Received: 16 August 2018 | Revised: 30 January 2019 | Accepted: 5 May 2019

DOI: 10.1111/jbi.13626

CORRESPONDENCE

Journal of Biogeography WILEY

Detecting outliers in species distribution data: Some caveats and clarifications on a virtual species study

Abstract

Liu et al. (2018) used a virtual species approach to test the effects of outliers on species distribution models. In their simulations, they applied a threshold value over the simulated suitabilities to generate the species distributions, suggesting that using a probabilistic simulation approach would have been more complex and yield the same results. Here, we argue that using a probabilistic approach is not necessarily more complex and may significantly change results. Although the threshold approach may be justified under limited circumstances, the probabilistic approach has multiple advantages. First, it is in line with ecological theory, which largely assumes non-threshold responses. Second, it is more general, as it includes the threshold as a limiting case. Third, it allows a better separation of the relevant intervening factors that influence model performance. Therefore, we argue that the probabilistic simulation approach should be used as a general standard in virtual species studies.

In an article by Liu, White, and Newell (2018), the authors used virtual species to test 10 different methods for identifying potential outliers from a set of occurrences in the context of species distribution modelling (SDM). Understanding how we can improve model predictions is very much needed to anticipate the effects of climate change on the distribution of biodiversity and to advance predictive macroecology in general (Araújo & Guisan, 2006; Bellard, Bertelsmeier, Leadley, Thuiller, & Courchamp, 2012; Guisan & Thuiller, 2005). As Liu et al. (2018) state, many authors have advocated for a stronger move towards the use of simulations in ecology (Zurell et al., 2010), including in macroecology and biogeography (Meynard & Kaplan, 2013; Moudry, 2015). In this context, we applaud Liu et al. (2018) for the diversity of methods tested, as well as the thoroughness with which they tested the factors that can affect the success of the different outlier detection methods. Indeed, the use of virtual species in this field has great potential, as it allows controlling and varying targeted aspects of the species properties, sampling schemes, landscape properties and/or modelling steps, to isolate and understand different sources of uncertainty in SDM studies in a way that would be unfeasible using real-world species.

However, we would like to point out shortcomings in their virtual species simulation method. To understand the basis of our criticism, it is therefore necessary to understand the first few steps in the virtual species simulation process. Generally, the first step involves defining a suitability response (also called a functional response) to one or more environmental gradients. This function indicates what the preferred environments are for the simulated virtual species. For example, we can think about a species that has a normally distributed response to mean annual temperature, meaning that it prefers intermediate temperatures and avoids extremely warm or extremely cold sites. However, when we go out in the field, we do not sample probabilities or suitabilities. We sample presences and absences, and we would like to estimate what the probability of occurrence of the species is, given the environmental conditions on those sites. That is why the simulation of virtual species to test SDMs requires transforming the simulated suitability function into presences and absences before the SDM strategy can be tested. At this step, we can distinguish two main approaches to translate the simulated suitability into a distribution pattern: we can interpret the suitability function as a probability of occurrence (hereafter, the 'probabilistic simulation approach'), or we can define a threshold under which the species will always be absent, and above which it will always be present (hereafter, the 'threshold simulation approach'). The strategy used by Liu et al. (2018) was the threshold simulation approach. In Meynard and Kaplan (2013), we discussed extensively why such a threshold simulation approach can be problematic, and advocated instead to interpret the functional response as a probability of presence. Liu et al. (2018) recognized the limitations of the threshold simulation method, but decided not to use the probabilistic approach for the following main two reasons: (a) '*it is clear that one disadvantage of the "probabilistic approach" is that it is very difficult to create virtual species with a specified prevalence*'; and (b) '*Studies using both threshold and probabilistic approaches have reached the same conclusions on the behaviour of the favoured threshold-selection method* (Liu, Newell, & White, 2016; Liu, White, & Newell, 2013)'. Here, we will argue that (a) controlling for species prevalence when using a probabilistic simulation approach is relatively easy; and (b) threshold and probabilistic approaches do not yield the same results. Furthermore, we will argue that implementing the probabilistic approach provides at least three important advantages, namely, that the probabilistic approach is in agreement with ecological theory, that it is more general and therefore allows studying the threshold response as a special case and finally that it allows separating different relevant factors in a way that is not feasible using the threshold simulation approach. We will conclude by advocating the use of the probabilistic simulation approach as the golden standard in any virtual species study aiming at advancing the SDM field.

First, controlling for the virtual species prevalence is not difficult. Meynard and Quinn (2007) had already provided an approach to control species prevalence which was based on adjusting the

overall prevalence of the simulated species using a simple linear transformation. Meynard and Kaplan (2013) provided another simple way to simulate virtual species with a probabilistic approach and a desired species prevalence based on the slope of the logistic curve and its inflection point. Scripts were provided in the appendix as well as suggestions on how this could be implemented in a multivariate environmental space rather than a single environmental variable. Additionally, a package allowing the implementation of the probabilistic approach was then published (*virtualspecies*, Leroy, Meynard, Bellard, & Courchamp, 2016), including the function 'convertToPA' that implements another approach to controlling species prevalence based on a numerical solution and a logistic transformation of the habitat suitability values. The implementation of the probabilistic approach using this package does not require any complicated code and allows simulating in both univariate and multivariate environmental frameworks. The three examples above show that fixing the desired prevalence of the simulated species can be accomplished in different ways and is relatively simple. The method chosen to do so should reflect the objectives of the study and/or the real mechanisms at work behind species' prevalence. For example, the linear transformation approach in Meynard and Quinn (2007) will uniformly increase the probability of presence throughout space preserving the shape of the original functional response, thereby best representing a species that is diffusely distributed over all habitat areas. Adjusting the location of the inflection point of a logistic functional response will expand or contract the area over which the species has a high probability of presence while minimally changing probability of presence outside this area. This flexibility represents important choices about the species–environment relationship and, therefore, represents an advantage to the probabilistic simulation approach.

Second, threshold and probabilistic approaches do not yield the same results. Previous studies comparing both approaches have revealed that, although the general tendencies may be the same, there are considerable risks to assuming equivalence of results without testing (see the examples in Meynard & Kaplan, 2012, 2013). Indeed, the probabilistic approach will introduce stochasticity in the generation of the presence–absence patterns, making it more difficult to predict specific occurrences, even under circumstances where the probability of occurrence is estimated accurately (Meynard & Kaplan, 2012). That stochasticity is often reflected in wider confidence intervals around the mean of classification rate performance indices when the probabilistic approach is used, whereas the threshold approach will often yield very narrow confidence intervals and perfect classification statistics, provided that the sample is not biased and is large enough. Such optimistic evaluation can have important implications regarding how we interpret the results of simulation studies, and what contrasts end up being statistically significant. For example, in Meynard and Kaplan (2013), we presented a case study dealing with thresholds used in transforming probabilities of occurrence to presence–absence predictions. Our results showed that there is no single threshold strategy that is recommendable across the board to generate presence–absence predictions. Indeed, when a non-threshold functional response is used, sensitivity and specificity are generally lower and a stronger function of sample prevalence than they are for a threshold functional response to the environment (Meynard & Kaplan, 2013). Moreover, confidence intervals around classification performance are also wider, making many of the comparisons statistically non-significant. In other words, the conclusion that one particular threshold is better in all cases, which is often the result of using a threshold simulation approach, is not valid any more when considering a probabilistic simulation approach (Meynard & Kaplan, 2013). These notable differences illustrate that a threshold approach can lead to a false impression of robustness and generalizability of results and, therefore, should not be used unless there is a specific reason to do so.

Given these differences, the question of which simulation strategy is more appropriate is fundamental to understand the generality of any virtual species study in an SDM context. There is at least one situation where the use of a threshold simulation approach is clearly inappropriate, and that is whenever we are interested in the effect of the shape of the functional response on model performance (Meynard & Kaplan, 2013). Indeed, no matter how complex the method used to generate the functional response was, the fact that a hard threshold is applied to determine presence versus absence removes much of the functional complexity and replaces the originally simulated suitability function with a threshold probability response (i.e. the probability is 0 below the threshold suitability value, and 1 above that threshold) (Leroy et al., 2016; Meynard & Kaplan, 2012, 2013). The main difference between the two 'functional responses' used by Liu et al. (2018) is that, in one case, they built a compound gradient based on the sum of suitabilities of each predictor independently (additive response), whereas in the second case, they allowed interactions between predictors by multiplying one gradient with the other (multiplicative response). The use of the terms 'linear' and 'nonlinear' by Liu et al. (2018) suggest complex functional relationships between environment and species distributions, whereas after applying the threshold, this functional complexity is only reflected in the linear versus nonlinear *border* between presence and absence areas in environmental variable space. The additive-threshold response is easier to predict than the multiplicative-threshold response due to the simple linear nature of this border in the prior case, but when non-threshold responses are studied, results can be considerably more difficult to predict in terms of presence–absence (see figure 4 in Meynard & Quinn, 2007, or figure 5 in Meynard & Kaplan, 2012). Avoiding a threshold simulation approach when studying the effects of the functional response on SDM performance is therefore key to providing appropriate general guidelines in this field. At the other extreme, there might be some limited situations where the threshold approach may provide a good approximation for the presence–absence patterns of interest. For example, when we are dealing with large temporal and spatial scales, the threshold approach may be valid because aggregating years of occurrence data at large spatial resolutions erases the probabilistic component of the response. This hypothesis is plausible, but needs further testing. However, in those cases, the relevance of the simulation study should be clearly tied to large spatial and temporal scales,

and we hypothesize that the scale at which the dynamics becomes threshold-like will depend on how rare the species of interest is, along with characteristics of the surveys and other aspects of the species ecology. In short, the use of a threshold or a probabilistic approach to simulate species occurrence depends, in part, on the objectives of the study, and, therefore, this choice should be clearly justified by the question at hand.

Outside of specific scenarios justifying the use of the threshold approach, we argue that the probabilistic approach should be preferred for three reasons. First, it is more in line with ecological theory, which overwhelmingly assumes non-threshold responses to the environment. For example, niche theory often assumes bell-shaped responses to environmental gradients, and metapopulation theory assumes that some environmentally suitable sites are empty and some unsuitable sites are occupied at any given time. Neither one of these cases can be simulated with a threshold simulation approach. Second, the probabilistic approach is more general because it allows testing a variety of response curves, including the threshold case, as particular cases of interest within a wider range of possibilities. Indeed, a threshold can be interpreted mathematically as a logistic curve with a slope equal to infinity. In Meynard and Kaplan (2012), we used this property to study what happens when the steepness of this slope varies, including very shallow and very steep slopes. If one is interested in the threshold response, then that case can be studied within the probabilistic framework while maintaining the perspective of what happens when the response is not a threshold. If the end result is highly dependent on the response being a threshold, we could detect this possibility within a probabilistic simulation framework but not within a threshold simulation framework. This allows clearly separating the effects of the threshold versus the effect of other factors that are involved in model performance. This brings us to our third reason to prefer the probabilistic simulation approach in general, which is that it allows clearly separating processes of interest that are not possible to separate using a threshold simulation approach. Liu et al. (2018)'s work provides a clear example. The overarching objective of their work was to identify outliers, potentially eliminating them from the model calibration phase in order to increase model performance. In statistics, outliers are defined as records (in this case occurrence records) that are distributed at the tails or outside of the statistical distribution of the dataset and are therefore different from mean observations. Outliers can be due to observation errors (e.g. misidentification of the specimens or geo-localization errors), but, importantly, they can also represent true observations in marginal habitats. In their simulations, Liu et al. (2018) generated outliers as occurrence records that were randomly located outside the species range. Therefore, 100% of their outliers represented observation errors. Indeed, the threshold simulation approach does not allow placing occurrences in marginal habitat because the probability of occurrence is either 0 or 1 everywhere. Using a probabilistic simulation approach instead would have allowed them to use outliers that represented a mixed bag of real observations in extreme environments along with observation errors anywhere in the distribution of the species, clearly distinguishing observation error from occurrences in marginal habitat in their characterization of outliers.

Because virtual species provide for such a unique opportunity to isolate and test how individual factors affect model performance, such an approach should be used wisely. In order to advance, we strongly believe that the threshold approach should be avoided when testing the effects of functional responses on SDM outcomes. In other situations, one should be explicit about the limits of the threshold approach and its pertinence for the case at hand. The probabilistic simulation approach is more general than the threshold simulation approach, and includes the threshold as one particular case within a continuum of possibilities, providing further flexibility to understand the advantages and shortcomings of different methods. All of these are, in our view, compelling reasons why the probabilistic approach should be taken as the general golden standard for virtual species simulations when testing SDM methodology.

Keywords

ENM, observation errors, outliers, prevalence, probabilistic approach, sample bias, simulations, species distribution models, virtual ecology, virtual species

ORCID

Christine N. Meynard https://orcid.org/0000-0002-5983-6289
David M. Kaplan https://orcid.org/0000-0001-6087-359X
Boris Leroy https://orcid.org/0000-0002-7686-4302

Christine N. Meynard[1]
David M. Kaplan[2]
Boris Leroy[3]

[1]*CBGP, INRA, CIRAD, Montpellier SupAgro, Univ Montpellier, Montpellier, France*
[2]*IRD, MARBEC (Univ. Montpellier, CNRS, Ifremer, IRD), Sète, France*
[3]*Unité Biologie des organismes et écosystèmes aquatiques (BOREA), Muséum National d'Histoire Naturelle, Sorbonne Université, Université de Caen Normandie, Université des Antilles, CNRS, IRD, Paris, France*

Correspondence

Christine N. Meynard, CBGP, 755 avenue du Campus Agropolis CS 30016, 34988 Montferrier sur Lez cedex, France.
Email: christine.meynard@inra.fr

REFERENCES

Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, *33*, 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., & Courchamp, F. (2012). Impacts of climate change on the future of biodiversity. *Ecology Letters*, *15*, 365–377. https://doi.org/10.1111/j.1461-0248.2011.01736.x

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*, 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2016). virtualspecies, an R package to generate virtual species distributions. *Ecography*, *39*, 599–607. https://doi.org/10.1111/ecog.01388

Liu, C., Newell, G., & White, M. (2016). On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and Evolution*, *6*, 337–348. https://doi.org/10.1002/ece3.1878

Liu, C., White, M., & Newell, G. (2013). Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, *40*, 778–789. https://doi.org/10.1111/jbi.12058

Liu, C., White, M., & Newell, G. (2018). Detecting outliers in species distribution data. *Journal of Biogeography*, *45*, 164–176. https://doi.org/10.1111/jbi.13122

Meynard, C. N., & Kaplan, D. M. (2012). The effect of a gradual response to the environment on species distribution modeling performance. *Ecography*, *35*, 499–509. https://doi.org/10.1111/j.1600-0587.2011.07157.x

Meynard, C. N., & Kaplan, D. M. (2013). Using virtual species to study species distributions and model performance. *Journal of Biogeography*, *40*, 1–8. https://doi.org/10.1111/jbi.12006

Meynard, C. N., & Quinn, J. F. (2007). Predicting species distributions: A critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, *34*, 1455–1469. https://doi.org/10.1111/j.1365-2699.2007.01720.x

Moudry, V. (2015). Modelling species distributions with simulated virtual species. *Journal of Biogeography*, *42*, 1365–1366. https://doi.org/10.1111/jbi.12552

Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Munkemuller, T., … Grimm, V. (2010). The virtual ecologist approach: Simulating data and observers. *Oikos*, *119*, 622–635. https://doi.org/10.1111/j.1600-0706.2009.18284.x

How to cite this article: Meynard CN, Kaplan DM, Leroy B. Detecting outliers in species distribution data: Some caveats and clarifications on a virtual species study. *J Biogeogr.* 2019;46:2141–2144. https://doi.org/10.1111/jbi.13626