



# Peer Community In Ecology

## Biodiversity databases are ever more numerous, but can they be used reliably for Species Distribution Modelling?

**Nicolas Schtickzelle**  based on peer reviews by 2 anonymous reviewers

Thibaut Ferraille, Christian Kerbiriou, Charlotte Bigard, Fabien Claireau, John D. Thompson (2023) Integrating biodiversity assessments into local conservation planning: the importance of assessing suitable data sources. *bioRxiv*, ver. 3, peer-reviewed and recommended by Peer Community in Ecology.

<https://doi.org/10.1101/2023.05.09.539999>

Submitted: 13 May 2023, Recommended: 03 October 2023

### Cite this recommendation as:

Schtickzelle, N. (2023) Biodiversity databases are ever more numerous, but can they be used reliably for Species Distribution Modelling?. *Peer Community in Ecology*, 100539. [10.24072/pci.ecology.100539](https://doi.org/10.24072/pci.ecology.100539)

Published: 03 October 2023

Copyright: This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

---

Proposing efficient guidelines for biodiversity conservation often requires the use of forecasting tools. Species Distribution Models (SDM) are more and more used to predict how the distribution of a species will react to environmental change, including any large-scale management actions that could be implemented. Their use is also boosted by the increase of publicly available biodiversity databases[1]. The now famous aphorism by George Box "All models are wrong but some are useful"[2] very well summarizes that the outcome of a model must be adjusted to, and will depend on, the data that are used to parameterize it. The question of the reliability of using biodiversity databases to parameterize biodiversity models such as SDM –but the question would also apply to other kinds of biodiversity models, e.g. Population Viability Analysis models[3]– is key to determine the confidence that can be placed in model predictions. This point is often overlooked by some categories of biodiversity conservation stakeholders, in particular the fact that some data were collected using controlled protocols while others are opportunistic.

In this study[4], the authors use a collection of databases covering a range of species as well as of geographic scales in France and using different data collection and validation approaches as a case study to evaluate the impact of data quality when performing Strategic Environmental Assessment (SEA). Among their conclusions, the fact that a large-scale database (what they call the "country" level) is necessary to reliably parameterize SDM. Besides this and other conclusions of their study, which are likely to be in part specific to their case study –unfortunately for its conservation, biodiversity is complex and varies a lot–, the merit of this work lies in the

approach used to test the impact of data on model predictions.

### **References:**

1. Feng, X. et al. A review of the heterogeneous landscape of biodiversity databases: Opportunities and challenges for a synthesized biodiversity knowledge base. *Global Ecology and Biogeography* 31, 1242–1260 (2022). <https://doi.org/10.1111/geb.13497>
2. Box, G. E. P. Robustness in the Strategy of Scientific Model Building. in *Robustness in Statistics* (eds. Launer, R. L. & Wilkinson, G. N.) 201–236 (Academic Press, 1979). <https://doi.org/10.1016/B978-0-12-438150-6.50018-2>.
3. Beissinger, S. R. & McCullough, D. R. *Population Viability Analysis*. (The University of Chicago Press, 2002).
4. Ferraille, T., Kerbiriou, C., Bigard, C., Claireau, F. & Thompson, J. D. (2023) Integrating biodiversity assessments into local conservation planning: the importance of assessing suitable data sources. *bioRxiv*, ver. 3 peer-reviewed and recommended by Peer Community in Ecology. <https://doi.org/10.1101/2023.05.09.539999>

## **Reviews**

### **Evaluation round #1**

DOI or URL of the preprint: <https://doi.org/10.1101/2023.05.09.539999>

Version of the preprint: 1

### **Authors' reply, 26 September 2023**

[Download author's reply](#)

[Download tracked changes file](#)

### **Decision by [Nicolas Schtickzelle](#) , posted 14 September 2023, validated 14 September 2023**

#### **Revision needed**

First, I want to apologize for the delay in getting a decision on your manuscript. A very large number of contacted reviewers declined, likely because the summer is often not a good period and more fundamentally because of the high workload of our research community.

I have now received the remarks of two reviewers, with complementary background: the first reviewer, a specialist, gave a detailed set of remarks and suggestions to further improve the manuscript, while the second reviewer, with a more generalist background, provides the feedback of a reader interested by the general question but without the detailed knowledge of the methodology used in the paper. Both reviews are detailed and demonstrate the commitment of the reviewers.

Both reviewers see great merit in your study and are expecting we can recommend the manuscript once it has been further improved according to their detailed set of suggestions. I agree with their view about the interest of the study, but also about the potential to improve it for both specialist and less specialist readers.

A few extra specific comments:

- L104-106: this seems a duplicate of previous sentence.
- At several places there is a mismatch in which region is T1, T2 and T3. Please check everywhere. One reviewer also asks meaningful questions on the regional context.
- L151: change "for assess" into "to assess".

## Reviewed by anonymous reviewer 1, 14 August 2023

[Download the review](#)

## Reviewed by anonymous reviewer 2, 25 August 2023

Dear Editor, Dear Authors,

I reviewed the preprint downloaded on bioRxiv under the doi <https://doi.org/10.1101/2023.05.09.539999>, Integrating biodiversity assessments into local conservation planning: the importance of assessing suitable data sources.

As a first introduction, I would like to thank the authors for the pertinence of the study. In a context of a huge amount of data available, shared or not but often under (even not) analyzed, the assessment of the suitability and guidelines to perform correct analysis seems crucial and welcome. Although my impression on the paper and its writing is positive, I may suggest some recommendations that could significantly improve the manuscript.

I ordered my remarks according to the line number :

L 27. I found the term "stakeholder" quite large. I would suggest a better explanation to whom this paper is addressed, and consequently, define more precisely the stakeholders, among scientists, public authorities and managers. If I write this remark at L.27, a more accurate term or definition would have a better place in the Introduction. I suggest to reduce the use of verb-ing at the passive voice.

L.28. "was done" please use a more accurate word (I may suggest "performed")

L35-36 : I would rephrase for: "Second, the collection of individual databases at the national 35 scale is necessary to complete local data and ensure the suitability of SDMs in a local context"

L50 : You present the biotic homogenization. I also understand the invasive species under this term. Although presented, it underlines one of major lack of the study regarding the SCP. Within each taxonomic group, some species have a particular importance, like the invasive species and the protected/rare/threatened species. As far as I know, stakeholders give more prioritization to areas with rare or threatened species. Conversely, if the biodiversity of an area included a large proportion of invasive species it can mismatch with an adequate choice of SCP. I would suggest two ways to deal with that : 1/ You can redo SDMs taking account the status of the species and/or 2/ you can present the % of observations including invasive or threatened/protected species, therefore if the % of species and the % of the observations are low, it should not have an impact on the global conclusion of your analysis. If they are high, I would highly recommend to perform the 1/.

L78-79. I would insist on a potential lack of knowledge to manage such tools

L81-85-+87 : About the check the suitability of models. A lot of data available was not gathered in order to answer a particular question. Therefore, analysis performed on data that was not designed to may induce several problems. ++ Some protocols are not completely transferable in each context, therefore add some words about the stat assessment seem to me crucial.

L94-95 : The role of stakeholders is not well defined -> how do they manage data are they enough competent to assess it? REgarding stakeholders, it depends also of conflict of interest e.g. between stakeholders and promoters/investors that require a less constraining mitigation measure. Some word about it would be welcome.

L.97-106: I had some difficulties to understand from where to where you are dealing your point. I would suggest to rephrase some sentences to be more precise and to clearly see each point. For each point, please

indicate a more concise problematic.

L.100: taxonomic groups. Which one are chosen? Why? What hypothesis do you suggest on these tax group?

L.104: A second third point is indicated. Maybe a fourth one?

L.110, Figure 1 and Table 1 : there is a mismatch between T2 and T3. Please correct it

Table 1 : Urbanization : A quick check on a satellite map show me that the T2 area is surrounded by two big cities (La Rochelle and Niort), whereas the T1 area seems quite far from an urban area. Is the city Lodève? Millau or Montpellier? Do you have any hypothesis on the urbanization context of each zone, since it is highly detailed, but not discussed. I also found an highway (N11 and a main railroad between the two main cities), whereas the fragmentation context as you describe it seems higher in T1 and T3.

L.152: for test -> to test

L.156: mismatch between the period and the time span you present. I think you should indicate "11 years"

L.158: Which ref did you choose? TaxRef? Did you use infra/supra specific taxa?

L.172: All databases, you mean all 3 databases combined, right?

L.180-182: IT should be more explained as hypothesis, therefore maybe a list of species may be useful? Does it include migrating species? if not, please move or discuss it in discussion.

L185 : Do you mean a different dispersal zone

L.189 : How do you deal with the buffer zone in the sea?

L.174-182: If you discuss about the migrating species, I am wondering why did you not include plants as 'not migratory' (at least at individual level), and bird as migratory. They have also a higher observational data, which could improve the SDM. Moreover, Papilionidae distribution is highly dependant on plant distribution as hosts.

L.198: I find it very discutable since the observer of a point is maybe not able to identify other taxa. (I am wondering if the bird data is often more accurate and with a better sampling effort than plants or insects) Therefore i think that a large inter-taxa identification variation occurs.

L.216,L.224: Please do not repeat the value of the threshold.

L.221. Please rephrase to : "The aim of SEA biodiversity conservation strategies.."

L.227,L.229 please change for active voice

L.236. Does the original community included deleted species? ex Nobs<15?

L.251: Please mention directly that Aves follows a Cauchy distribution

L.254: How did you specify the null model

L.258: Should not it be written as (1 | Studysite + Database) ?

L.274: I am wondering if a comparison with another contry would be complementary (eg. comparing a "european database")

L.283 : I think that the number of species observed comparatively to total species number would be useful.

L.325 : Did you interpret some behaviour or group like moth and butterflies? Could we expect different conclusions between these two groups?

L.374 : What are the advantages and disadvantages of the use of expert data vs the use of non-expert data, and the implication of sharing data with the scientific community and also with the public

L.396-397: Please rephrase the sentence as it is not very understandable.

L.402-404 : This sentence seems to generalize the conclusions though I understand that this fact can be only extended to certain conditions (taxa, data...)

L405-406 : (in regard to my comment on L.374). Could you provide some discussion about an expected model performance. ie discuss about the performance of the model about only but numerous opportunist data? To complete though it is partly discussed L.423. Maybe consider reordering?

L.409: ... therefore how to deal with absences (or pseudo-absence)

L.414-428: I am very curious about the consequences of your study to define conservation priorities regarding protected, rare or threatened species. Since these areas are the one to prioritize.

L.427-428: Opportunistic data tends to provide more observation of rare species. Indeed, people (naturalists) focus more on target rare or beautiful species whereas common species, though often observed, are not always reported in databases. I would be very please to have a short discussion about more biases like this one.

L.433-434: I think this is one of the most important point, with the hardest remaining issue is found I.439 about sharing data. A deeper discussion about this issue and even some solutions to deal with it could be useful.

L.460 : I would include a discussion/conclusion about the integration of site managers that can provide some tools about the suitability. However, these tools are not systematically shared, or are often published elsewhere. Solutions to deal with that should be underlined.

Appendix A : Some errors should be corrected, the same as explained above about the site names (T2, T3), and the time span (11 years). Some helpful information is provided in this appendix that, I think, should be move into the main text. (LL. 819,825, 828,829, 831)

In a general way, I can suggest to clearly detail the hypothesis and the expectations at the end of the introduction. The discussions could also be improve detailing some unused information (about the site and the species). Some concepts are not discussed and it could very interesting to give further details and discussion.

I look forward this paper soon published.