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

Xavier Rotllan-Puig[1,*] & Anna Traveset[2]

[1]ASTER Projects. Barri Reboll, 9, 1r. 08694 - Guardiola de Berguedà (Barcelona). Spain

[2]Terrestrial Ecology Laboratory. Global Change Research Group. Institut Mediterrani d'Estudis Avançats (CSIC-UIB). C/ Miquel Marqués, 21. 07190 - Esporles (Mallorca - Illes Balears). Spain

*Correspondence: Xavier Rotllan-Puig xavier.rotllan.puig@aster-projects.cat

Abstract

1. One of the crucial choices when modelling species distributions using pseudo-absences approaches is the delineation of the background area to fit the model. We hypothesise that there is a minimum background area around the centre of the species distribution that characterizes well enough the range of environmental conditions needed by the species to survive. Thus, fitting the model within this area should be the optimal solution in terms of both quality of the model and execution time.
2. MinBAR is an R package that calculates the optimal background area. The version 1.0.0 is implemented for MaxEnt and uses Boyce Index as a metric to assess models performance.
3. Two case studies are presented to assess the hypothesis and to illustrate the package.
4. Partial models trained with part of the species distribution often perform better than those fitted on the entire extension. MinBAR is a versatile tool that helps modellers to objectively define the optimal solution.

Introduction

Species distribution modelling (SDM) has become an essential tool in the fields of ecology and biodiversity conservation. Its popularity, among other reasons, is due to the ease of use of software such as MaxEnt (S. J. Phillips, Anderson, & Schapire, 2006) or BIOMOD (Thuiller, Lafourcade, Engler, & Araujo, 2009), but also because of the development of the R-programming community and the free availability of biodiversity data in public repositories like GBIF (https://www.gbif.org/). However, such data is often limited to species presences only and with a lack of occurrences in poorly sampled areas. These facts 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 crucial 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). Defining its extent, however, remains a challenge. Elith et al. (2011), for instance, argued that it has to be defined by the ecologist and limited by geographic boundaries or by how far the species can disperse. More recently, other authors have considered the interactions with other species or the sampling biases in the data set as constraints (Jarnevich et al., 2017). Yet, in many situations it is difficult to accurately define a background area, either owing to limited knowledge of the species biology or to the lack of available data (R. P. Anderson & Raza, 2010; N. Barve et al., 2011). In addition, studies are usually performed at a country or regional level and, then, the background area is constrained to an artificial or political boundary despite the species distribution might be wider (El-Gabbas & Dormann, 2018). Finally, another limitation may appear when the extent of the species is so large that it makes computations to fit the model and generate predictions highly resource-demanding and time-consuming. These limitations are particularly important when the study encompasses a high number of species with a large geographical range. Any of these situations usually lead to fit partial models, which might or might not imply a reduction of model performance (El-Gabbas & Dormann, 2018).

In this work, we hypothesize that there is a minimum background area around the centre of the species distribution (minimum buffer) that characterizes 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 execution time.

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 whether 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, Lobo, & Lucia Hernandez-Manrique, 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 about the best way to evaluate the performance of SDMs. One of the most widely used metrics is the AUC or area under the receiver operating characteristic (ROC) curve, although it has received several critiques because of its misuse (J. M. Lobo, Jimenez-Valverde, & Real, 2008). In particular, for the purpose of MinBAR, AUC is not the best choice because its scores are highly influenced by the defined background area, then it is only useful for assessing the performance of different models with exactly the same extent. For that reason, this package uses the Boyce Index

(Hirzel, Le Lay, Helfer, Randin, & Guisan, 2006), implemented in the R package *ecospat* (Di Cola et al., 2017). 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 (Di Cola et al., 2017). 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) and Di Cola et al. (2017) for further details on how BI is calculated as well as its strengths and weaknesses.

In order to evaluate the predictive performance of the models, this package includes two different metrics. On the one hand, Boyce Index Partial (BI_part) evaluates the accuracy of predictions within the buffer, or what is the same, in the training area. On the other hand, Boyce Index Total (BI_tot) assesses 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.

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 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, Gerstner, Kass, & Anderson, 2018). It uses 70% of the presences to calibrate the model and 30% for evaluation, 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). For this assessment, it uses 30% of all presences excluding those used to calibrate the model.

At this point, 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 on 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 aim is to project the model in time or space, the focus should fall on the Boyce Index Total. In turn, both approaches have two possibilities: either (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. BI_part and BI_tot accept two possibilities: either *NULL* (default), which deactivates the condition, or a number below 1 (it makes no sense a higher BI), which establishes 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 all four arguments are *NULL* (default), all buffers are modelled; alternatively, if both BI_par and BI_tot are defined as a number, and so are SD_BI_part and SD_BI_tot, the process stops when the first of them is reached. Any combination of them is allowed.

Outputs

At the end of the modelling process, *minba* outputs different information in the form of tables and charts to let the user know 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 S1 in Supplementary Material 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 *rankFinalWithTime*, which rank 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.

Implementation (Case Studies)

To test the hypothesis on the existence of an optimal background area, we used two different case studies. For each one we selected several common plant species of different typology (i.e. herbaceous, shrubs, broad-leaved trees, conifers). The function *minba*, by default, defines 10 buffers, with 3 model replicates per buffer, and lets the process produce models for all of them. By doing so, one can appreciate 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 to limit interferences in the results as much as possible for all the species. We used 19 climatic variables available from WorldClim at different resolutions for each case of study. Equally, we downloaded the occurrences of the species from public repositories by means of the PreSPickR package (Rotllan-Puig, 2018).

All the R scripts used in these case studies, including the code for the generation of the manuscript, can be found in https://github.com/xavi-rp/MinBA. In turn, the source code of the MinBAR package can be downloaded from https://github.com/xavi-rp/MinBAR.

Case 1: Entire distribution

We modelled 25 species native from Eurasia and North of Africa (see the list in Supplementary Material Table S2.1). The presences were downloaded from GBIF. We discarded those occurrences out of the native areas as they were introduced, and this was out of the scope of this case study.

The output graphs produced by *minba* for instance for *Fraxinus excelsior* and *Linaria alpina* can be seen in Figure 1 and Figure 2, respectively. Both BI_tot and BI_part for the two species did not notably improve when increasing the buffers after the second one. A similar pattern was seen for almost all the species studied (see all plots in Supplementary Material S3). Actually, the results (Table 1, Figure 3) 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 so (72%).

Case 2: Partial distribution on islands

We modelled the distribution of 10 species on the Balearic Islands (Western Mediterranean), although their native distribution also includes other continental areas (see the list in Supplementary Material Table S2.2). The occurrences were downloaded from Bioatles (http://bioatles.caib.es).

As two examples, the output graphs produced by *minba* for *Arbutus unedo* and *Asphodelus aestivus* can be seen in Figure 4 and Figure 5, respectively. Both BI_tot and BI_part for the two species did not improve very much when increasing the buffers after the first half, and a similar pattern was seen for almost all the studied species (see all plots in Supplementary Material S4). The results (Table 2, Figure 6) also showed that the best models for most of the species were those fitted with only part of their distribution, specially taking into account the execution time (90%) but also not doing so (70%).

Conclusions

The package MinBAR has been developed, so far, to work with MaxEnt. It includes the Boyce Index as the main evaluator of the models predictive performance. In coming

195 versions, however, it would be interesting to include other threshold-dependent
196 evaluators based on sensitivity and specificity, as well as the option to pass arguments to
197 the *maxent* function, or to decide the centre from where to start delimiting buffers for
198 modelling. In addition, the inclusion of an index that would take into account at the same
199 time the accuracy in the training area and after transferring to further areas, such as the
200 one described by Duque-Lazo et al. (2016), might also be quite useful for the users.
201 Furthermore, the implementation of other algorithms and modelling techniques would be
202 highly convenient.

203 In short, delimiting the background area can strongly affect the results of SDMs
204 (Acevedo, Jimenez-Valverde, Lobo, & Real, 2012). Both case studies presented here
205 show that the model including the presences from all the species distribution does not
206 always perform the best. Therefore, the tool developed here will help modellers to
207 objectively define an optimal solution.

208

Tables

Table 1: Best and second best buffer with and without taking into account execution time, for each species in Case Study 1

Species	Best Buffer No-Time	2nd Buffer No-Time	Best Buffer With-Time	2nd Buffer With-Time
pin_syl	10	9	1	10
que_ile	10	7	7	10
fag_syl	9	10	9	10
fra_exc	4	3	3	4
que_pet	9	8	9	8
que_rob	10	9	2	5
que_pyr	9	10	9	5
que_sub	9	7	8	9
abi_alb	9	8	3	4
ace_pla	10	8	1	2
aln_glu	8	10	8	7
jun_oxy	8	9	8	9
arb_une	8	10	8	10
cra_mon	10	7	4	10
pru_spi	6	8	4	6
bux_sem	8	9	8	4
cot_tom	10	7	5	10
vio_mir	7	9	7	9
dip_eru	10	9	10	5
cen_alb	9	4	4	5
ger_luc	9	10	2	9

lin_alp	8	6	8	5
pis_ter	9	10	5	9
leo_com	8	10	6	8
lot_edu	9	10	4	5

212

213

214 *Table 2: Best and second best buffer with and without taking into account execution time,*
215 *for each species in Case Study 2*

Species	Best Buffer No-Time	2nd Buffer No-Time	Best Buffer With-Time	2nd Buffer With-Time
arb_une	8	3	8	3
asp_aes	10	8	1	2
cha_hum	5	7	5	2
eph_fra	8	10	6	8
hel_sto	9	6	6	9
jun_oxy	4	3	3	4
pis_len	10	9	10	3
que_coc	6	7	6	5
rha_ala	10	6	2	6
vib_tin	7	5	5	7

216

217

Figures

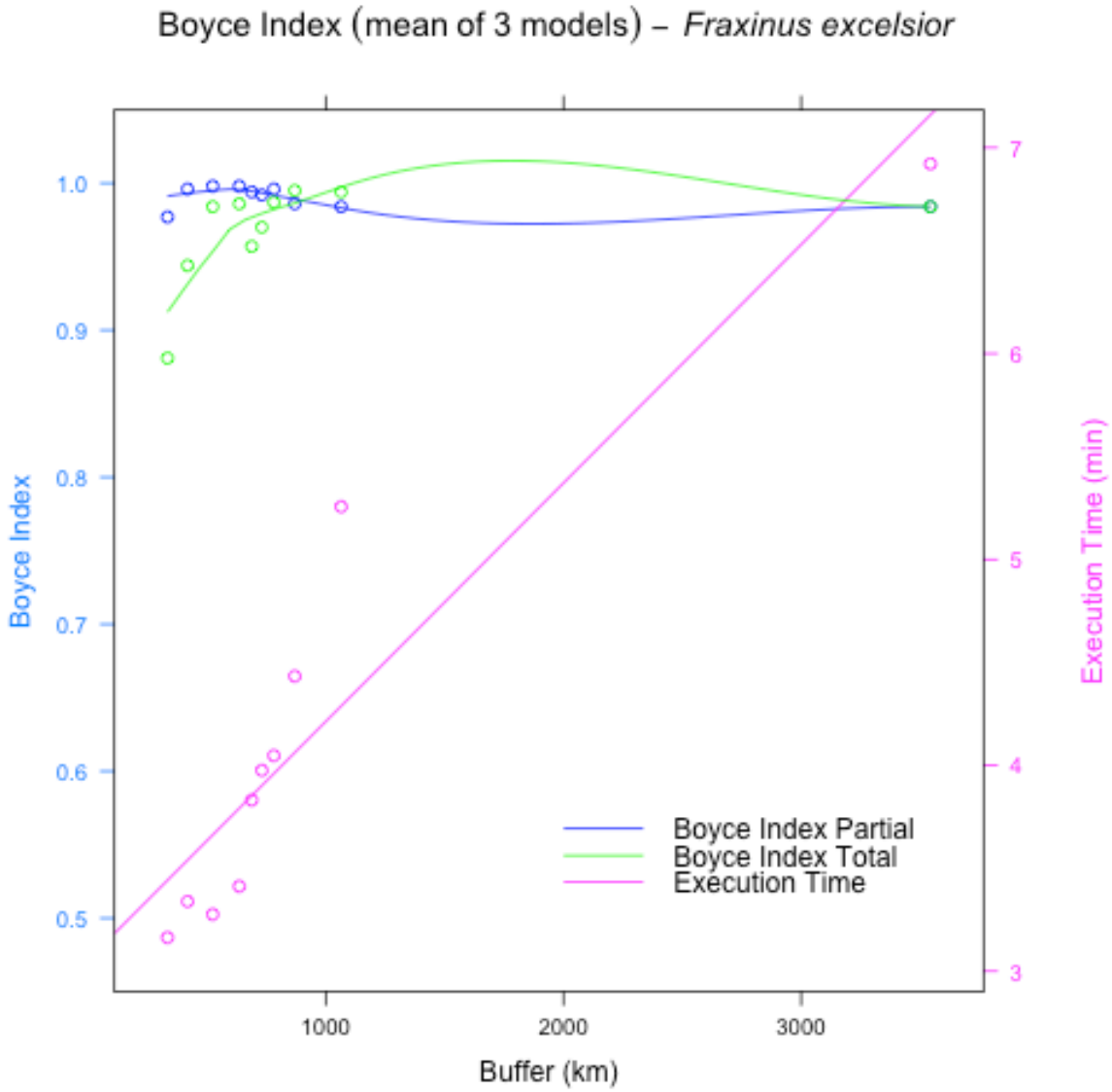
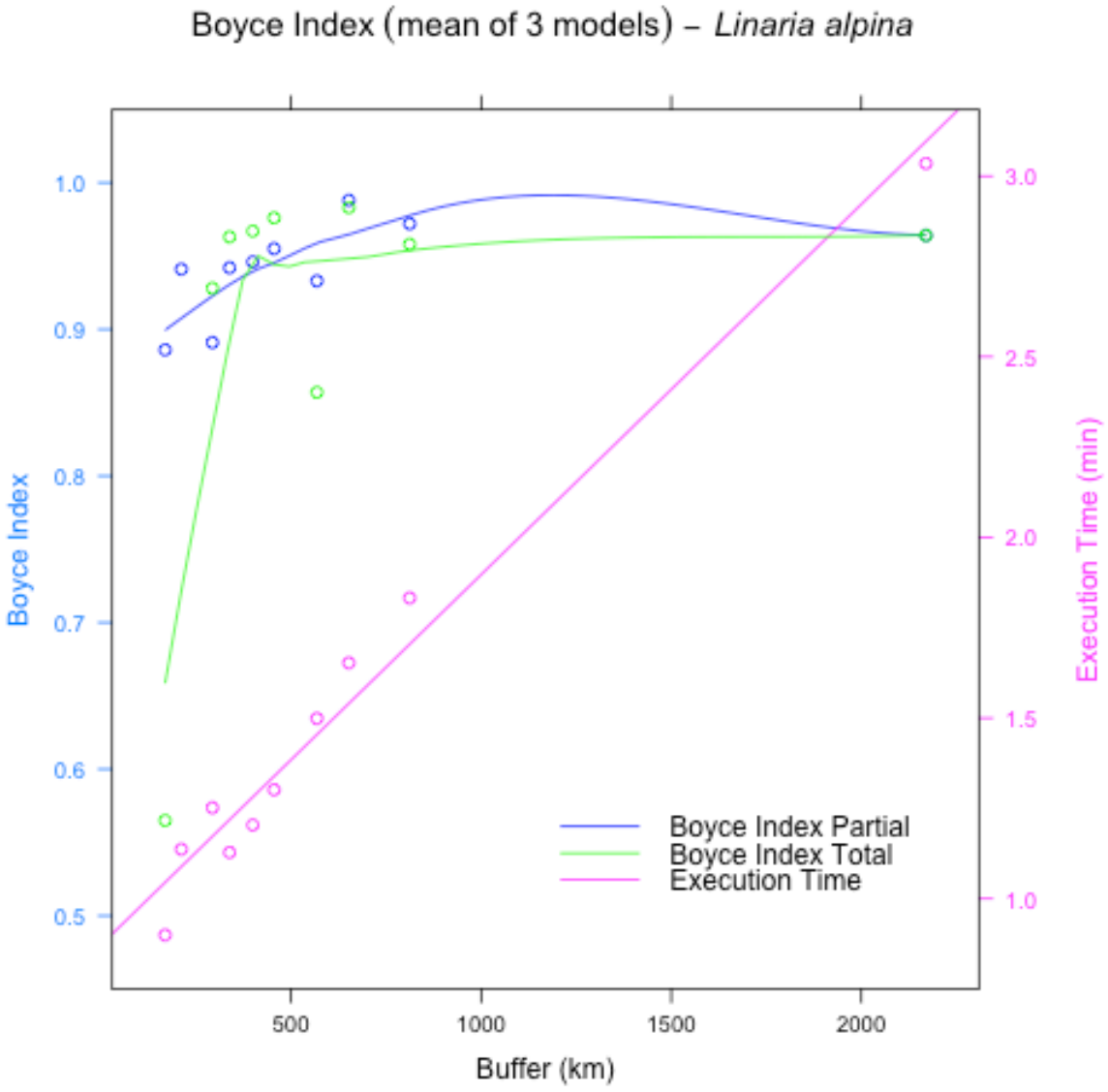
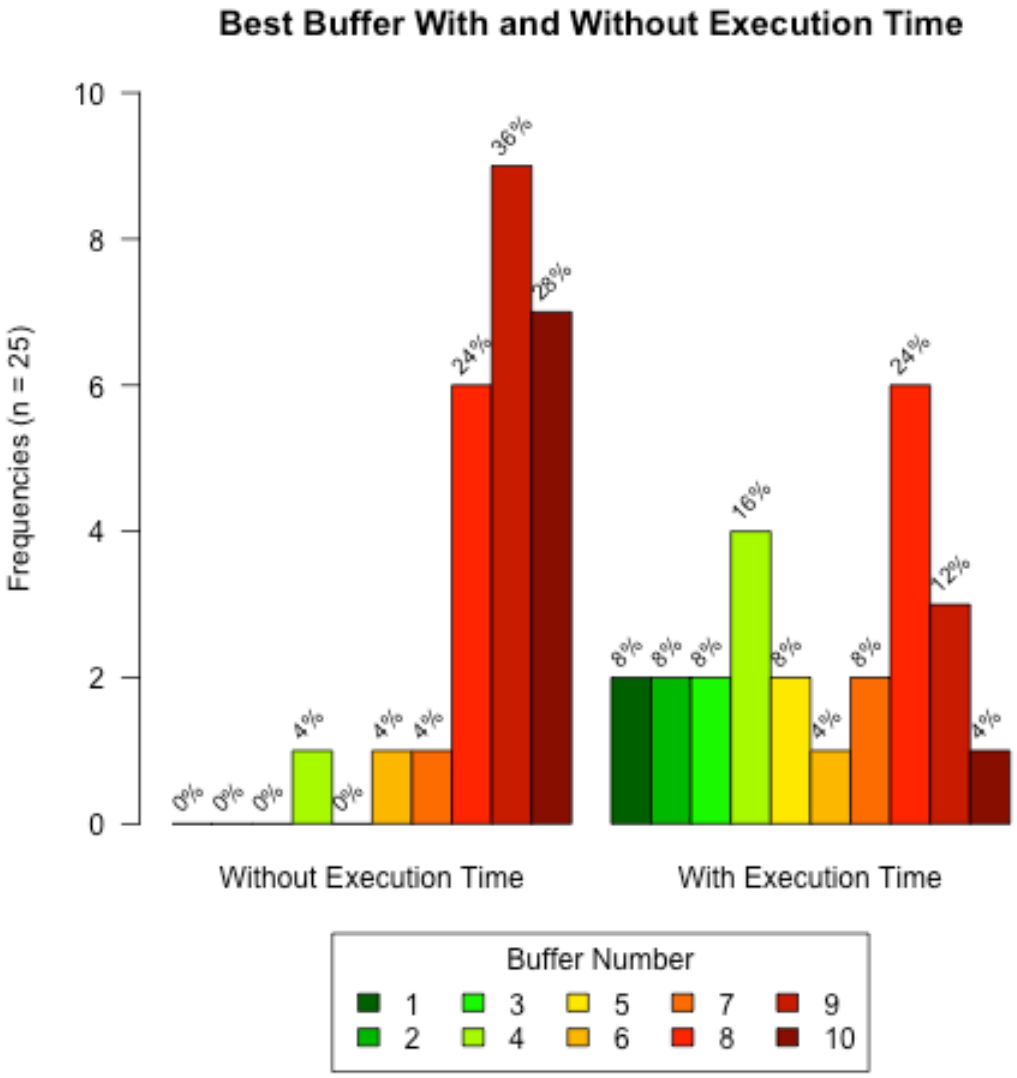


Figure 1: Evolution of Boyce Index Total (green) and Parcial (blue) and the execution time in minutes (pink) for Fraxinus excelsior



222

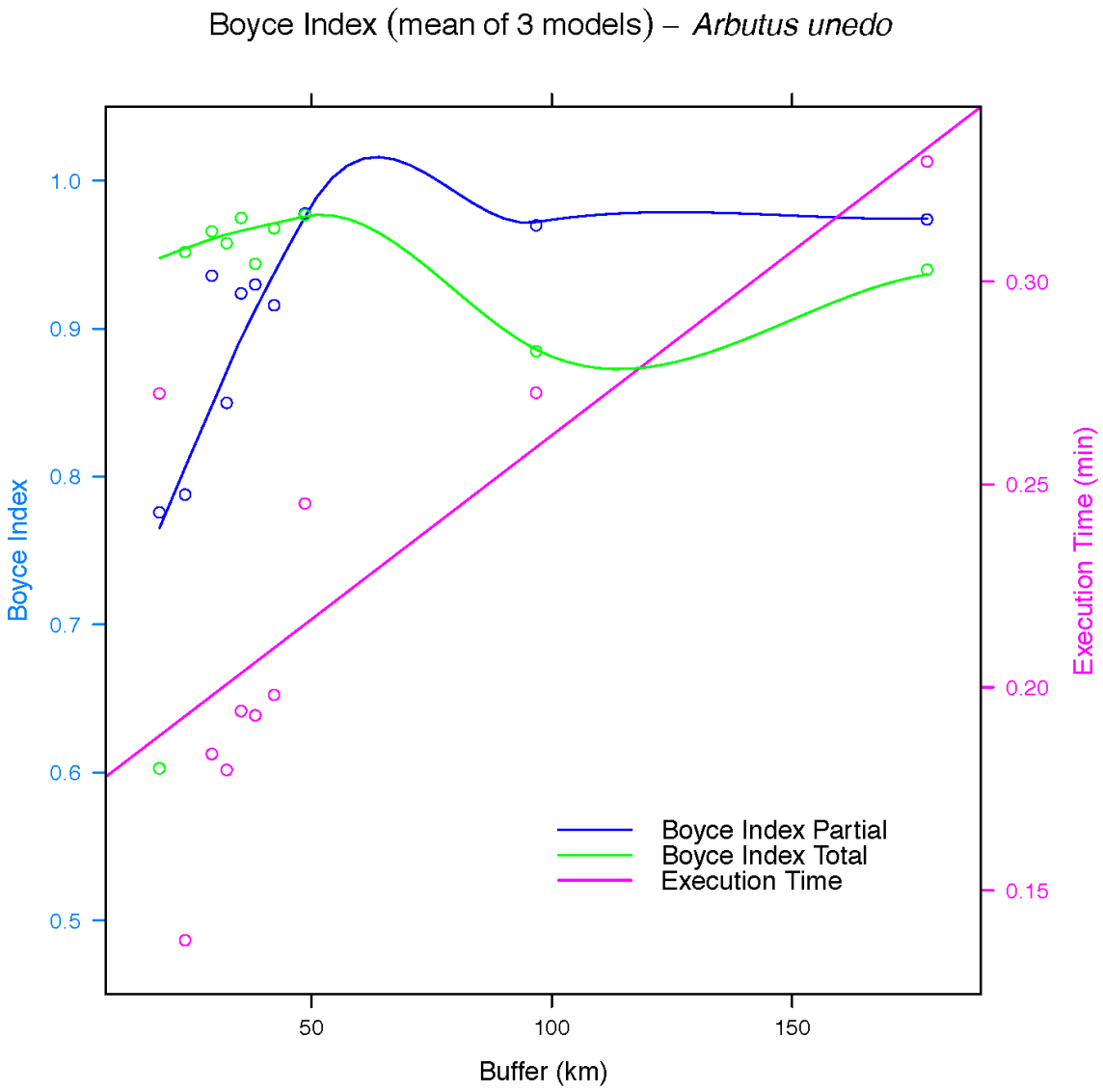
223 *Figure 2: Evolution of Boyce Index Total (green) and Parcial (blue) and the execution*
224 *time in minutes (pink) for Linaria alpina*



225

226 *Figure 3: Frequencies of best buffer with and without taking into account execution time*

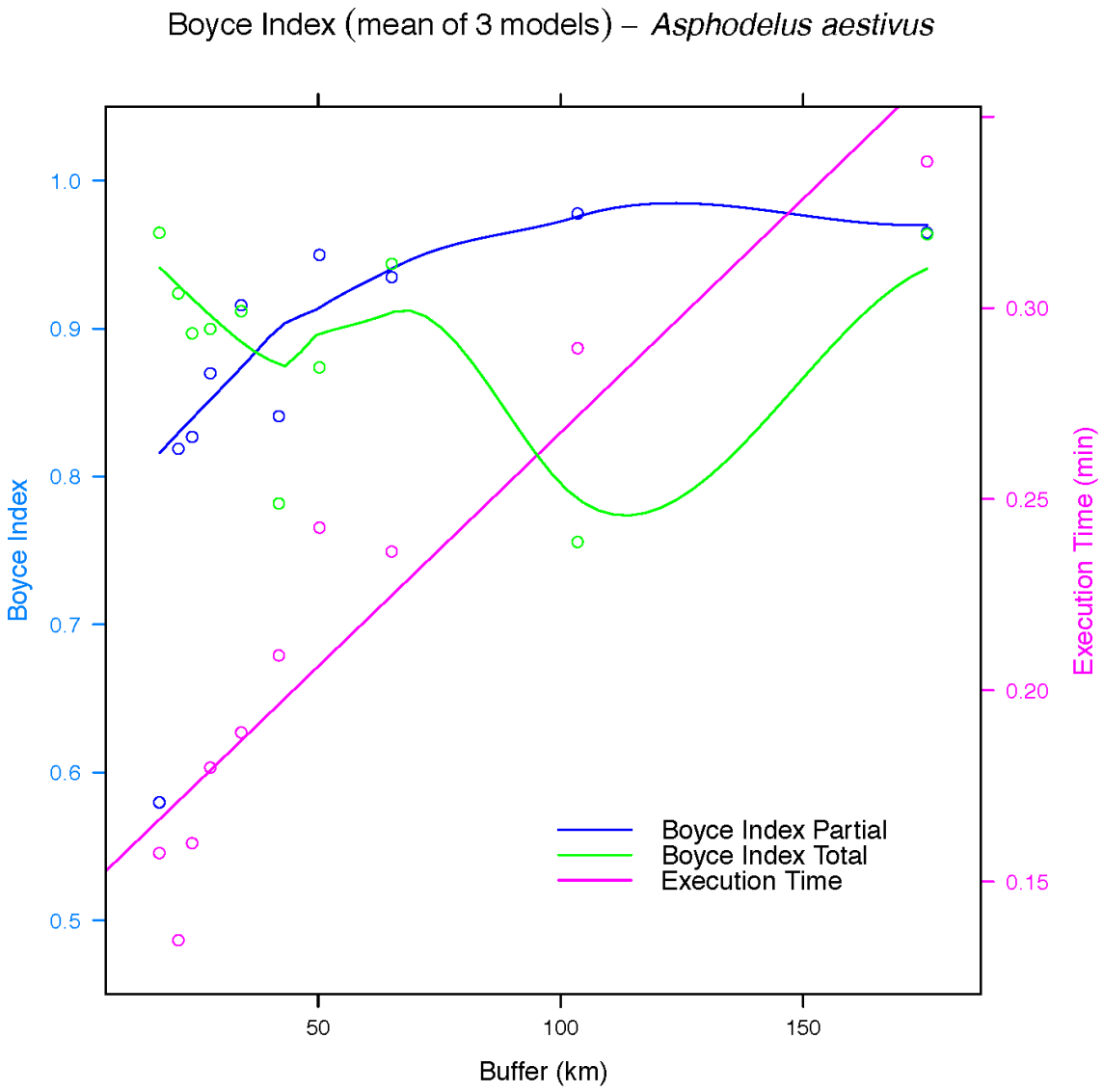
227 *in Case Study 1*



228

229 *Figure 4: Evolution of Boyce Index Total (green) and Parcial (blue) and the execution*

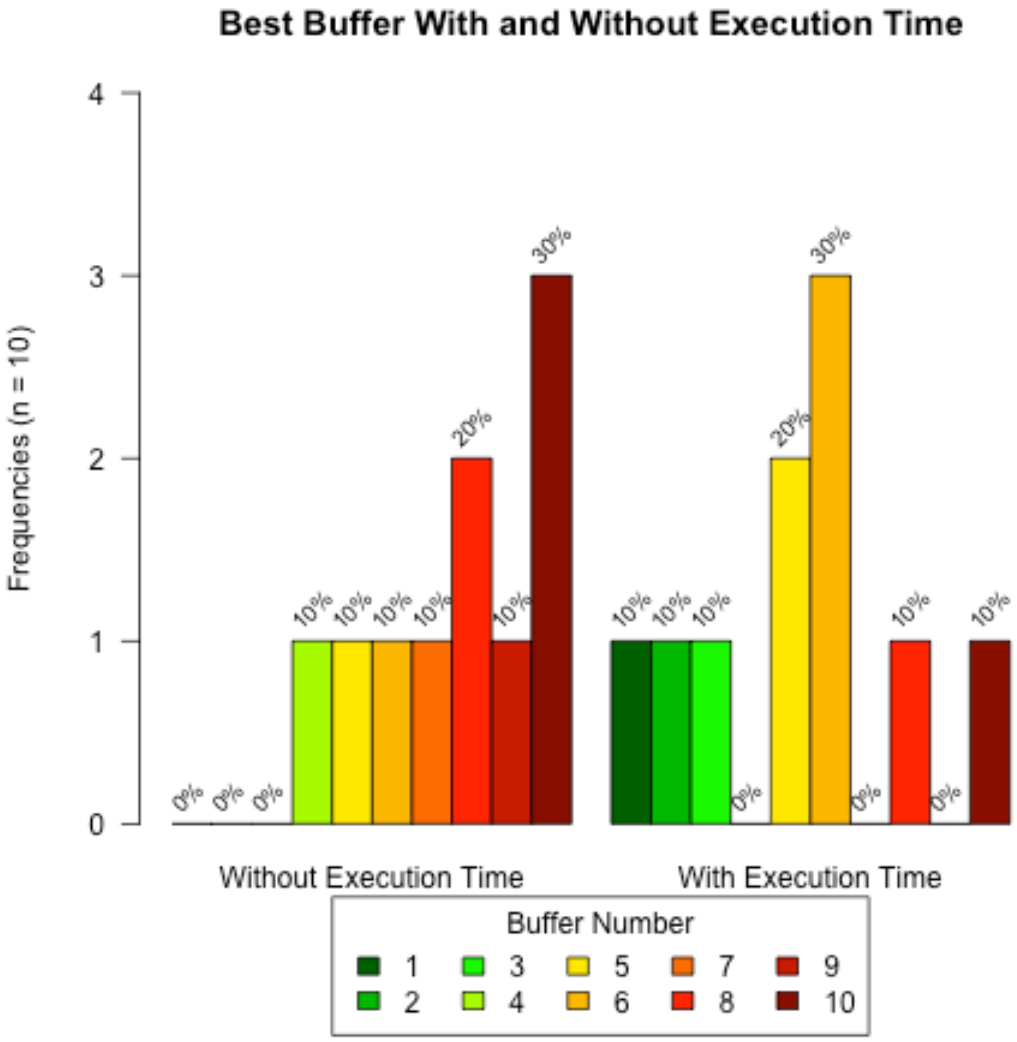
230 *time in minutes (pink) for Arbutus unedo*



231

232 *Figure 5: Evolution of Boyce Index Total (green) and Parcial (blue) and the execution*

233 *time in minutes (pink) for Asphodelus aestivus*



234

235 *Figure 6: Frequencies of best buffer with and without taking into account execution time*

236 *in Case Study 2*

237

References

Acevedo, P., Jimenez-Valverde, A., Lobo, J. M., & Real, R. (2012). Delimiting the geographical background in species distribution modelling. *Journal of Biogeography*, *39*(8), 1383–1390. Journal Article. doi:10.1111/j.1365-2699.2012.02713.x

Anderson, R. P., & Raza, A. (2010). The effect of the extent of the study region on gis models of species geographic distributions and estimates of niche evolution: Preliminary tests with montane rodents (genus nephelomys) in venezuela. *Journal of Biogeography*, *37*(7), 1378–1393. Journal Article. doi:10.1111/j.1365-2699.2010.02290.x

Barve, N., Barve, V., Jimenez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., … Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, *222*(11), 1810–1819. Journal Article. doi:10.1016/j.ecolmodel.2011.02.011

Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F. T., D'Amen, M., Randin, C., … Guisan, A. (2017). Ecospat: An r package to support spatial analyses and modeling of species niches and distributions. *Ecography*, *40*(6), 774–787. Journal Article. doi:10.1111/ecog.02671

Duque-Lazo, J., Gils, H. van, Groen, T. A., & Navarro-Cerrillo, R. M. (2016). Transferability of species distribution models: The case of phytophthora cinnamomi in southwest spain and southwest australia. *Ecological Modelling*, *320*, 62–70. Journal Article. doi:10.1016/j.ecolmodel.2015.09.019

El-Gabbas, A., & Dormann, C. F. (2018). Wrong, but useful: Regional species distribution models may not be improved by range-wide data under biased sampling. *Ecology and Evolution*, *8*(4), 2196–2206. Journal Article. doi:10.1002/ece3.3834

264 Elith, J., Phillips, S. J., Hastie, T., Dudik, M., Chee, Y. E., & Yates, C. J. (2011). A
265 statistical explanation of maxent for ecologists. *Diversity and Distributions*, *17*(1),
266 43–57. Journal Article. doi:10.1111/j.1472-4642.2010.00725.x

267 Guevara, L., Gerstner, B. E., Kass, J. M., & Anderson, R. P. (2018). Toward ecologically
268 realistic predictions of species distributions: A cross-time example from tropical
269 montane cloud forests. *Global Change Biology*, *24*(4), 1511–1522. Journal Article.
270 doi:10.1111/gcb.13992

271 Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the
272 ability of habitat suitability models to predict species presences. *Ecological*
273 *Modelling*, *199*(2), 142–152. Journal Article. doi:10.1016/j.ecolmodel.2006.05.017

274 Jarnevich, C. S., Talbert, M., Morisette, J., Aldridge, C., Brown, C. S., Kumar, S., …
275 Holcombe, T. (2017). Minimizing effects of methodological decisions on
276 interpretation and prediction in species distribution studies: An example with
277 background selection. *Ecological Modelling*, *363*, 48–56. Journal Article.
278 doi:10.1016/j.ecolmodel.2017.08.017

279 Lobo, J. M., Jimenez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of
280 the performance of predictive distribution models. *Global Ecology and*
281 *Biogeography*, *17*(2), 145–151. Journal Article. doi:10.1111/j.1466-
282 8238.2007.00358.x

283 Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of
284 species geographic distributions. *Ecological Modelling*, *190*(3-4), 231–259. Journal
285 Article. doi:10.1016/j.ecolmodel.2005.03.026

286 Raes, N. (2012). Partial versus full species distribution models. *Natureza & Conservacao*,
287 *10*(2), 127–138. Journal Article. doi:10.4322/natcon.2012.020

288 Rotllan-Puig, X. (2018). PreSPickR: Downloading species presences (occurrences) from
289 public repositories. Computer Program. doi:10.13140/RG.2.2.10574.97607/1

290 Sanchez-Fernandez, D., Lobo, J. M., & Lucia Hernandez-Manrique, O. (2011). Species
291 distribution models that do not incorporate global data misrepresent potential

distributions: A case study using iberian diving beetles. *Diversity and Distributions*, *17*(1), 163–171. Journal Article. doi:10.1111/j.1472-4642.2010.00716.x

Thuiller, W., Lafourcade, B., Engler, R., & Araujo, M. B. (2009). BIOMOD - a platform for ensemble forecasting of species distributions. *Ecography*, *32*(3), 369–373. Journal Article. doi:10.1111/j.1600-0587.2008.05742.x