DETERMINING THE MINIMAL BACKGROUND AREA FOR MAXENT SPECIES DISTRIBUTION MODELS: MinBAR PACKAGE

22/9/2018

# Introduction

Species Distribution Models (SDM) is a widely used tool especially because of the ease of use of several software, such as MaxEnt (Phillips, Anderson, and Schapire 2006) or BIOMOD (Thuiller et al. 2009). Also, thanks to the development of the R-programming community, and the free availability of biodiversity data in public repositories like GBIF (cita), SDMs have a growing popularity in the fields of ecology and biodiversity conservation. However, such availability of data is often limited to species presences only (also called occurrences) and with a lack of data in poorly sampled areas. These facts, among others, limit the use of some techniques or algorithms and force to make critical assumptions and choices, which introduce different levels of uncertainty to model predictions (Jarnevich et al 2017). One of the main choices when using pseudo-absences approaches is the delineation of the background area to fit the model, also called “landscape of interest” or “study area” (Elith et al, 2011; Raes, 2012). However, it is not clear how to define its extent. Elith et al (2011), for instance, said that it has to be defined by the ecologist and limited by geographic boundaries or by how far the species can be dispersed. Other authors also included as constraints the interactions with other species or sampling biases in the data set (Jarnevich et al 2017). Despite that, there are still some situations in which it is difficult to define exactly an accurate background area, often due either to the lack of available data or to the limited knowledge of the species biology (e.g. species inhabiting on islands or archipelagos). In addition, it is also common that the study is focused in a country or region. Then the background area is constrained to an artificial or political boundary, although the species distribution might be wider (El-Gabbas & Dormann, 2018). Finally, another difficult situation is when the extent of the species is so big that it makes the computations to fit the model and make predictions highly resource-demanding and time-consuming. These two limitations are particularly important when the study encompasses a big number of species with a large geographic range. Any of these three situations usually lead to fit partial models. Nevertheless, sometimes this does not suppose a reduction of model performance (El-Gabbas & Dormann, 2018). Therefore, the hypothesis here is that there is a minimum background area around the centre of the species distribution (minimum buffer), which characterize well enough the range of environmental conditions needed by the species to survive. Thus, fitting the SDM within this area should be the optimal solution in terms of both quality of the model and time of execution.

# MinBAR overview

MinBAR is an R package that aims at (1) defining what is the minimum or optimal background extent necessary to fit good partial SDMs and/or (2) determining if the background area used to fit a partial SDM is reliable enough to extract ecologically relevant conclusions from it.

## Problem

On the one hand, fitting partial SDMs might lead to underestimated predictions of species’ distribution or to biased descriptions of their niches (Sanchez-Fernandez et al, 2011). On the other hand, making model calibrations and predictions of species with a large geographic range can demand a huge amount of computer resources in terms of time and memory. To solve these problems, the idea behind the MinBAR package is to sequentially fit several concentric SDMs, with different diameter each (i.e. buffers), from the centre of the species distribution to the periphery, until a satisfactory model is reached.

## Evaluation metrics

A certain controversy exists in the best way to evaluate the performance of SDMs. One of the most widely used is the area under the ROC curve (AUC), although it has received several critiques because of its misuse. For the purpose of MinBAR, AUC is not the best choice because it is only useful for assessing the performance of different models that use exactly the same background area. For that reason, this package uses the Boyce Index (Hirzel et al. 2006), implemented in the R package ecospat. However, AUC is also calculated and gathered in the outputs, although it is not used to derive conclusions. Boyce Index (BI) is a presence-only and threshold-independent evaluator for SDMs. Among others, it is adequate in situations where the model uses background data instead of true absences, as is the case of MaxEnt (). It varies between -1 and 1, where positive values indicate consistent model predictions; values close to zero indicate predictions not better than those from a random model; and negative values imply bad predictions. See Hirzel et al. (2006), Di Cola et al. (2017) and references therein for further details on how BI is calculated as well as its strengths and weaknesses. Here, in order to evaluate the predictive performance of the models, the package includes two different metrics. On the one hand, Boyce Index Partial (BoyceIndex\_part or BI\_part) evaluates the accuracy of the predictions within the buffer, or what is the same, in the training area. On the other hand, Boyce Index Total (BoyceIndex\_tot or BI\_tot) assesses the predictions beyond the training area, across the whole distribution of the species (i.e. transferability of the model). ##*minba*: The main function The main function of MinBAR is minba. In the version 1.0.0 of the package presented here, minba is implemented for MaxEnt models and uses the Boyce Index as a measure of model performance. This function firstly loads the presences data set and the explanatory variables. Secondly, it calculates the centre of the species distribution, the most distant occurrence and the buffers. The buffers are not defined by equal distance, but by % of presences equally distributed. This is particularly useful for very discontinuous distributions (e.g. introduced or invasive species), while not affecting more aggregated populations. Thirdly, minba makes n models for each buffer in a loop and calculates averages. In this third step, it crops the variables to the extent of the buffer +5%, and calculates the number of necessary pseudo-absences to cover the 50% of the pixels within the buffer (Guevara et al, 2017). It uses 70% of the presences to calibrate the model and 30% to evaluate it, all from within the buffer (Boyce Index Partial). It also makes predictions and evaluations for the whole extent of the species +5% (Boyce Index Total). To do that, it uses 30% of all presences excluding those used to calibrate the model. At this point, while the process is looping over the buffers, minba gives several options. The user can choose either (1) to run the models for all the buffers to see if the selected background area is accurate and how the quality of the models evolves, or (2) to stop the process when it reaches certain conditions, which can be defined by the user as well. The latter option is adequate for very large species distributions. In this case, the user also has several options, mainly depending in the aim of the study. On the one hand, if the interest is related to the characteristics of the population (e.g. description of the ecological niche, etc.), the focus should be more in the Boyce Index Partial. On the other hand, if the model wants to be projected in time or space, the focus should fall on the Boyce Index Total. In turn, both approaches have two possibilities: (a) fixing a minimum Boyce Index to stop the process when it is reached; or (b) to automatically stop it when the standard deviation (SD) of the last four calculated buffer’s Boyce Index is small. Thus, the user has four arguments (i.e. BI\_part, BI\_tot, SD\_BI\_part and SD\_BI\_tot) to pass to minba in order to define how to proceed. The first two arguments accept two possibilities: NULL (default), which deactivates the condition, or a number below 1 (it makes no sense a BI higher), which establish the minimum limit. Similarly, SD\_BI\_part and SD\_BI\_tot accept NULL (default) to deactivate the condition, or a number to establish the minimum SD. After checking the results of the case studies presented in this document (see below), a recommended minimum SD could be 0.006. Therefore, there are several combinations to choose from. For instance, if the four arguments are NULL (default) all the buffers are modelled; alternatively, if both BI\_par and BI\_tot are a number and also SD\_BI\_part and SD\_BI\_tot, the process stops when one of the BI or the SD\_BI is first reached; and so forth. ##Outputs The function minba, at the end of the modelling process, outputs different information in form of tables and charts to let the user check the optimal buffer. It writes out three tables in csv files: selfinfo\_mod\_, info\_mod\_ and info\_mod\_means\_ (all followed by the name of the species). The first two tables are merely informative about how the modelling process has been developed and the results of each model. Whereas info\_mod\_means\_ shows the means of the n models run for each buffer. See Table 1 as an example of info\_mod\_means\_. It contains the Boyce Index Partial, the Boyce Index Total and the execution time. Additionally, it also has columns with rankings of the buffer derived from these three metrics, plus two more ranking columns: rankFinalNoTime and rankFinalNoTime, which ranks for the best buffer with and without taking into account the execution time, respectively. Finally, minva draws scatterplots, smoothed by fitting a Loess regression curve, of the two BI to show the evolution of them with the increase of the buffer diameter in kilometres. It also plots the execution time by fitting a linear regression model.

## Future developments

# Implementation (Case Studies)

To assess the hypothesis formulated at the beginning of this paper about the existence of an optimal background area, I developed two different case studies. For each case I selected several common plant species of different typology (i.e. herbaceous, shrubs, broad-leaved trees, conifers). The function minba, by default, defines 10 buffers, 3 model repetitions per buffer and lets the process produce models for all of them. Running it so, it can be appreciated the evolution of the metrics along the different buffers. MaxEnt was run with the default parameters, except for the number of background points. The intention of that was limiting interferences in the results as much as possible for all the species. I used the 19 climatic variables available from WorldClim at different resolutions for each case of study. Equally, I downloaded the occurrences of the species from public repositories by means of the PreSPickR package. All the code used can be found in <https://github.com/xavi-rp/MinBA>. ##Case 1: Entire distribution I modelled 25 species native from Eurasia and North of Africa (see list in Supplementary materials Table S1). The presences were downloaded from GBIF (). I discarded those out of the native areas as they were introduced and this was out of the scope of this case study. The results (Table ??, Figure ??) showed that the best models for most of the species were those fitted with only part of their distribution, both taking into account the execution time (96%) and not doing it (72%) ##Case 2: Partial distribution on islands

# Conclusions

The package MinBAR has been developed, so far, to work with MaxEnt. It includes the Boyce Index as a main evaluator of the models predictive performance. In coming versions, however, it would be interesting to include other threshold-dependent evaluators based on sensitivity and specificity, as well as the option to pass arguments to the maxent() function, or to decide the centre from where to start delimiting buffers for modelling. In addition, the inclusion of an index that would take into account at the same time the accuracy in the training area and after transferring to further areas, such as the one described by Duque-Lazo et al. (2016), might also be quite useful for the users. Furthermore, the implementation of other algorithms and modelling techniques would be highly convenient. In conclusion, delimiting the background area can strongly affect the results of SDMs (Acevedo et al, 2012). However, it is generally difficult to define it accurately because the biological information required is often incomplete or unavailable (Anderson & Raza, 2010; Barve et al., 2011). In both case studies presented in this document it can be seen how the model including the presences from all the species distribution does not always perform the best. Therefore, here I present a methodology to help modellers to objectively define an optimal solution.

## Tables

|  |  |  |  |
| --- | --- | --- | --- |
| **fwe** | **WEFwfe** | **wfwfwf** | **fWEFqw** |
| 3451 | 653 | 2451 | 3241234 |
| 15 | 151 | 6565 | 12542 |

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

# References

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. “Maximum Entropy Modeling of Species Geographic Distributions.” Journal Article. *Ecological Modelling* 190 (3-4): 231–59. doi:[10.1016/j.ecolmodel.2005.03.026](https://doi.org/10.1016/j.ecolmodel.2005.03.026).

Thuiller, Wilfried, Bruno Lafourcade, Robin Engler, and Miguel B. Araujo. 2009. “BIOMOD - a Platform for Ensemble Forecasting of Species Distributions.” Journal Article. *Ecography* 32 (3): 369–73. doi:[10.1111/j.1600-0587.2008.05742.x](https://doi.org/10.1111/j.1600-0587.2008.05742.x).