Increasing reality of species distribution models of consumers by including its food resources

Gabriel Preuss¹, André Andrian Padial¹

¹ Laboratório de Análise e Síntese em Biodiversidade, Departamento de Botânica. Programa de Pós-graduação em Ecologia e Conservação, Universidade Federal do Paraná, 81531-980, Curitiba-PR, Brazil

Corresponding author: André Andrian Padial (aapadial@gmail.com)

Academic editor: A.M. Leal-Zanchet | Received 23 February 2021 | Accepted 7 July 2021 | Published 6 August 2021

Citation: Preuss G, Padial AA (2021) Increasing reality of species distribution models of consumers by including its food resources. Neotropical Biology and Conservation 16(3): 411–425. https://doi.org/10.3897/neotropical.16.e64892

Abstract
Species distribution models are not usually calibrated with biotic predictors. Our study question is: does the use of biotic predictors matter in predicting species distribution? We aim to assess the importance of biotic predictors in the output of distribution models of the Brazilian squirrel (Sciurus aestuans) throughout South America based on fruits of Syagrus romanzoffiana – the most consumed food resource. We hypothesized that the distribution model of S. aestuans using its main food resource as a biotic predictor will be more accurate in comparison with the output of the model without the biotic predictor. We built three different distribution models: (i) distribution of S. romanzoffiana; (ii) distribution of S. aestuans without biotic predictor; and (iii) distribution of S. aestuans with biotic predictor. We evaluated performance scores, number of presence pixels and concordance between suitability maps. We found that performance scores may not vary between models with different predictors, but the output map changed significantly. We also found that models with biotic predictors seem to vary less in presence pixels. Furthermore, the main variable in the distribution model was the biotic variable. We conclude that the knowledge of a species’ biology and ecology can make better predictions of species distribution models mainly by avoiding commission errors.

Keywords
biotic interaction, biotic predictor, diet specialization, Eltonian Noise Hypothesis, hutchinsonian niche, realized niche, variables importance
Introduction

One of the most important characteristics of species distribution modelling (SDM) is the capacity to generate accurate predictions of past, present and future species distributions (Guisan and Thuiller 2005; Barry and Elith 2006). Assuming that models have a minimally reliable predictive capacity, conservation planning can be elaborated with support of these findings, especially in a changing world (Margules and Usher 1981; Haila and Kouki 1994). Historically, SDMs are calibrated only with abiotic variables as predictor, assuming that biotic interactions do not influence species range pattern, have influence only at small spatial scales, or are indirectly represented by abiotic variables (Huntley et al. 1995; Bakkenes et al. 2002; Pearson and Dawson 2003; Heikkinen et al. 2007). In fact, the Eltonian Noise Hypothesis (ENH) assumes that biotic interactions do not influence species distribution (Soberón and Nakamura 2009). Modeling exercises using abiotic features thus consider species distribution inside the Grinnellian niche, which comprises the scenopoetic variables (Grinnell 1917). Even so, some studies suggest the important role of biotic interactions as a predictor in spatial distribution even on a large scale (Meier et al. 2011; Araújo et al. 2014). By including variables representing biotic interactions, modeling exercises can thus assume a Hutchinsonian view of the niche: predation, competition, facilitation and mutualism together with environmental filtering compose the realized niche (Hutchinson 1957).

One difficulty to insert biotic interactions in SDMs is the lack of suitable data to represent biotic interaction on a large-scale (Araújo and Luoto 2007; Zimmermann et al. 2010; Guillera-Arroita et al. 2015), especially when species interactions are poorly known or its interactors are numerous (Meier et al. 2011). Since occurrence data are increasing in open access databases, obtaining large-scale data has become easier and cheaper (Fielding and Bell 1997; Wisz et al. 2013). In this way, studies that use a well-distributed species with sufficient sampling effort and its biotic interactions may fill the information gap in biotic interactions’ role in SDMs (Meier and Dikow 2004).

Indeed, robust models require two principal components: sufficient sampling effort in occurrence data for the target biological group and a set of meaningful predictor variables with low or no collinearity (Lobo et al. 2002; Synes and Osborne 2011). There are different ways to assess the performance of models, and the most common are model features such as prevalence, Kappa, True Skill Statistic and AUC scores (Allouche et al. 2006; Liu et al. 2011). Performance scores aims to measure the accuracy of species distribution models through discrimination capacity and reliability, and there are different applications for each type of performance score (Pearce and Ferrier 2000; Liu et al. 2011). Those scores can be divided in three: (i) Threshold-independent scores are those which can be assessed with continuous values (suitability models); (ii) Threshold-dependent are scores only assessed with binary data (presence-absence models); and (iii) overall ones are those applied to both cases (Liu et al. 2011).
The assessment of a model’s performance may provide the possibility to compare models with alternative algorithms and predictors and evaluate how different variables affect the model’s predictive performance (Kadmon et al. 2003; Loiselle et al. 2003; Segurado and Araujo 2004; Pearson et al. 2007). Furthermore, each study area may have its particular source of variation. In North America, for example, the variation in final results of SDM exclusively by algorithm may be high, while in South America the variation reaches lower oscillations (Diniz-Filho et al. 2009), suggesting that for South America, the studies might focus on better predictor variables instead of considering algorithm variation.

Keeping the focus on biotic predictors, a previous study that incorporated competition in predictor variables showed that biotic interactions influence directly the predicted range of tree species (Meier et al. 2011). These findings suggest that one species may be determinant in another species distribution. It is known that the addition of biotic interactions usually improves the predictive performance of SDMs (Araújo and Luoto 2007). Furthermore, the inclusion of biotic interactions in SDMs has restricted the predicted distribution of a species in relation to another (Schweiger et al. 2012), reducing the possibility of commission error (false positives). However, the inclusion of biotic interactions in SDMs has been concentrated in competition (Guisan and Thuiller 2005; Meier et al. 2011). Other kinds of biotic interactions such as predation, herbivory and facilitation still need to be tested (Guisan and Thuiller 2005; Baselga and Araújo 2009).

One way to guarantee that one kind of interaction really occurs in nature, is the dependency of the species with that interaction (Meier et al. 2011), a feature that usually occurs with food specialist species. In this way, the relation resource-consumer may be a good way to understand the role of biotic interactions in species distributions; furthermore, this relation was recorded as necessary additional information for the knowledge of the role of biotic interactions in SDMs (Anderson 2017). Herein we used a food specialist species of squirrel to predict its actual distribution once its distribution is poorly known in Brazil, but its foraging behavior is marked by the strong interaction with a palm tree fruit (Alvarenga and Talamoni 2006).

*Sciurus aestuans* Linnaeus, 1776 is a squirrel species widely distributed along South America (Amori et al. 2013). In spite of the generalism in squirrels’ diet, in the Brazilian Atlantic Forest, the foraging behavior of *S. aestuans* is characterized by the use of the fruit of *Syagrus romanzoffiana* (Cham.) Glassman (Paschoal and Galetti 1995; Bordignon and Monteiro-Filho 2000; Alvarenga and Talamoni 2006). In a study developed in the Brazilian Atlantic Forest, *S. romanzoffiana* presented up to 70% of *S. aestuans* diet (Alvarenga and Talamoni 2006). The same study shows that the palm tree is essential for the reproduction of squirrels, where lactating females spend their time feeding and not searching for food, suggesting elevated habitat fidelity near the palm tree (Alvarenga and Talamoni 2006). Furthermore, the caching habit of squirrels allows them to consume *S. romanzoffiana* fruits throughout the year, showing a strong interaction between the species (personal observations; Paschoal and Galetti 1995; Bordignon and Monteiro-Filho 2000; Alvarenga and
Talamoni 2006). Considering all this, we believe *S. aestuans* and *S. romanzoffiana* to have a close relationship in relation to ecology and spatial distribution.

Assuming that the distribution pattern of *S. aestuans* is regulated by a bottom-up control, we hypothesized that: I – The distribution model of *S. aestuans* with *S. romanzoffiana* as biotic predictor will be more accurate in comparison with the output of the model without biotic predictor, and II – The Eltonian Noise Hypothesis is not determinant in our distribution models once biotic interaction will affect the distribution area.

**Methods**

**Data preparation**

A recent taxonomic review suggests the division in Sciurus genus. However, in GBIF database the occurrence points are available for *S. aestuans*. To obtain the occurrence records of *S. aestuans* and *S. romanzoffiana*, we used personal data and the Global Biodiversity Information Facility (GBIF) database (GBIF 2020), with data recorded from years 1979 to 2013. We selected the data for South America, which resulted in 149 occurrence points of *S. romanzoffiana* and 132 of *S. aestuans* before data cleaning (presence records of *S. romanzoffiana* and *S. aestuans* can be seen in Suppl. materials 1, 2, respectively).

The environmental variables of each model were obtained from Climatologies at High Resolution for the Earth’s Land Surface Areas (CHELSA) database (Karger et al. 2017), collecting all the 19 bioclimatic variables under 30 sec arc for the same period of occurrence records (1979–2013). Since climatic data is available for the whole world, but our analysis was in South America, we cropped the raster files to our study area. Hence, for data cleaning we excluded points inside the same pixel, and those out of the confirmed distribution of the species following a Brazilian checklist established by specialists by *S. romanzoffiana* (Soares, 2020) and out of the boundaries of South America for *S. aestuans* where there is no information about its distribution. We were left with 136 occurrence points of *S. romanzoffiana* and 74 of *S. aestuans*. To avoid collinearity among variables, which would overfit the model, we used Spearman correlation and excluded the variables most correlated by a threshold of 0.6 (Loiselle et al. 2003). Only six variables were retained (mean diurnal range, isothermality, max temperature of warmest month, mean temperature of wettest quarter, precipitation seasonality, precipitation of warmest quarter and precipitation of coldest quarter). To finish data preparation, we ran a background selection with the same number of occurrence points in the entire area of distribution of the species.

**Models**

Our first model aims to predict the distribution of *S. romanzoffiana* that will be used as predictor variable of *S. aestuans*. Our second model predicts the distribution of *S. aestuans* only considering climatic predictors. Finally, our third model comprises
the predicted distribution of S. aestuans with climatic and biotic predictor. The last model is built with the predictors of model 2 (mean diurnal range, isothermality, max temperature of the warmest month, mean temperature of wettest quarter, precipitation seasonality, precipitation of warmest quarter and precipitation of coldest quarter) plus the distribution of S. romanzoffiana as a biotic predictor.

Modelling and validation

Considering that the data selected for this study was presence data, we used the Maximum Entropy Model available in dismo package in R (Hijmans et al. 2011). Furthermore, this is a model of presence-background, which according to Elith and Graham (2009) performs better than some presence-absence models. Moreover, we emphasize that our purpose is not to find the best algorithm model for the above-mentioned species, but to compare performance of the model with and without the variable representing resources for the consumer. For each model we used 75% of location points as training data and 25% for test validation of the Maxent model. Also, for each model we saved suitability and presence/absence (using the threshold which gets higher specificity and sensibility) raster files, values of variables contribution, values of AUC, kappa and TSS which was built according to the best threshold following the ROC-AUC curve.

With those saved files, we made ensembles with 10 simulations for each model and built the final raster of suitability and presence/absence. We also built a table with the values of model performance (AUC, kappa, prevalence, TSS) for each simulation from models 2 and 3 in order to compare the model performance with and without biotic predictor. Afterward, we made the multiplication of ensembles to assess the concordance of species suitability between maps with and without biotic predictor. Using this method, we can assess the area with greater probability for S. aestuans to occur. All codes used may be seen in Suppl. material 3.

Results

Our maps show that the distribution of S. aestuans differs according to predictor variables used (Figure 1). For scores of model performance, there was no difference between models without and with biotic predictor use considering standard deviation (Table 1).

Apart from the metrics of performance, models with biotic predictor show an area of suitability smaller than that one without biotic predictor (Fig. 1A, B respectively). Furthermore, looking for variables’ contribution of the model, the presence of S. romanzoffiana (the resource used as biotic predictor) was the most important variable in species distribution (Fig. 2). Variables contribution and performance scores of all simulations can be seen in Suppl. material 4: Tables S1, S2.

To represent the difference between maps without and with biotic predictors in the same figure, we multiplied the pixel values of both maps (Fig. 3). There was
Figure 1. Ensemble from cross validation maxent maps of suitability, higher values present higher habitat suitability and higher chance of species occurrence, where: (A) model without resource presents only climatic predictor variables, and (B) model with resource presents climatic variables plus distribution of *Syagrus romanzoffiana*.

Figure 2. Variable contribution of models without and with biotic interaction, where: (A): contribution of variables in species distribution suitability without biotic predictor and (B): contribution of variables in species distribution suitability with biotic predictor. Predictor variables represents: bio2 = Mean diurnal range, bio3 = Isothermality, bio8 = Mean temperature of wettest quarter, bio15 = Precipitation seasonality, bio18 = Precipitation of warmest quarter, bio19 = Precipitation of coldest quarter, syagrus = predicted distribution of *Syagrus romanzoffiana*.
Table 1. Results of each model performance scores inside cross-validation (n=10) and number of presence pixels for models without and with biotic predictors. Values next to 1 present higher model accuracy and SD represents Standard Deviation.

<table>
<thead>
<tr>
<th>Values</th>
<th>Kappa</th>
<th>AUC</th>
<th>TSS</th>
<th>Presence pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model with biotic predictor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.527</td>
<td>0.983</td>
<td>0.527</td>
<td>2819726</td>
</tr>
<tr>
<td>SD</td>
<td>0.069</td>
<td>0.013</td>
<td>0.069</td>
<td>1518457</td>
</tr>
<tr>
<td>Without biotic predictor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.531</td>
<td>0.991</td>
<td>0.531</td>
<td>4660709</td>
</tr>
<tr>
<td>SD</td>
<td>0.053</td>
<td>0.007</td>
<td>0.053</td>
<td>1939159</td>
</tr>
</tbody>
</table>

Figure 3. Output of concordance between maps with and without biotic predictor. Values above 20 and under 40 indicate high suitability for the species only in the model without biotic predictor. Values above 40 indicate high suitability in both models.

concordance between models without and with biotic predictors. However, a model without biotic predictor suggests a larger possible occurrence area for the species. Besides, the variation in presence pixels among the models of ensemble is higher without biotic predictor (Suppl. material 4: Figure S1).

Discussion

The results indicate that our hypotheses are not rejected. The importance of including a variable representing a biotic interaction was clear in our modeling exercise. Suitability maps were more restricted by the biotic predictor, which was the most determinant variable in the distribution model. This indicates that commission errors were probably reduced considering the predominant food resource of
the consumer. Concerning performance scores, both models seem to be similar. This means that model refinement considering biotic interactions did not affect the performance of the modeling exercise. However, practical results such as the presence/suitability area for species distribution suggest the contrary. This leads to an important finding: models can reach high values of accuracy in scores but represent distinct results in terms of species distribution. In this way, we suggest that before making a modeling exercise, one should ask: do most used performance scores truly represent the reality and reliability of a model?

Even though the insertion of a biotic predictor may lead to neutral or worse model performance (Araújo and Luoto 2007), our models performed similarly. We used the three types of performance scores: threshold-independent, threshold-dependent and overall test of SDM performance (AUC, Kappa and TSS respectively) to compare the difference in score values (Liu et al. 2011). However, there was no difference in performance score values, which may tell us that they should not be considered alone. More important than selecting the best variables for prediction, we suggest that researchers should question themselves if the most important variables really matter considering the biology of the target species.

Furthermore, comparing the variation in presence pixels of each simulation inside the ensemble, the model with the biotic predictor tends to be more accurate and closer to the final ensemble. The lower variation may also indicate lower commission errors (Peterson et al. 2008). In addition, the resulting suitability map changed significantly, where the model with the biotic predictor presents an area of presence pixels that is half the size of the model without the biotic predictor. Considering that for conservation priority areas we cannot reach all areas and need to choose those which concentrate the higher probability of species occurrence, practical results may be more, or as crucial as, model performance. In this way, what can make the difference to prioritize an area or not are the importance of predictors used. Therefore, biotic predictors may not significantly influence performance scores, but certainly influence the result of species distribution area.

The fact that the biotic predictor was by far the most important indicates that biotic interactions may have a central role in species distribution, differently from what is expected in ENH. Surely, we cannot discard the possibility that the correlation of occurrences in data from the squirrel and from the fruit can be due to a sampling bias – researchers sample the same grids. We recommend that this should be further investigated. Even so, inventories are not usually done simultaneously by researchers from different groups and we then raised the possibility that results highlight the importance of knowing the biology of the modelled species. Notwithstanding the controversy of considering biotic predictors as determinant in species distribution, this seems to be a pattern corroborated by several studies (Meier et al. 2011; Meineri et al. 2012; Araújo and Rozenfeld 2013; Giannini et al. 2013; Palacio and Girini 2018; Simoes and Peterson 2018). Such a pattern seems to repeat itself especially for endotherms, since the distribution of ectotherms may be directly related
to climate conditions instead of biotic interactions, and the addition of biotic predictors may increase model complexity and decrease model performance (Engler et al. 2013), unless species have a strong biotic interaction. Anyway, we advocate that the knowledge of species interaction is essential for SDMs.

Since SDMs have been the most applied method to predict species distribution facing a changing world, our results may be considered as a call for attention for researchers. Accurate and reliable predictions are increasingly needed and, as it seems to be, models which combine climate and biotic variables are closer to reality (Palacio and Girini 2018). This kind of data may be more easily found and applied for species with specialized habits such as those which feed mainly in one kind of food resource (Giannini et al. 2013). Furthermore, the food web control may also be a driver in model accuracy. In a bottom-up control for instance, higher levels need the lower ones but not necessarily the other way around (Vázquez et al. 2007; Giannini et al. 2013).

*Sciurus aestuans* bases its diet in, and is the main disperser of, *S. romanzoffiana*, which indicates high specialization between these two species (Alvarenga and Talamoni 2006). This fact may be responsible for the high contribution of *S. romanzoffiana* distribution in the prediction model of *S. aestuans* distribution. This kind of result is more easily found in strong and specialized interactions (Giannini et al. 2013). However, understanding distribution of generalist species is also needed, making the knowledge of species ecology a bottleneck in SDMs (Araújo et al. 2014). Thus, we encourage new studies that predict species distribution matching climate and biotic predictors even though it is difficult to access this kind of data for some species.

There are different ways to establish conservation planning (Margules and Pressey 2000). However, the concern about a future with climate change is a congruence (Menon et al. 2002; Duarte et al. 2013). Hence, each attempt to reach a good conservation planning needs to go through the future prediction, which is the objective of most SDMs (Thuiller et al. 2003; Synes and Osborne 2011). This way, new approaches and methodologies to accurately predict the future are appreciated. Our study may contribute to predict species distribution accurately and close to reality and, therefore, to conservation planning in a future of climatic changes.

**Conclusion**

We suggest an integrative approach for SDMs which include biotic and climatic predictors, as indicated by other authors (Palacio and Girini 2018; Simoes and Peterson 2018). Our results also highlight that SDMs applied without a biological purpose do not make sense. We call the researchers’ attention to the application and assessment of SDMs, where each objective needs to be taken into account and the ecological sense can never be forgotten. We cannot just turn biological and ecological knowledge into numbers. Considering all this, research may be performed with different algorithms and scores but the first premise needs to be the previous knowledge of species and ecology.
References


Lobo JM, Lumaret JP, Jay-Robert P (2002) Modelling the species richness distribution of French dung beetles (Coleoptera, Scarabaeidae) and delimiting the predictive capacity


**Supplementary material 1**

**Occurrence data – *Syagrus romanzoffiana***
Authors: Gabriel Preuss, André Andrian Padial
Data type: occurrence
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/neotropical.16.e64892.suppl1

**Supplementary material 2**

**Occurrence data – *Sciurus aestuans***
Authors: Gabriel Preuss, André Andrian Padial
Data type: occurrence
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/neotropical.16.e64892.suppl2

**Supplementary material 3**

**Code used to perform all analysis***
Authors: Gabriel Preuss, André Andrian Padial
Data type: code
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/neotropical.16.e64892.suppl3
Supplementary material 4

Tables S1, S2, Figure S1
Authors: Gabriel Preuss, André Andrian Padial
Data type: measurements
Explanation note: Table S1. Values of variables importance for each simulation inside the ensemble. Table S2. Values of performance scores and presence pixels for each simulation inside ensemble. Figure S1. Variation in presence pixels among the models of the cross validation (n=10), where: x axis represents the model type and y axis represents the number of presence pixels in the map.
Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.
Link: https://doi.org/10.3897/neotropical.16.e64892.suppl4