Environmental Modelling & Software 92 (2017) 269–316

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Input variable selection with a simple genetic algorithm for conceptual species distribution models: A case study of river pollution in Ecuador

Sacha Gobeyn [a, *], Martin Volk [b], Luis Dominguez-Granda [c], Peter L.M. Goethals [a]

[a] Ghent University, Laboratory of Environmental Toxicology and Aquatic Ecology, J. Plateaustraat 22, B-9000 Ghent, Belgium
[b] UFZ - Helmholtz Centre for Environmental Research, Department of Computational Landscape Ecology, Permoserstr. 15, 04318 Leipzig, Germany
[c] Centro del Agua y Desarrollo Sustentable, Escuela Superior Politécnica del Litoral (ESPOL), Campus Gustavo Galindo, Km. 30.5 Via Perimetral, PO Box 09-01-5863, Guayaquil, Ecuador

ARTICLE INFO

Article history:
Received 8 July 2016
Received in revised form
6 January 2017
Accepted 15 February 2017

Keywords:
Conceptual species distribution models
Input variable selection
Simple genetic algorithms
Species response curves
River pollution
Freshwater management

ABSTRACT

Species distribution models (SDMs) have received increasing attention in freshwater management to support decision making. Existing SDMs are mainly data-driven and often developed with statistical and machine learning methods but with little consideration of hypothetic ecological knowledge. Conceptual SDMs exist, but lack in performance, making them less interesting for decision management. Therefore, there is a need for model identification tools that search for alternative model formulations. This paper presents a methodology, illustrated with the example of river pollution in Ecuador, using a simple genetic algorithm (SGA) to identify well performing SDMs by means of an input variable selection (IVS). An analysis for 14 macroinvertebrate taxa shows that the SGA is able to identify well performing SDMs. It is observed that uncertainty on the model structure is relatively large. The developed tool can aid model developers and decision makers to obtain insights in driving factors shaping the species assemblage.

Software availability

Name: Species Distribution Model Identification Tool (SDMIT)
Developer: Sacha Gobeyn, Coupure Links 653, B-9000 Ghent, sacha.gobeyn@ugent.be, sachagobeyn@gmail.com
Program language: Python 2.7.10
Software requirement: Anaconda Python distribution package (Continuum Analytics, https://www.continuum.io/why-anaconda)
Source code: https://github.com/Sachagobeyn/SDMIT/releases/tag/v1.0.0, https://github.com/Sachagobeyn/SDMIT
License: CC BY 4.0 Creative Commons

1. Introduction

Species distribution models (SDMs) are increasingly used by scientists and policy makers as tools to investigate a wide variety of ecological problems. With respect to freshwater management, different SDMs are developed and used to quantify the impact of anthropogenic pressures (e.g. increased nutrient concentrations, loss of habitat) on the distribution and diversity of macroinvertebrate taxa (Van Broekhoven et al., 2006; Domisch et al., 2013; Boets et al., 2015; Gies et al., 2015). Statistical and machine learning methods, used to identify SDMs with data, are useful, but have also been criticised because they often lack in conceptual background of ecology theory (Austin, 2007; Fukuda et al., 2013).

All SDMs are positioned along an axis of purely data-driven (data influenced) to conceptual (hypothetic influenced) models (Beale and Lennon, 2012; Mount et al., 2016). No model is solely data-driven or conceptual, but current SDM approaches often rely on statistical and machine learning approaches rather than on hypothetic knowledge (Austin, 2007). This causes the SDMs being characterised by statistical assumptions and properties of the used ecological data. Consequently, the models are vulnerable to imperfections and uncertainties in the data.

In contrast, conceptual SDMs primarily aim to reflect ecological concepts. For instance, BIOCLIM, one of the first SDMs, aims to delineate a number of environmental envelopes reflecting the total

* Corresponding author.
E-mail address: Sacha.Gobeyn@UGent.be (S. Gobeyn).

http://dx.doi.org/10.1016/j.envsoft.2017.02.012
1364-8152/

range of environmental space permitting a positive growth of a species (Booth et al., 2014). This approach is directly linked with the concepts of niche theory presented by Hutchinson (1957). In another example, the Genetic Algorithm for Rule Set Production (GARP) modelling system of Stockwell and Noble (1992) is an approach explicitly searching for heuristic rules of environmental boundaries so to explain species presence. In this approach, environmental thresholds, which reflect the boundaries in the environmental space limiting species distribution, are searched. Although both methods know many applications, they are often outperformed by novel machine learning approaches, making them possibly less interesting for application (Elith et al., 2006; Hamblin, 2013).

The interface between data-driven and conceptual models poses the challenge to develop perfomant SDMs with sufficient conceptual ground (Austin, 2007; Guisan and Rahbek, 2011; Bennetsen et al., 2016). Starting with conceptual approaches, one is challenged to identify well performing models. Identifying (an) alternate model(s) from a set of models with data is often a difficult task, especially when many candidate models can be formulated (i.e. different possible input variables, interactions, model parameters). Computer algorithms, using metaheuristics, have proven valuable to tackle this challenge (Maier et al., 2014).

In this paper, an approach using a simple genetic algorithm (SGA), a type of heuristic search algorithm, is presented to identify alternative formulations for a conceptual SDM. The process of identifying alternative formulations is defined here as model identification, and consists of identifying values for the parameters, selection of input variables and interaction functions. The approach is illustrated by implementing and using an SGA to perform input variable selection (IVS), for a case study of ecological water quality (EWQ) assessment and river pollution in the Guayas River basin, Ecuador. The case study is selected because the anthropogenic pressures on the river system are relatively straightforward to interpret. The remainder of the paper is structured as follows: section 2 presents a methodological framework for model identification of conceptual models with SGAs and evolutionary algorithms (EAs) in general. In section 3, the methodology is illustrated with the Guayas River case study and in section 4, the approach and results are discussed.

2. Model identification in conceptual SDMs

In this section, we present an approach to identify alternative formulations for a conceptual SDM. In the first part (section 2.1), the concepts commonly used in species distribution modelling are shortly addressed. In section 2.2, an approach to develop a conceptual SDMs is presented, with specific focus on model identification. In the last part (section 2.3), the use of EAs as a tool to identify alternative models is presented. SGAs are classified under EAs and are used as specific implementation of EAs in section 3.

2.1. Preliminaries

The basis for every model is the definition of a conceptual model, which involves the formulation of assumptions and underlying ecological theories and processes (Guisan and Zimmermann, 2000). Filter theory is an attractive starting point for a conceptual model, because this theory attempts to structure the processes driving species absence in a number of elements (Poff, 1997; Guisan and Rahbek, 2011). In filter theory, it is assumed that the realized species assemblage is the results of hierarchical filters, functioning as limits or barriers for species to exist in a local community. Typically, three types of filters are distinguished, each (inter-)acting on a specific scale: the dispersal, abiotic and biotic filters (Guisan and Rahbek, 2011). These hierarchical filters are assumed to act upon species traits, which are functional attributes of the species (for instance, body size, weight, reproduction, .).

Conceptually, filter theory is closely related to niche theory of Grinnell (1917) and Hutchinson (1957). A niche is defined as the environmental space which permits an intrinsic growth rate of a species. In the note of Hutchinson (1957), a distinction is made between a fundamental and realized niche. A realized niche is defined as a subset of a fundamental niche, not only constrained by abiotic factors, but also by dispersal (and historical) limitations and biotic interactions between the species (Soberón and Nakamura, 2009). In species distribution modelling, it is assumed that a realized niche is fitted, because in reality, species are observed in this niche (Guisan and Thuiller, 2005). However, Beale and Lennon (2012) state that it is preferable to model a fundamental niche rather than a realized niche, because the narrower precision of a realized niche might underestimate model uncertainty. An insight in this uncertainty is of major importance in - for instance - estimating the effect of climate change on species, since simulations are run for environmental conditions outside the fitted range.

2.2. Model identification in species distribution modelling

In Fig. 1, an approach for the development of a conceptual SDM is presented. The first step is to define and use a number of theoretical ecological concepts for the basis of the model (see section 2.1). Typically, filter theory and/or niche theory are used as an initial basis, because of their structural approach, subdividing driving processes in a number of elements (see for instance Guisan and Rahbek (2011) and Poff (1997)). With respect to species distribution modelling, the concept of habitat suitability or species response curves, defining the biological response to abiotic conditions, is often used to reflect the concept of abiotic filters of filter theory. In a second step, information on the attributes of the system are collected from databases, field samples and experts. Next, the model is constructed by defining a number of structural elements; input and output variables, interactions and parameters with data (see Bennetsen et al. (2016)) and/or expert knowledge (see Adriaenssens et al. (2006)). In the final step, alternative model structures are identified, typically by testing the conceptual model with data. Machine learning and statistical approaches can be used as a means to explore the space of possible alternative models. The

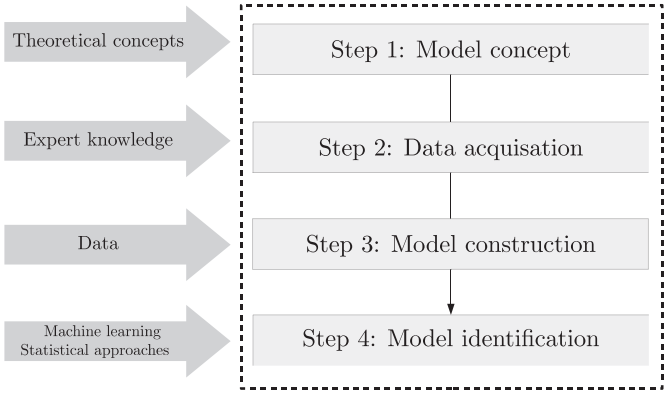

Fig. 1. Four step approach for model development of conceptual SDMs (adapted after Bennetsen et al. (2016)). In the first step, the used theoretical ecological concepts are formulated, whereas in the second step, information on the model attributes is collected. In step three, the model is constructed with the available information and finally, alternative model structures are identified. In all of these steps, ecological concepts, acquired information, expert knowledge and machine learning and statistical approaches can be used to justify and aid development.