Ecological Modelling xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Ecological Modelling

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

The effects of model and data complexity on predictions from species distributions models

David García-Callejas[a,b,*], Miguel B. Araújo[a,b,c,d]

[a] InBIO/CIBIO, University of Évora, Largo dos Colegiais, 7000 Évora, Portugal
[b] Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot, United Kingdom
[c] National Museum of Natural Sciences, Calle Jose Gutierrez Abascal, 2, 28006 Madrid, Spain
[d] Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Climate change
Data complexity
Model complexity
Species distributions models
Transferability
Virtual species

ABSTRACT

How complex does a model need to be to provide useful predictions is a matter of continuous debate across environmental sciences. In the species distributions modelling literature, studies have demonstrated that more complex models tend to provide better fits. However, studies have also shown that predictive performance does not always increase with complexity. Testing of species distributions models is challenging because independent data for testing are often lacking, but a more general problem is that model complexity has never been formally described in such studies. Here, we systematically examine predictive performance of models against data and models of varying complexity. We introduce the concept of computational complexity, widely used in theoretical computer sciences, to quantify model complexity. In addition, complexity of species distributional data is characterized by their geometrical properties. Tests involved analysis of models' ability to predict virtual species distributions in the same region and the same time as used for training the models, and to project distributions in different times under climate change. Of the eight species distribution models analyzed five (Random Forest, boosted regression trees, generalized additive models, multivariate adaptive regression splines, MaxEnt) showed similar performance despite differences in computational complexity. The ability of models to forecast distributions under climate change was also not affected by model complexity. In contrast, geometrical characteristics of the data were related to model performance in several ways: complex datasets were consistently more difficult to model, and the complexity of the data was affected by the choice of predictors and the type of data analyzed. Given our definition of complexity, our study contradicts the widely held view that the complexity of species distributions models has significant effects in their predictive ability while findings support for previous observations that the properties of species distributions data and their relationship with the environment are strong predictors of model success.

1. Introduction

Understanding why species distribute as they do is a central problem in ecology. Current methods for studying species distributions often involve statistical or numerical models that relate the distributions of species with layers of environmental information. The use of correlative species distributions models (also known as bioclimatic envelope models, habitat suitability models, and ecological niche models; for definitions of these seemingly related terms see Araújo and Peterson, 2012) is currently the most widespread approach due to their versatility, ease of use, and modest data requirements (e.g., Guisan and Zimmermann, 2000; Elith and Leathwick, 2009). Yet, despite widespread use of these models, the debate as to what is the best modelling approach is far from settled (Araújo and Rahbek, 2006), and predictions from alternative models can be markedly different in the context of spatial (e.g., Randin et al., 2006; Duncan et al., 2009; Heikkinen et al., 2012) and temporal transferability (e.g., Thuiller, 2004; Araújo et al., 2005a; Pearson et al., 2006; Zanini et al., 2009). Previous tests of performance of species distributions models have led to the conclusion that more complex models were generally better than simpler ones (e.g., Segurado and Araújo, 2004; Elith et al., 2006). However, model performance is typically inflated when test data that are not independent from data is used to train the models, such as when

* Corresponding author at: Centre for Ecological Research and Forestry Applications (CREAF), Universitat Autónoma de Barcelona, Bellaterra, Spain.
Tel.: +34 935868388.
E-mail address: david.garcia.callejas@gmail.com (D. García-Callejas).

http://dx.doi.org/10.1016/j.ecolmodel.2015.06.002
0304-3800/

data are randomly split between training and test sets (Araújo et al., 2005b; but see Madon et al., 2013). In the few cases in which models have been tested for transferability using independent data (from another region or another time), no clear relationship between the perceived complexity of the models and their performance was found (Araújo et al., 2005b; Randin et al., 2006; Dobrowski et al., 2011; Heikkinen et al., 2012; Smith et al., 2013).

Models perceived as 'simple' usually have procedures for fitting the data that are easier to grasp, and/or perform fewer and/or simpler operations with the data. In contrast, models perceived as 'complex' involve procedure for fitting the data that are more difficult to comprehend while usually performing a significant number of operations in order to produce the desired outcome. It is implicitly assumed that this loose definition of complexity is related to the capacity of different models to produce either 'simple' or 'complex' response curves (e.g., Elith et al., 2006; Merow et al., 2014). When selecting the best model for a given problem, it is expected that parsimony should lead to selecting models that minimize overall prediction error by finding an optimal balance between the error in fitting the training data (also referred to as parameter estimation error or bias) and the error in generalizing to new datasets (also referred to as approximation error, or variance). Models that are too simple would fit training data poorly (high bias), while overly complex models would generate low bias and high variance as they would capture random error or biases in the data.

The concept of model complexity is central to the endeavour of finding optimal models for predictive purposes. Yet, measuring model complexity is not straightforward. Here, we attempt to formalize one of the key aspects of model complexity (algorithmic or computational complexity, see Section 1.1: computational complexity), and test whether the principle of parsimony can guide identification of the optimal model complexity for predicting species distributions in space and time. In addition, we quantify structural complexity in the response data (i.e., presence and absence of species) and examine how data complexity affects the predictive abilities of models. To overcome problems of data availability, we simulated virtual species across a realistic geographical domain with different sets of environmental predictors, and compare some of the results with empirical presence–absence records within the same geographical domain.

1.1. Computational complexity

The computational complexity of an algorithm is defined by the amount of computational resources it requires to produce an output (Arora and Barak, 2009). This definition stems from the idea that an algorithm processes an input via a certain number of elementary operations, and these operations consume varying amounts of computing time. The computing time spent by the algorithm is, thus, an approximation to the complexity of the operations performed on the input. Complex algorithms inevitably perform more complex operations on their input than simple ones, thereby requiring more computation time to solve a particular task. Numerical analyses of computational complexity treat algorithms as black boxes, disregarding their internal structure, functional form or any other specificity. Such analyses are, therefore, suitable when the goal is to compare different methodologies on equal footing.

Computational complexity is also referred to as time complexity or algorithmic complexity, and it is commonly expressed by the O notation (read 'big o'). This notation identifies the time complexity of an algorithm by the highest-order term of its growth rate as a function of input size, suppressing lower-order terms and constants. It is an asymptotic measure of complexity; as input size increases, so does the importance of the dominant term in characterizing computation time. For example, an algorithm may take $3x^2 + 2x$ time units in solving a problem of size x. As x approaches infinity, the higher-order term (x^2) will tend to take over computation time, and the lower-order terms, as well as the multiplicative coefficients, will become irrelevant. This particular algorithm is thus said to have a time complexity of $O(n^2)$. If the computation time of an algorithm is independent of the dataset size, it is said to be a constant time algorithm, expressed as a time complexity of $O(1)$. As this methodology aims to estimate the asymptotical behaviour of a given algorithm, it also bypasses the issue of comparing algorithms written in different programming languages: the computational cost of a given algorithm implemented in two different languages is assumed to be proportional, up to a multiplicative constant that will become irrelevant asymptotically. The chief assumption of the method is that the algorithms being compared are efficiently programmed, i.e., there are no spurious tasks within the algorithms consuming computation time. A full treatment of computational complexity is out of the scope of this study (but see Arora and Barak, 2009; Papadimitriou, 1994).

Consistent with the principle of parsimony and assuming that algorithmic complexity is a good proxy for overall model complexity, the highest predictive capacity should be expected in models of intermediate algorithmic complexity. It is worth noting that the quantification of algorithmic complexity is independent of the modelling methodology. That is, the framework implemented herein with correlative species distributions models, could also be easily implemented with alternative mechanistic approaches for modelling species distributions (e.g., Fordham et al., 2013; García-Valdés et al., 2015).

1.2. Geometrical complexity of the data

Estimating the ecological niche of a species is an instance of the broad class of problems in which a set of points (in environmental space) must be classified into one of two opposing classes (presence/absence) according to some relationship between the dimensions of the space and the class to which each point belongs. The difficulty in estimating the ecological niche of a species can be assessed by evaluating the geometrical structure of the boundary between classes in the training data. Aside from deficiencies and biases in the data collection (Barry and Elith, 2006; Araújo et al., 2009), the internal structure of the data and its relationship with models predictive capacity has never been formally explored (but see Blonder et al., 2014). It has been, though, extensively addressed in other scientific fields; particularly within the machine learning community where the concept of *geometrical complexity* has been developed. Given a dataset with a two-class categorical response and N predictors, the geometrical complexity is defined as an approximation of the structural characteristics of the N-dimensional boundary separating the response classes (Basu and Ho, 2006). It is a general measure defined by a set of complementary metrics (see Section 2). When analyzed together, these metrics help differentiate datasets with geometrically simple class boundaries from those with complex and/or random class boundaries.

We predict that data complexity is related to the predictive capacity of the models. Specifically, simpler datasets will tend to reflect simpler occurrence–environment relationships thus being easier to model and yielding comparatively better performance than models trained with more complex datasets.

2. Materials and methods

2.1. Virtual species generation

We created three different types of virtual species. Their distributions were projected across mainland Spain by defining environmental suitability landscapes based on different sets of

Please cite this article in press as: García-Callejas, D., Araújo, M.B., The effects of model and data complexity on predictions from species distributions models. Ecol. Model. (2015), http://dx.doi.org/10.1016/j.ecolmodel.2015.06.002