rate of 100 %. This seems to suggest that the crossover rate of 85 % is an appropriate choice to reduce SGA analysis uncertainty.

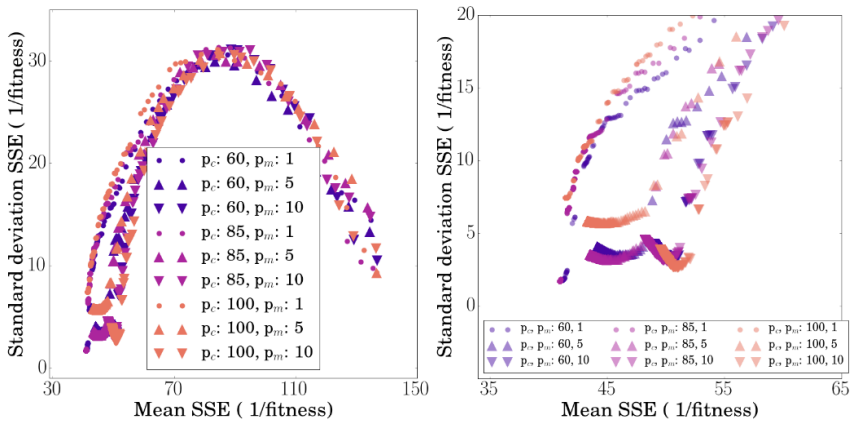

Fig. 1. Standard deviation on SSE as a function of the mean value of the SSE, for varying values of the hyper parameters (%). The right panel zooms in a narrower range of the left panel.

3.2 Analysis of identified SDMs

The acquired models with the SGA (pm= 1 %, pc = 85 %) are evaluated by calculating the Cohen's Kappa (Kappa) and area under the receiver operator curve (AUC) (see [20] for mathematical description). The acquired models are assessed to have a fair to satisfying performance. The mean Kappa is equal to 0.33±0.03 which is assessed as fair (Kappa $\in [0.2, 0.4]$, see [8]). The mean AUC is equal to 0.7±0.03 which is assessed as satisfying (AUC > 0.7, see [19]). In Figure 6, the model structure and accompanied uncertainty found by repeatedly running the SGA is shown. The support (%) for a model variable is calculated as a measure of variable importance by dividing the number of times a variable is selected by the SGA by the total number of SGA analysis. The support for the variable pH is very high (99 %) whereas the support for the river slope, catchment area, tides and geology is lower and uncertain.

In Figure 7, the species response curves and the accompanied uncertainty for the variable pH is shown. Either a response with very steep boundaries or a triangular response is observed. The uncertainty is shown for three values of the mutation rate (constant crossover rate). It is observed that the uncertainty on the acquired curves increases for higher mutation rate. This patterns is similar to the increase of uncertainty in the convergence of the objective function (Figure 5, right panel). When inspecting the uncertainty on the parameters of the categorical variables (not shown here), one observes a rather high uncertainty. In conclusion, the uncertainty in the objective function is reflected in the uncertainty of the model structure.

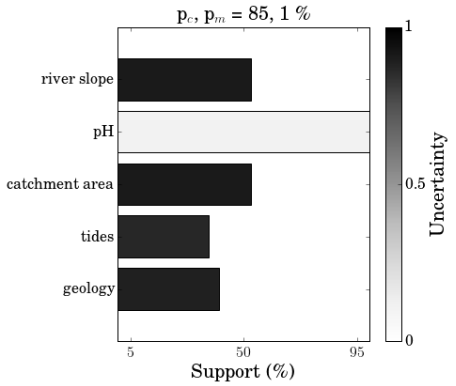

Fig. 1. Support for variable inclusion for repeated SGA analysis (p_c = 85 %, p_m = 1 %). The support is calculated by dividing the number of times a variable is selected by the SGA by the total number of analysis (i.e. 100). The uncertainty is estimated with the Shannon entropy [22].

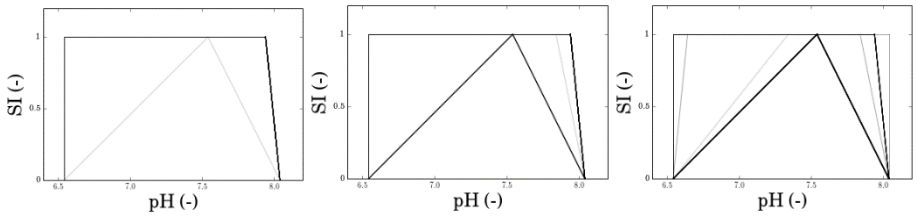

Fig. 1. Uncertainty on acquired species response curves with the repeated SGA analysis. From left to right, a mutation rate of 1, 5 and 10 % is used (constant crossover rate = 85 %).

4 Discussion and outlook

In this paper, a variable length chromosome SGA is implemented and used to jointly perform IVS and PE. The implemented algorithm is able to identify fair to satisfying models. The uncertainty on the acquired species response curve parameters is rather low, at least for the variable with a high support. In addition, it is observed that the uncertainty on the acquired near-optimal solution is not equal over different values of the mutation and crossover rate.

The accuracy of the models could be improved by increasing the precision of the binary encoding used for the algorithm. In the current implementation every three bits code one parameter of the species response curves (see Figure 3). This allows to encode eight discrete values for every parameter. The representation restricts the possible parameter values to a limited set defined by the lower and upper boundary of the parameter interval and the number of bits [25]. Increasing the number of bits for the binary encoding might increase the precision but will also increase the length of the chromosome. Consequently, different near-optimal values for the hyper parameters will be obtained with equation 5. When testing the required number of

chromosomes, for a fixed number of FE, one observes that the found number of chromosomes (and thus mutation rate, see [12]) is almost equal for higher chromosome lengths. For example, for a bit length of three, a maximum chromosomes length of 111 (three bits * (two parameters * 12 continues variables + 13 parameter for categorical variables)) leads to a population size of 112, whereas for a four and six bit problem (l = 148, l = 185) a number of 111 and 110 chromosomes is found. Since the determined near-optimal values for the hyper parameters for varying chromosome lengths does not vary, it is expected that the performance and uncertainty of the SGA will not vary as a function of the length used to encode the optimisation. This suggests that increasing the precision of the binary encoding will not influence the performance and uncertainty of the SGA analysis. Additional experiments with the SGA should confirm this hypothesis.

A hyperbolic relation is found between the uncertainty on the SGA analysis and the found near-optimal solution. At the start of the analysis, the uncertainty is rather small, and increases with the number of generations to finally converge to a value as the SGA converges. There are differences in the amount of uncertainty at convergence for varying values of mutation and crossover rate. For low mutation rates, the uncertainty on the found near-optimal solution declines as the crossover rate is lowered. For higher mutation rates, this relation is inverse but less apparent. In general the guidelines by [12] are assessed as appropriate for these type of problems, since with these settings the SGA is able to reduce the prediction error of the models (mean SSE declines from approximately 140 to 40). Options to further improve the algorithm performance can be to improve the exploitive character of the algorithm by combing the genetic algorithm with a hill climbing (HC) approach or to vary the mutation and crossover rates over the generations. Further research can investigate whether these implementation have a significant added value for SDM identification and whether they can reduce the uncertainty of the analysis.

Genetic algorithms have shown to be valuable for PE and IVS in species distribution modelling [3,5]. In this study, a variable length chromosome implementation of an SGA is presented to jointly perform PE and IVS. The results tested for one species are promising, however, it should be further investigated how the performance of the algorithm varies as function of the algorithm settings. In addition, the approach should be validated by applying the SGA for different species.

The current available software is an open source package implemented in the Python programming language [9,10]. Many other packages are available (Generalized Linear Models, GLM, in the R programming language or Genetic Algorithm for Rule set Production/Prediction (GARP) software [24]). For instance, the GLM R package is an user-friendly package useful for ecologist, however, automated running a number of analysis to estimate uncertainty is difficult. Even more, the statistical approaches present a number of boundary conditions to the shape of the species response. In the developed approach, these boundary conditions are relaxed as is the case for GARP. The difference with GARP is that the SDMIT approach is designed to run on high performance clusters whereas GARP was initially designed for single-run analysis in a graphical user interface environment. In addition, machine learning approaches like decision trees and support vector machines are

available [7]), however, the disadvantage of these approaches is that the tools are not implemented specifically for the optimisation of SDMs and thus often lack the ecological theoretical background. Consequently they are used as data mining approaches rather than model optimisation algorithms. The SDMIT packages is an answer to these limitations. With this, SDMs can obtained that improve the insight in species and community response to environmental changes.

Acknowledgments. Sacha Gobeyn is supported by a Bijzonder Onderzoeksfonds (BOF) project related to the Ecuador Biodiversity Network of the Vlaamse Interuniversitaire Raad-Universitaire Ontwikkelingssamenwerking (VLIR-UOS). This research was performed in the context of the VLIR Ecuador Biodiversity Network project. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government - department EW.

References

1. Austin MP (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol Modell 200:1–19. doi: 10.1016/j.ecolmodel.2006.07.005
2. Bennetsen E, Gobeyn S, Goethals PLM (2016) Species distribution models grounded in ecological theory for decision support in river management. Ecol Modell 325:1–12. doi: 10.1016/j.ecolmodel.2015.12.016
3. D'heygere T, Goethals PLM, De Pauw N (2003) Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. Ecol Modell 160:291–300. doi: Pii S0304-3800(02)00260-0
4. Elbersen JWH, Verdonschot PFM, Roels B, Hartholt JG (2003) Definitiestudie kaderrichtlijn water (KRW) I. Typologie nederlandse oppervlaktewateren. Alterra, Research Instituut voor de Groene Ruimte, Wageningen.
5. Fukuda S, De Baets B, Mouton AM, et al (2011) Effect of model formulation on the optimization of a genetic Takagi-Sugeno fuzzy system for fish habitat suitability evaluation. Ecol Modell 222:1401–1413. doi: 10.1016/j.ecolmodel.2011.01.023
6. Fukuda S, Mouton AM, De Baets B (2012) Abundance versus presence/absence data for modelling fish habitat preference with a genetic Takagi-Sugeno fuzzy system. Environ Monit Assess 184:6159–6171. doi: 10.1007/s10661-011-2410-2
1. Fukuda S, De Baets B, Waegeman W, Verwaeren J., Mouton, AM (2013) Habitat prediction and knowledge extraction for spawning European grayling (Thymallus thymallus L.) using a broad range of species distribution models. Environ Model Softw 47:1–6. doi: 10.1016/j.envsoft.2013.04.005
2. Gabriels W, Goethals PLM, Dedecker AP, 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. doi: 10.1007/s10452-007-9081-7
3. Gobeyn S. Species distribution model identification tool (SDMIT); software available at https://sachagobeyn.github.io/SDMIT/

4. Gobeyn S, Volk M, Dominguez-Granda L, Goethals PLM (2017) Input variable selection with a simple genetic algorithm for conceptual species distribution models: A case study of river pollution in Ecuador. Environ Model Softw 92:269–316. doi: 10.1016/j.envsoft.2017.02.012
5. Goldberg DE, Deb K (1991) A comparative analysis of selection schemes used in genetic algorithms. Found Genet Algorithms 1:69–93. doi: 10.1.1.101.9494
7. Gibbs MS, Dandy GC, Maier HR (2008) A genetic algorithm calibration method based on convergence due to genetic drift. Inf Sci 178:2857–2869. doi: 10.1016/j.ins.2008.03.012
8. Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. Ecol Modell 135:147–186. doi: 10.1016/S0304-3800(00)00354-9
9. Guisan A, Rahbek C (2011) SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. J Biogeogr 38:1433–1444. doi: 10.1111/j.1365-2699.2011.02550.x
10. Hamblin S (2013) On the practical usage of genetic algorithms in ecology and evolution. Methods Ecol Evol 4:184–194. doi: 10.1111/2041-210X.12000
11. Hirzel AH, Le Lay G (2008) Habitat suitability modelling and niche theory. J Appl Ecol 45:1372–1381. doi: 10.1111/j.1365-2664.2008.01524.x
12. Li X, Wang Y (2013) Applying various algorithms for species distribution modelling. Integr Zool 8:124–135. doi: 10.1111/1749-4877.12000
13. Maier HR, Kapelan Z, Kasprzyk J, et al (2014) Evolutionary algorithms and other metaheuristics in water resources: Current status, research challenges and future directions. Environ Model Softw 62:271–299. doi: 10.1016/j.envsoft.2014.09.013
6. Manel S, Ceri WH, Ormerod SJ (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. J Appl E 38:921–931. doi: 10.1046/j.1365-2664.2001.00647.x
14. Mouton AM, De Baets B, Goethals PLM (2010) Ecological relevance of performance criteria for species distribution models. Ecol Modell 221:1995–2002. doi: 10.1016/j.ecolmodel.2010.04.017
15. Poff NL (1997) Landscape filters and species traits: towards mechanistic understanding and prediction in stream ecology. J North Am Benthol Soc 16:391–409. doi: 10.2307/1468026
16. Sandin L, Verdonschot PFM (2006) Stream and river typologies - major results and conclusions from the STAR project. Hydrobiologia 566:33–37. doi: 10.1007/s10750-006-0072-9
17. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423. doi: 10.1145/584091.584093
18. Stockwell D (1999) The GARP modelling system: problems and solutions to automated spatial prediction. Int J Geogr Inf Sci 13:143–158. doi: 10.1080/136588199241391
19. Van Broekhoven E, Adriaenssens V, De Baets B (2007) Interpretability-preserving genetic optimization of linguistic terms in fuzzy models for fuzzy ordered classification: an ecological case study. Int J Approx Reason 44:65–90. doi: 10.1016/j.ijar.2006.03.003