Integrating biodiversity assessments into local conservation planning: the importance of assessing suitable data sources

Thibaut Ferraille*[1,2], Christian Kerbiriou[3,4], Charlotte Bigard[5], Fabien Claireau[3,4], John D. Thompson[1]

[1] CEFE, Université de Montpellier, CNRS, EPHE, IRD, Montpellier, France
[2] Naturalia environnement, Avignon, France
[3] Centre d'Ecologie et des Sciences de la Conservation (CESCO), Muséum national d'Histoire naturelle, Centre National de la Recherche Scientifique, Sorbonne Université, Paris, France
[4] Centre d'Ecologie et des Sciences de la Conservation (CESCO), Muséum national d'Histoire naturelle, Station de Biologie Marine, Concarneau, France
[5] Agence Bretonne de la Biodiversité, Brest, France

*Corresponding author
Correspondence: thibaut.ferraille@gmail.com

ABSTRACT

Strategic Environmental Assessment (SEA) of land-use planning is a fundamental tool to minimize environmental impacts of artificialization. In this context, Systematic Conservation Planning (SCP) tools based on Species Distribution Models (SDM) are frequently used for the elaboration of spatially exhaustive biodiversity diagnostics. Despite the paradigm of "garbage in – garbage out" that emphasises the importance of testing the suitability of data for SDM and priority conservation areas, the assessment of database sources remains relatively rare. In addition, the lack of practical recommendations for the use of open-access databases by SEA stakeholders remains a problem. The aim of this study is to explore the quality of data sources that can be used by stakeholders to assess priority conservation areas in SEA. The study was done using data for nine taxonomic groups (commonly used in inventories for environmental impact assessment) and three databases available to SEA stakeholders. Three local administrative entities in very different socio-ecological contexts were used to examine three main issues : (i) the suitability of local versus national databases for assessing conservation priorities, (ii) differences among taxonomic groups or territories in terms of the suitability of databases, (iii) the importance of the quality of databases for the application of SDM to assess priority conservation areas. Our study provides several clear messages for potential users of open-access databases. First, the need for prudence in the interpretation of biodiversity maps. Second, for SDM, the collection of individual databases at the national scale is necessary to complete local data and ensure the suitability of models in a local context. Third, a data driven approach can lead to the use of notably different species communities to identify priority conservation areas when compared to the community in the original database. Finally, we propose a workflow to guide SEA stakeholders through the process of data rationalization and use in conservation planning.

Keywords: Data-driven approach, Species Distribution Models (SDM), Strategic Environmental Assessment (SEA), Systematic Conservation Planning (SCP)

1 - Introduction

Land-use change, in particular urban land expansion, leads to artificialization of habitats and soils and is one of the major causes of the loss of biodiversity (Maxwell et al., 2016; IPBES, 2019). The reduction and fragmentation of natural habitats leads to population declines and species extinction (Fahrig, 1997; Horváth et al., 2019; Lino et al., 2019), as well as biotic homogenization that involves an increase in genetic, taxonomic and functional similarity (Olden and Rooney, 2006; Zambrano et al., 2019).

A major tool to limit artificialization is the mitigation hierarchy used in environmental assessments studies. This approach consists of three sequential steps: “avoid” impacts, “reduce or minimize” impacts not avoided and “offset” residual impacts (Bull et al., 2016; Maron et al., 2016). However, in the current application of the mitigation hierarchy several weaknesses prevent it from achieving the goal of “No Net Loss” of biodiversity (Quétier et al., 2014; Bezombes et al., 2019). Avoidance is poorly implemented despite the fact that it is the first and most efficient step of the hierarchy (Bigard et al., 2017; Phalan et al., 2018). What is more, the mitigation hierarchy is mostly applied in a project-by-project approach without scaling up (Pope et al., 2013; Bigard et al., 2017), which limits proper consideration of fragmentation issues (Gontier et al., 2006) and cumulative impacts (Whitehead et al., 2017), including those of multiple small projects (Bigard et al., 2017).

To anticipate avoidance measures, Strategic Environmental Assessment (SEA) of land-use planning is a global and fundamental tool to minimize environmental impacts (Baker et al., 2005). SEA provides for the integration of avoidance measures early in the land-use planning process through environmental assessment of policies, plans and programs (Fundingsland Tetlow and Hanusch, 2012; Bigard et al., 2020). However, the implementation of SEA is often based on biodiversity diagnostic maps that are rarely complete and exhaustive. Indeed, biodiversity diagnostics are rarely exclusively based on empirical observations from field surveys (Phalan et al., 2018) and usually use areas and documents already known (e.g., protected areas and green infrastructures).

Spatial modelling provides a tool for the elaboration of spatially exhaustive diagnostics of biodiversity maps for land-use and conservation planning (Almenar et al., 2019; Tarabon et al., 2019; Bigard et al., 2020; Tulloch et al., 2019; Baker et al., 2021; Boileau et al., 2022). Among these methods of biodiversity modelling, Species Distribution Models (SDM) are widely used to predict suitable habitat for species based empirical observations (Guisan et al., 2017; Zurell et al., 2020) and are increasingly used in conservation planning (Guisan et al., 2017; Domisch et al., 2019; Baker et al., 2021). Systematic Conservation Planning (SCP) tools are also particular pertinent to identify priority biodiversity stakes and avoid the adoption of an *ad hoc* approach (Margules and Pressey, 2000; Pressey and Bottrill, 2008) in order to inform SEA (Tulloch et al., 2019).

The management of databases and their use for conservation planning is a critical issue for the application of such methods to practical conservation planning. The databases available for stakeholders are often limited because of data sensitivity or ownership issues, although more and more programmes contain data that are publicly available and use of them can be made without any particular attention to their quality (Costello and Wieczorek, 2014; Tittensor et al., 2014). Surprisingly however, despite the prevailing recognition of the “garbage in – garbage out” that emphasises the critical importance of the quality of data (Sanders and Saxe, 2017; Canbek, 2022), an examination of data suitability is relatively rare in local conservation planning (Rondinini et al., 2006; Hermoso et al., 2015a). In this context, some authors argue the necessity of examining the sensitivity of model results to the nature of the datasets that are used (Sanders and Saxe, 2017; Clare et al., 2019; Velazco et al., 2020). Another particularly important point that can influence distribution modelling is the spatial extent of the data, and in particular the question of whether to use only local data or those collected on a larger scale (Baker et al., 2021; Meyer, 2007). The choice and possible combination of data sources is part of this problem due to the fact that they often vary considerably in their design, the gradients covered, and potential sampling biases (Fletcher et al., 2019; Boyd et al., 2023). Basically, the use of available databases requires a rigorous test of their quality and pertinence (Zuckerberg et al., 2011), especially when used for analyses such as SDM (Tulloch et al., 2016; Domisch et al., 2019). As recognised by Clare et al., (2019), the lack of practical recommendations for the use of databases that differ in terms of their quality and pertinence by public authorities or other institutions remains a serious problem.

The overall goal of this study is to test the influence of different database sources that can be used by stakeholders to map priority conservation areas in SEAs based on SCP. To do so, we studied three local administrative territories that occur in different socio-ecological contexts in France. The study has three main objectives. First, in terms of the quantity of data available for nine taxonomic groups commonly used in

naturalist inventories in environmental impact assessment studies, we assess the variation in the suitability of the available databases at three scales (local, regional and national) for application of SDM. Second, by focusing on two taxonomic groups with different dispersion capacities, home range sizes and environmental response scales, we explore the influence of databases for the application of SDM to assess priority conservation areas. Third, we analyse the influence of this data-driven approach on the composition of species communities that are ultimately considered in priority conservation areas relative to the actual communities in the original databases. Third, we analyse the influence of this data-driven approach on the composition of species communities that are ultimately considered in the analysis relative to the actual communities in the original databases.

2 - Methods

2.1 - Study sites

To assess the availability and suitability of pertinent data sources, we selected three French local administrative entities in charge of land-use planning: Lodévois-Larzac (T1), Brocéliande (T2), La Rochelle (T3) (Figure 1, Table 1). We selected these territories on the basis their contrasting social, ecological and geographical contexts in order to examine patterns of variation of data suitability among sites. For example, each of three territories is composed of different ecosystems and bioclimates and the sites vary in terms of urbanization pressures (from 3% to 28% artificial land-use for T1 and T2 respectively) and the presence of protected areas in their territories (from 2% to 70% of the protected territory for T3 and T1 respectively).

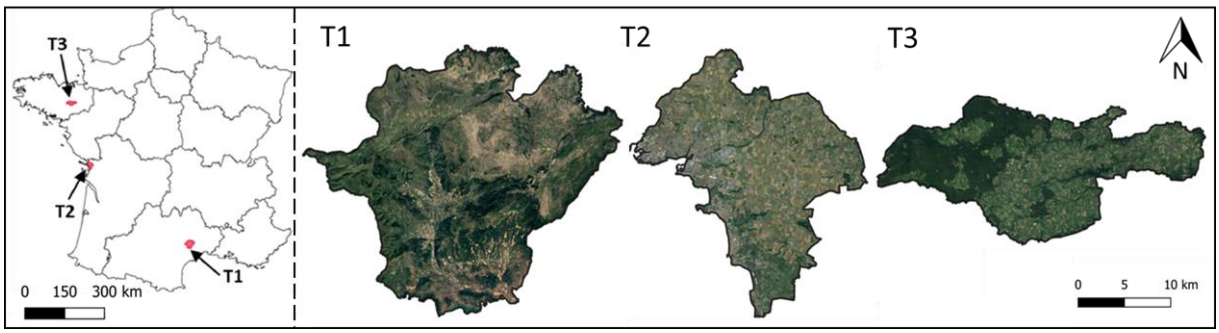


Figure 1 - Localisation of the study sites in France: T1 is Brocéliande, T2 is La Rochelle, T3 is Lodévois-Larzac. Source: IGN, Google, 2023.

Table 1 - Description of three study sites in France

Territory		**Lodévois-Larzac** (T1)	**La Rochelle** (T2)	**Brocéliande** (T3)
Country / Region		France / Occitanie	France / Nouvelle-Aquitaine	France / Bretagne
Main city coordinates		43°43'57"N 3°19'02"E	46°09'35"N 1°09'05"W	48°00'08"N 2°05'48"W
Area		554 km^2	327 km^2	298 km^2
Climate		Mediterranean	Oceanic	Oceanic
Major habitats / land-use		Mediterranean forest and scrubland	Intensive agricultural field, swamps, few forests	Pastures, temperate forest
% land-use (1)	Artificial	3%	28%	7%
	Farmland	3%	55%	29%
	Semi-natural	94%	17%	64%
Landscape and dynamics		Rural, woodland expansion	Coastal, intensive agriculture, urban expansion	Rural, agricultural intensification
Urbanization		Close to a major city, across by a highway	Economic and tourist dynamism	Close to a major city, across by a highway
% protected area (2)	Regulatory	0%	327 ha – 0.99%	0%
	Land control	1618 ha – 0.6%	148 ha – 0.45%	80 ha – 0.25%
	Contractual	38096 ha – 68.8%	3534 ha – 10.7%	586 ha – 2%
	Total	38602 ha – 69.7%	3546 ha – 10.8%	592 ha – 2%

(1) source: OSO Land Cover (Inglada et al., 2019). Artificial is impermeable surfaces; crop field is annual crop and orchards and vineyards; semi-natural is hardwood and soft wood and grasslands and water.
(2) source: https://inpn.mnhn.fr/

2.2 – Workflow of analysis

A methodological framework was developed to test the influence of different database sources in mapping priority conservation areas in SEAs thanks to a SCP approach, all steps are summarized in Figure 2.

2.3 – Database available for SEA stakeholders

2.3.1 – Content of available databases

We focused on nine taxonomic groups commonly used for naturalist inventories for environmental impact assessment studies: Amphibia, nesting Aves (hereafter name Aves), Chiroptera, Flora, Mammalia aptera (hereafter name Mammalia), Orthoptera, Odonata, Papilionidae and Reptilia (Bigard et al., 2017; Guillet et al., 2019; Iorio et al., 2022). We selected three open-access databases containing these groups that can be widely used by SEA stakeholders for the assessment and hierarchy of conservation priorities (Figure 2, step 1.1).

The first of these databases concerns the French Natural and Landscape Information System (SINP) that is structured at the regional scale in France. This database is composed of opportunist observations and only contains presence data for taxa for which identification is confirmed by experts (Jomier et al., 2018). Access to this database requires a data extraction request to the regional administrative structure in charge of the database. Generally, a special request is necessary above a 10km buffer zone around the study site (maximum 2000 km2), the data remains accessible only locally. Each site has its own database (see Appendix A.1).

The second database is the Global Biodiversity Information Facility (GBIF) an international platform for the provision of biodiversity data that is based on information collected from various databases (Telenius, 2011). It is composed of observation data that are not based on protocols and for which presence data and identification are not subject to expert confirmation. The data downloading is autonomous from the website (see Appendix A.1).

The third database is a French biodiversity monitoring scheme (Vigie Nature) dedicated to assess spatio-temporal populations trends. Within this monitoring scheme data collection is based on a standardized biodiversity survey. Despite local spatial distribution heterogeneities, the sampling plan ensures a representation of the current national distribution of habitats and landscapes across France (Julliard and Jiguet, 2002; Mariton et al., 2022). Homogeneity in identification criteria and compliance with the protocol are ensured by offering training to volunteers. This database is composed of presence/absence and abundance data. Access to these databases requires a data extraction request to the person in charge (see Appendix A.1).

These three databases were combined in two ways: All databases (i.e. SINP, GBIF and Vigie Nature) used in section 2.3 (i.e. for assess which taxa are enough documented within each dataset) and National databases (i.e. GBIF and Vigie Nature) used specifically in section 2.4 (i.e. for test the effect of database sources on SDM performance).

The databases were collected for continental France except for the SINP that was collected in 10km buffer zone around the study sites due to the access restrictions explained above. The databases were collected over a period of 10 years (i.e. from 01/01/2010 to 31/12/2020) and data with spatial inaccuracy greater than 50 meters were not considered. We made a series of operations to standardize, correct and homogenize species names, transform data into occurrences, limit their spatial biases and identify the presence of species into each study site (Appendix A.1).

2.3.2 – Quantitative description of databases for SDM

For the nine taxonomic groups, four metrics were selected to quantitatively describe the amount and thus the suitability of each database (i.e. SINP, GBIF, Vigie Nature and All databases) for the realization of SDM in presence-only (Figure 2, step 1.2): (i) the number of species observed in the study site; (ii) the proportion of species with < 15 observations which represents the minimal threshold for the utility of SDM with more accurate predictions than in a random model (Støa et al., 2019); (iii) the proportion of species with between 15 and 50 observations, i.e. the minimum number of presences recommended for SDM (Merow et al., 2014 in Guisan et al., 2017); (iv) the proportion of species with > 50 observations, i.e. highly suitable for modelling (Støa et al., 2019). These metrics were calculated at three different scales: (a) for each territory including a 10km buffer zone (local scale) that is the maximum extent for a SINP data request; (b) on a regional scale that is used for the structure of biodiversity data in France; (c) for continental France. SINP database is analysed individually only at the local scale due to the previously mentioned restriction of access, nevertheless it integrated the three scales of the All databases.

2.4 - Systematic conservation planning process

For the three study sites, we identified priority conservation areas with a Systematic Conservation Planning (SCP) tool based on Species Distribution Models (SDM) (Figure 2, steps 2.1 and 2.2). Several variants of SDM were made using different database sources (i.e. GBIF, Vigie Nature and National databases) and two methods of generating pseudo-absences (named individual database or mixed databases). Among the nine taxonomic groups studied above, only Aves and Papilionidae taxa were analysed to compare the tests in the SCP process. The data available in France for these groups seems to be sufficient to realize SDM with each database. The use of these two groups allows for a comparison between one group of highly mobile taxa with a large home range (Aves) and another group with a smaller home range and whose movement closely tracks local environmental variation (Papilionidae).

2.4.1 - Species distribution modelling (SDM)

We modelled favourable habitats for birds and butterflies in the three study sites using SDM (Figure 2, step 2.1). A resolution of 50m was used to meet the needs of the SEA of land-use planning. A buffer zone of 10km around each of the study sites (i.e. local scale) was used for the SDM prediction to limit any border edge effects and to increase the number of species that could be modelled and evaluated. Indeed, species with less than 15 data points for the calibration (threshold explain above, Støa et al., 2019) and/or less than 10 data for performance evaluation (threshold defined by expert opinion) were not modelling.

Biodiversity data used for the SDM came from the databases decribed above at the national scale, according to the results of section 2.3 (Table 2). These data were separated into two independent datasets that allow for robust validations with independent data (Matutini et al., 2021). National without local data were used for model calibration and local data were used only for model performance evaluation. Therefore, the SINP database, which is only available at local scale, was not used for model calibration.

For model calibration, pseudo-absences were generated with two methods. First, methods to generate pseudo-absences in the individual databases for all their data (i.e. GBIF, Vigie Nature and National databases). For databases with a protocol for sampling (i.e. Vigie Nature) to optimize species detection (day and year periods), the absence points were defined as all the points without the observed species. For databases without such sampling protocols for all their data (i.e. GBIF and National databases), pseudo-absence data were generated with the target-group (TG) approach, which infers the sampling bias from the aggregated occurrences of (TG) species, i.e. the respective taxonomic groups (Ponder et al., 2001; Anderson, 2003; Phillips et al., 2009). Second, a method to mix the presence data in the GBIF and National databases with the absence from Vigie Nature (named mixed databases) was applied (Hermoso et al., 2015a).

Three types of environmental variables used for SDM: biogeographic, human occupancy and pollution and fragmentation (Appendix A.2).

SDM were calibrate by Random Forest down-sample (Valavi et al., 2021a) which according to Valavi et al. (2021b) is among the best performing models for presence-only data. Although the Random Forest is not very sensitive to the non-independence of the variables and over-parametrization (Matsuki et al., 2016; Srisa-An, 2021), in order to be parsimonious, the collinear variables were removed (Pearson >0.7, Appendix A.2, Brun et al., 2020). Thirty bootstraps were performed for each SDM (Guisan et al., 2017) using a calibration for 70% of the data at the national scale outside of local scale. The thirty Random Forest bootstraps were combined with mean to provide an ensemble prediction of habitat suitability for all species.

The performance evaluation of the models was done using the Boyce's index (CBI), the most suitable metric for model in presence-only (Boyce et al., 2002; Leroy et al., 2018), with local data (i.e. All databases combining SINP, GBIF and Vigie Nature). Dubos et al. (2022) reveal the CBI turns out to be misleading in some cases, thus we used a threshold of 0.3 to define good or poor model quality.

2.4.2 – Systematic conservation planning (SCP) tool

Priority conservation areas in the three territories were analysed from SDMs for each database source in using a SCP tool (Figure 2, step 3.1). To meet the needs of SEAs, the study site was restricted to administrative boundaries with a buffer zone of 1km to maintain coherence between administrative entities and a resolution of 50m. The SEAs biodiversity conservation strategies aim is to establish priorities for the whole territory and all the cells have the same cost value of 1. The objective was a maximum coverage objective that seeks to maximize the number of features, i.e. the SDMs (Church et al., 1996). The features were the predictions of SDM with a Boyce's index > 0.3 to limit the influence of poor models. The priority conservation areas decisions

were between 0 and 1. To obtain a priority gradient, we cumulated Ferrier importance scores (Ferrier et al., 2000) from nine targets of the total amount of each feature (from 0.1 to 0.9 every 0.1). The package "prioritizr" was used (Hanson et al., 2021) with the open-source solver SYMPHONY (Kim et al., 2023).

2.5 - Comparative analysis of SDM and priority conservation areas

The influence of database sources on SDM predictions and priority conservation areas were analysed (Figure 2, step 2.2 and 3.2). SDM performance evaluations were analysed between database sources. The SDM prediction and priority conservation areas maps were compared with the Spearman's rank coefficient (Phillips et al., 2009) and the Schoener's D index as a measure of projection overlap (Schoener, 1968) which was calculated with the ENMTool R package (Warren et al., 2008; Warren and Dinnage, 2022).

2.6 - Species community analysis

We assessed the influence of the complete data driven workflow on the composition of species communities, i.e. differences between the original community in the database and the final community used to identify priority conservation areas (Figure 2, step 4).

To do so, we developed a workflow resistance score for each of the methodological filters for all species. A score of 1 is allocated to species observed in the study area that did not cross any of the stepwise filters. A score of 2 is allocated to species with sufficient data to calibrate SDM, i.e > 15 national observations, **or** evalue the performance of SDM, i.e. > 10 local observations. A score of 3 is allocated to species with sufficient data to calibrate **and** evaluate the performance of the SDM. A score of 4 is for species that were present in the final analysis as a priority species for conservation planning (i.e. with the two previous filters and a Boyce's index > 0.3). In order to assess species composition bias ultimately considered in priority conservation areas, species communities were analysed through traits that can influence species detection (mass, displacement capacity, period activity) and ecological traits (habitats, specialisation) (Appendix A.3). Missing data were completed with a trait imputation procedure generated using the R package "missForest" (Stekhoven and Bühlmann, 2012) by considering evolutionary relationships in the imputation process (see Carmona et al., 2021) using the R script of Toussaint et al. (2021). Due to the nature of the response variable (i.e. ordinal scoring including four modalities), we used ordinal regression mixed models with cumulative link using the clmm function of "ordinal" R package (Christensen, 2022). We adapted the link function to the data distribution for Aves and Papilionidae, using the "cauchit" and "logit" links, respectively. Species traits were used as fixed effects, while the random effects selected were the study sites for Aves and the combination for study sites and database sources for Papilionidae. Finally, we evaluated the quality of the full model by comparing to the null model with Akaike's information criterion (AIC) (Mac Nally et al., 2018).

Thus, our models were structured in the following way:

For Aves $$Resistance \sim \beta_0 + \beta_i \,.'trait'_i + 1|StudySites$$

For Papilionidae $$Resistance \sim \beta_0 + \beta_i \,.'traits'_i + 1|StudySites_Databases$$

Where β is the parameter estimates, i correspond to the variables of 'traits' using in fix effect, 0 is the shift between ordinal class of resistance (i.e. 1|2, 2|3 and 3|4) and "1|" is the random effect.

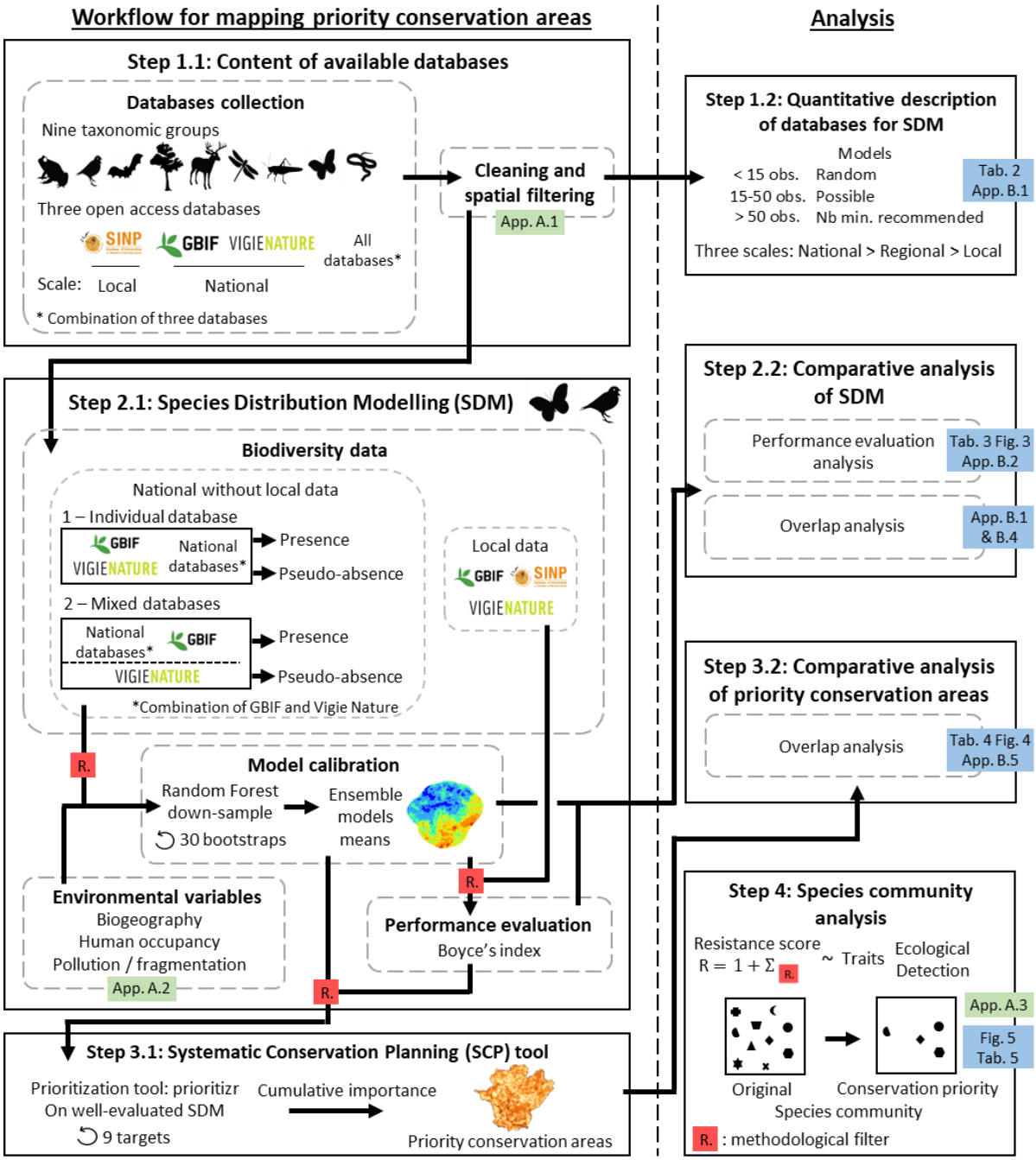


Figure 2 - Methodological framework applied to test the influence of different database sources to map conservation planning areas in SEAs based on SCP. The green and blue boxes are the appendices detailing the methodology and the results of the analyses, respectively. The red boxes "R." are the methodological filters used for generate workflow resistance score.

3 - Results

3.1 - Quantitative description of databases for SDM

The use of individual observation databases provided a limited number of observation data for distribution modelling of many species, regardless of the taxonomic group (Table 2, Appendix B.1). None of the GBIF and SINP individual databases have more than 50 observations for all nine taxonomic groups, and there are currently no programs for three of the nine taxonomic groups in the Vigie Nature databases. Local databases and, to a lesser extent, regional databases are not equivalent in terms of the amount of data available for different taxa for the three study sites. The difference is particularly pronounced for the SINP databases, with no data available for five taxonomic groups in the T3 study site, whereas in the T1 study site, six taxonomic groups have sufficient data to model over 50% of the species. At the regional scale, the number of observations per species is highly variable among territories for the GBIF databases and is more similar for the Vigie Nature databases.

The use of combined databases (i.e. All databases) increased the total number of species that can be studied. At the local scale, the proportion of species in each suitability class showed only small changes, while at the national scale of France, it allowed a significant gain in species with suitable data (Table 2). Indeed, on a national scale, the GBIF and Vigie Nature databases are complementary with each other. For example, the GBIF has few Chiroptera data, which is complemented by Vigie Nature data, and vice versa for Amphibian data.

At the national scale, aggregation of the databases seem to provide the most suitable setup (databases and scale) for SDM analysis. Using these compiled, national databases provides a large amount of data for a large number of species present in the three territories (Table 2 and Appendix B.1).

3.2 - SDM and priority conservation areas analysis

The evaluation of SDM revealed differences among the database sources; none of which produced more than 87% of satisfactory models for the two studied taxonomic groups and some had less than 20% of satisfactory models (Table 3, Appendix B.2). Use of the GBIF data led to a higher proportion of well-evaluated SDM, ranging between 48 and 79% of satisfactory models for the species in the two taxonomic groups. GBIF data are also more suitable than Vigie Nature data, they produced between 11% and 37% more satisfactory models than the latter database (Table 3). Nevertheless, between 4% and 9% of species provide well-evaluated models from Vigie Nature and poorly evaluated by GBIF database. The combination of National databases (i.e. GBIF and Vigie Nature) decreased the performance of SDMs with GBIF data, but still yield better results than SDM based on the Vigie Nature database. The substitution in individual databases of pseudo-absences for the absences from Vigie Nature (i.e. mixed databases) reduced the performance of models based on GBIF, but increased the performance of the combined National databases (Figure 3, Appendix B.2). Regardless of the database used, our analyses revealed significant differences between territories (Table 3). No SDMs for butterflies in T2 could be evaluated due to insufficient local data. For the T3 study site, over 50% of the Papilionidae and Aves models perform poorly, whereas for the T1 study site poor models occur in less than 50% of the evaluations.

Although important differences in model performance between the database sources used for SDM showed a high degree of overlap, as indicated by Shoener's D index with values above 0.8, the ranking of habitat suitability was highly variable. This was in particularly the case for the GBIF and Vigie Nature databases that had median spearman's rank coefficient values between 0.2 and 0.5 and a very wide distribution (Appendix B.3). The substitution of pseudo-absence data in GBIF and National databases by absence data from Vigie Nature, showed a similar situation (Appendix B.4).

For priority conservation areas, whatever the individual databases used, the overlaps with Schoener's D index were above 0.72 and similar in each territory. Nevertheless, Sperman's rank coefficients showed a greater difference in prioritization ranks in particular between GBIF and Vigie Nature and for the T3 study site (Table 4). Between maps of priority conservation areas, we observed similarities in overlap, despite a significant difference in the hierarchy of areas to be prioritized (Figure 4, Appendix B.5).

3.3 - Species community analysis

The distribution of workflow resistance scores showed that only 30% and 42% respectively of Aves and Papilionidae species were integrated in priority conservation areas maps. Among species not integrated, the workflow steps filtering the most species concern the amount of suitable data for model evaluation followed

by the quality of the models (Figure 5). The analysis of the species community composition observed in each of the three study sites in comparison with the species community integrated in priority conservation areas revealed significant differences. For Aves communities, the differences concern an under-representation of nocturnal species, large species with high dispersal capacity, and species of swamp habitats and deciduous forests in relation to the observed species community in the databases for the three study sites. Conversely, species that favour urban habitats, shrubland, grassland and coniferous forests are over-represented in the species community of the final maps, as are species with specialized diets and foraging strata (Table 5). For Papilionidae communities, common species with long flight periods are over-represented in the final community used for analysis. Species related to anthropogenic and thermo/meso Mediterranean habitats, and species that use a wide range of hostplants are over-represented in relation to the original species community, while supra-Mediterranean species and those of montane environments are under-represented (Table 5).

Table 2 - Quantitative description of the observation databases (Local SINP, GBIF, Vigie Nature) for nine taxonomic groups in the three study sites ("T1": Lodévois-Larzac, "T2": La Rochelle, "T3": Brocéliande) for "local" (buffer of 10km), "Regional" and "France" (continental). "nSp" is the number of species observed at least 3 times in a given territory within a 10 km buffer zone, ">15obs" is the percentage of species with more than 15 observation and ">50obs" the percentage with more than 50 observations (red < 25%, orange 25-75%, green >75%, grey - no data).

Taxa		GBIF									SINP			Vigie Nature									All databases combined								
		Local			Regional			France			Local			Local			Regional			France			Local			Regional			France		
		T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3
Amp	nSp	11	1	11	13	2	11	14	2	11	13	2	0	No program									14	2	11	14	2	11	14	2	11
	>15obs																														
	>50obs																														
Ave	nSp	100	73	77	198	159	120	204	170	131	197	94	82	114	155	101	182	143	120	201	170	132	205	171	132	205	166	132	205	171	132
	>15obs																														
	>50obs																														
Chi	nSp	4	0	5	13	13	5	19	19	21	21	6	18	25	22	24	25	22	24	25	22	25	26	22	26	26	22	26	26	22	26
	>15obs																														
	>50obs																														
Flo	nSp	786	476	639	1428	1093	825	1573	1261	995	1703	1335	843	286	177	136	825	657	413	1336	1105	848	1844	1460	1093	1844	1158	1093	1844	1460	1093
	>15obs																														
	>50obs																														
Mam	nSp	10	1	14	23	6	21	23	6	28	23	7	30	No program									23	7	30	23	6	30	23	7	30
	>15obs																														
	>50obs																														
Odo	nSp	38	5	23	56	7	28	57	7	30	57	1	0	0	1	24	56	7	27	57	7	30	57	7	30	57	7	30	57	7	30
	>15obs																														
	>50obs																														
Ort	nSp	53	3	24	93	25	26	95	25	34	91	8	0	25	19	19	25	19	19	25	19	19	98	25	34	98	25	34	98	25	34
	>15obs																														
	>50obs																														
Pap	nSp	115	15	33	144	18	43	147	18	52	154	11	0	0	0	50	132	18	51	149	18	52	154	18	52	154	18	52	154	18	52
	>15obs																														
	>50obs																														
Rep	nSp	18	2	7	20	4	7	21	4	7	21	4	0	No program									22	4	7	22	4	7	22	4	7
	>15obs																														
	>50obs																														

Table 3 - Proportion of well-evaluated SDM corresponding to Boyce's index greater than 0.3. National combines GBIF and Vigie Nature databases.

Taxa	Study site	Number of species	Individual database			SDM with Vigie Nature absence	
			GBIF	Vigie Nature	National	GBIF	National
Aves	T1	114	79 %	61 %	70 %	87 %	82 %
	T2	87	68 %	31 %	47 %	57 %	46 %
	T3	56	48 %	38 %	29 %	52 %	43 %
Papilionidae	T1	127	75 %	52 %	64 %	75 %	73 %
	T3	25	52 %	16 %	20 %	12 %	20 %

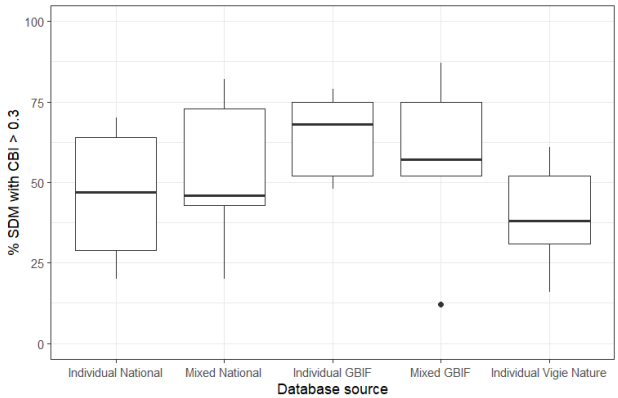


Figure 3 - Proportion of SDM with Boyce's index (CBI) greater than 0.3 by database source ("Individual" and "Mixed" with Vigie Nature absence), by combining territories (T1, T2, T3) and taxonomic groups (Aves, Papilionidae). "National" combines GBIF and Vigie Nature databases.

Table 4 - Overlap of priority conservation areas between database sources using two metrics: Schoener's D index (D) and Spearman's rank coefficient (S cor). VN is Vigie Nature database and National combines GBIF and Vigie Nature databases.

Taxa	Study sites	Individual database						Pseudo-absence of individual database - absence from VN			
		National – GBIF		National – VN		GBIF - VN		GBIF		National	
		D	S cor	D	S cor	D	S cor	D	S cor	D	S cor
Aves	T1	0.82	0.56	0.81	0.53	0.78	0.36	0.75	0.88	0.79	0.38
	T2	0.82	0.74	0.76	0.49	0.72	0.42	0.76	0.64	0.79	0.64
	T3	0.77	0.34	0.81	0.51	0.78	0.29	0.73	0.19	0.79	0.39
Papilionidae	T1	0.85	0.64	0.84	0.65	0.83	0.59	0.72	0.66	0.82	0.66
	T3	0.80	0.49	0.77	0.31	0.78	0.22	0.76	0.17	0.72	0.06

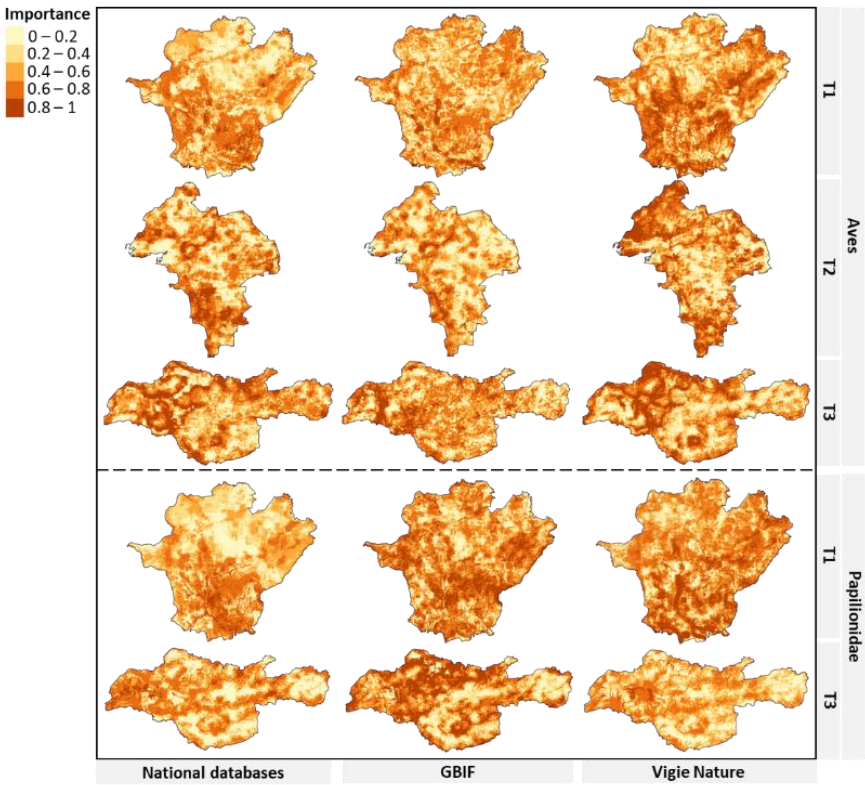


Figure 4 - Maps of priority conservation areas for Aves and Papilionidae of tree study sites (T1, T2, T3) from different individual database source (All, GBIF, Vigie Nature).

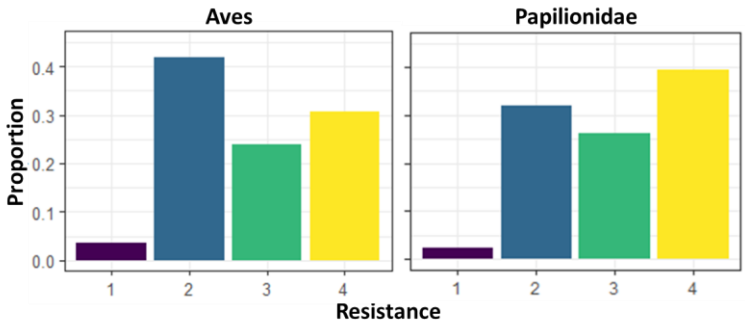


Figure 5 - Workflow resistance scores for Papilionidae and Aves. 1 - species observed in the study area that did not cross any of the stepwise filters. 2 - species with sufficient data to calibrate SDM, i.e > 15 national observations, or evalue the performance of SDM, i.e. > 10 local observations. 3 - species with sufficient data to calibrate and evaluate the performance of the SDM. 4 - species present in the final analysis, i.e. Boyce's index > 0.3.

Table 5 - Parameter estimates (β), standard error (se) and P-values for the full model of Aves and Papilionidae species resistance to the workflow. Appendix B.6, the evaluation of the quality of the model.

	Aves				Papilioniadae			
	Variables	β	se	P-value	Variables	β	se	P-value
Full model	β_0 1|2	-7.84	1.12	/	β_0 1|2	-2.54	1.28	/
	β_0 2|3	0.14	0.33	/	β_0 2|3	1.39	1.26	/
	β_0 3|4	1.4	0.34	/	β_0 3|4	2.97	1.27	/
	Mass	-0.21	0.13	.	WingspanM	0.01	0.09	
	Avian hand-wing index	-0.31	0.07	***	FMoMean	0.38	0.11	***
	Nocturn	-2.28	0.59	***	Hostplant N	0.23	0.11	*
	Deciduous	-0.27	0.15	.	Hostplant Spe	0.16	0.11	
	Coniferous	0.38	0.15	*	HPG Bi	0.28	0.50	
	Woodland	0.17	0.14		HPG Th	-0.21	0.22	
	Shrub	0.55	0.13	***	HPG Sb	-0.32	0.23	
	Grassland	0.47	0.13	**	HPG Tr	0.22	0.45	
	Mountain meadows	0.23	0.24		HPG Li	0.02	0.36	
	Reed	0.08	0.33		SSI	0	0	
	Swamps	-0.84	0.28	**	AltVeg 'A'	0.06	1.12	
	Rocks	-0.14	0.21		AltVeg 'Mo'	-0.93	0.31	**
	Urban	0.83	0.14	***	AltVeg 'SupMed'	-0.79	0.46	.
	Spe. Diet	0.92	0.35	**	AltVeg 'ThMeMed'	1.37	0.37	***
	Spe. Foraging behav.	-0.08	0.4		Rarity '2'	0.86	0.29	**
	Spe. Diet strat	1.49	0.4	***	Rarity '3'	1.42	0.34	***
	Spe. Habitat	-0.14	0.38					
	Spe. Nest	-0.28	0.72					
	Spe. Mean	-2.26	1.63					

P-value: *** P < 0.001, ** P < 0.01, * P < 0.05, . P < 0.1, / P-value not applicable
Trait description: Spe. is the specialization, FMoMean is duration of yearly flight period, Hostplant N is hostplant specificity, HostplantSpe is hostplant specificity index, HPG is hostplang growth form (Bi: short herb, Th: tall herb, Sb: shrub, Tr: tree, Li: Liana), SSi is Species Specialization Index, AltVeg is altitudinal vegetation (A=Anthropogenic, Co: Foothill, TheMeMed: Thermo/Meso-Mediterranean, Med: Mediterranean, SupMed: Supra-Mediterranean , Mo: Montane, Asa: Alpine and Subalpine). For more details on traits see Appendix A.3

4 - Discussion

The absence of recommendations for the use of available databases that differ in terms of their quality and pertinence by public authorities or other institutions remains a serious problem for local conservation planning (Clare et al. 2019). The goal of this study was to test the suitability of different database sources that can be used by public stakeholders to map priorities for biodiversity stakes in SEAs and SCP. We found that the compilation of databases at the national scale is the most suitable procedure to apply SDM to a large number of species. For Aves and Papilionidae, the GBIF database provided the highest proportion of well-assessed SDM. We detected a significant overlap in species distributions in different database sources despite significant variability in the order of habitat suitability and similar spatial predictions for priority conservation areas. Finally, we showed that the composition of the species community used for priority conservation areas in three study sites were clearly not representative of the observed species communities in the original database (in terms of ecological traits).

Open-access biodiversity data provide a valuable source of information for conservation and land-use planning stakeholders; they contain vital information on species locations compared to expert knowledge and unshared datasets that are inaccessible to most users (Sousa-Baena et al., 2014; Meyer et al., 2015). Through their use and careful application of SDM, they contribute to the estimation of relative habitat suitability in a given study site (Baker et al., 2021). Our study showed however that SDM for a large majority of species observed locally requires their compilation on a national scale. Local and regional data are not suitable for model calibration but remain important for assessing the suitability of models in a local context. Indeed, this performance evaluation step is one of the most restrictive filters in the workflow we proposed, as evidenced by SDM's for butterflies in the T2 study site, where no species could be evaluted. The spatial extent of data collection can influence distribution modelling (Meyer, 2007), and our study emphasises this importance for SDM and the use of data available on national scale in France.

Different types of databases are constructed in different way – with opportunistic data collection or, in some cases, as part of a scientific monitoring scheme – allowing the use of as complementary data sources (Beck et al., 2013; Shirey et al., 2021). The three individual databases examined in this study are indeed complementary in that when they are combined they provide a suitable source for modelling the distribution of many species. Nevertheless, some groups commonly have a low number of data in such bases, e.g. Insecta (Troudet et al., 2017). The study of data with a fine spatial grain, as required for SEA and SCP (Guisan et al., 2013), reveals information gaps for more taxonomic groups at the global scale, e.g. Amphibians and Mammals (Witté and Touroult, 2017). The construction of an overall database at national scale is therefore the most appropriate way to have suitable data for SDM of different taxonomic groups.

SDM of Aves and Papilionidae species clearly revealed differences between the databases used for modelling, with differential impacts on the identification of conservation priorities. Indeed, species distribution overlap is high between database sources as indicated by the Schoener's D index, i.e. regardless of the data source species, are predicted in similar environments (Warren et al., 2008). However, Spearman's ranking of habitat suitability between data sources was highly variable, indicative that species' responses to environments are highly variable, as are the location of favourable habitats (Warren et al., 2008). Although the use of presence-absence data is advocated for SDM (Guillera-Arroita et al., 2015; Valavi et al., 2021b; Dubos et al., 2022), we showed that opportunist data from GBIF provided a greater number of well-assessed models at the local scale. Models using opportunist data with a target-group approach to generate pseudo-absences provides a sufficient quality of information on species distribution (Phillips et al., 2009; Barber et al., 2022) and can be correctly used in SCP (Sofaer et al., 2019; Baker et al., 2021). The lack of data at the regional scale, whatever the database, does not allow us to explain a better fit of models using GBIF data. Evaluating the models with a large proportion of opportunist data could however bias the evaluation, but only independent data were used, which provides robust validation of SDM (Matutini et al., 2021). Moreover, in contrast to Hermoso et al. (2015a), we found that mixing presence-only data with absence data increased the number of misjudged models. In addition to the use of the ROC curve (AUC) as a model evaluation metric by Hermoso et al. (2015a), the different results can be explained by different sampling biases between the two data types (Baker et al., 2022; Barber et al., 2022). Finally, the GBIF data seem to be more adapted to model the distribution of a large number of species.

The notion of "garbage in – garbage out" emphasises the critical importance of the quality of data (Sanders and Saxe, 2017), nevertheless, the examination of data suitability for conservation planning remains rare. In addition to the above issues our study revealed the importance of attention that should be paid to the representativeness of the species communities used in the models compared with the actual species communities observed in the study sites. This is particularly important in the light of the finding that there are marked differences between conservation priorities when different database sources are employed. Indeed, the number and composition of species in the community used can influence conservation priorities. Elsewhere it has been shown the difference will decline as the number of species increases (Kujala et al., 2018). The methodology tested in our study is based on a data-driven approach that attempts to use all available biodiversity data. This approach is data intensive, but is necessary to ensure the best representation of the observed local biodiversity. We revealed that such an approach can nevertheless induce a significant bias in the species community that is ultimately studied. Indeed, the prevalence of data affects the composition of the modelled species as well as the accuracy of the models and the evaluation of the species response (Fukuda and De Baets, 2016). Particular attention should thus be paid to the representativeness of the species communities used in the models in relation to the actual species communities observed in the territory under study.

Our study presents a workflow (Figure 2) for identifying biodiversity stakes using a data-driven approach from open-access database sources. Land-use planning stakeholders can use this workflow as a step towards the rationalization of data in order to reduce the biases mentioned above. The confrontation of the limits of such a workflow with the needs of stakeholders could illustrate how to precisely target new sources of database that should be collected according to the suitability of current databases for priority groups. Hermoso et al. (2015b) revealed that evaluation models using a new collection of field data does not necessarily reduce the problems of model uncertainty. However, other databases can be examined by stakeholders as well as other monitoring schemes (e.g. "PopAmphibien" for Reptilian and Amphibian populations in France http://lashf.org/popamphibien-2/) or negotiate the use of databases that are not yet shared. An important issue is thus the integration of stakeholders in the workflow we propose, and their appropriation of the procedure. This could be done by a form of participatory modelling (Lagabrielle et al., 2010; Lees et al., 2021), where stakeholders are consulted for issues and choices such as the species to be examined. In such participatory modelling it is important to avoid arbitrary choices that are neither reproducible nor representative of local diversity, but rather the result of administrative or political interest. Finally, it is currently recommended to use these tools to elaborate a more holistic approach to SCP (Cadotte and Tucker, 2018).

5 – Perspectives: operational implementation by SEA stakeholders

Spatially exhaustive and ecologically representative priority conservation areas are crucial for the elaboration of SEAs that aim to limit artificialization impacts as early as possible in the planning process. Empirical observations are major sources of information on biodiversity that are still rarely used by SEAs. The collection of open-access databases for SEA territories provides important but incomplete knowledge on species occurrence. Furthermore, their use is particularly interesting to help strategically direct inventory campaigns (especially for under sampled taxa and areas). These additional data would clearly improve the assessment of the SDMs suitability in local territories. What is also interesting here for stakeholders is that the process of filtering species and attributing them a score allows us to identify different groups of species in terms of their needs for additional data in order to undertake SDM.

The influence of database sources on the identification of priority conservation areas reveals the importance of examining their suitability. Thus, it is necessary to be prudent in the interpretation of biodiversity maps. The integration of local experts may help limit any misjudgements in the workflow procedure.

A data-driven approach that considers as many species as possible requires a large amount of data, biases the species communities considered and does not highlight species of particular interest as their threats and regulatory protections. It is therefore necessary to rationalize this approach, by integrating the needs and issues of local SEA stakeholders.

Acknowledgements

We thank Nicolas Dubos and Boris Leroy for our exchanges about SDM. This work would not have been possible without the help of many voluntary field who continually provide open-access data. We also thank Eric Durand for helpful discussion of this study.

Data, scripts, code, and supplementary information availability

Observation data from SINP and Vigie nature databases are not available for confidentiality reasons, but their link and the request process are detailed in Appendix A1. The data of GBIF are available online: https://doi.org/10.15468/dl.ry6uw7.

The data, R scripts, outputs of steps 1.2 to 4 of figure 2 are available online: https://doi.org/10.5281/zenodo.7883973.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Funding

This study was funded by “Naturalia Environnement” and “Association Nationale de la Recherche et de la Technologie” (grant number: 2020/0584).

References

Aiello-Lammens, M.E., Boria, R.A., Radosavljevic, A., Vilela, B., Anderson, R.P., 2015. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. Ecography 38, 541–545. https://doi.org/10.1111/ecog.01132

Almenar, J.B., Bolowich, A., Elliot, T., Geneletti, D., Sonnemann, G., Rugani, B., 2019. Assessing habitat loss, fragmentation and ecological connectivity in Luxembourg to support spatial planning. Landscape and Urban Planning 189, 335–351. https://doi.org/10.1016/j.landurbplan.2019.05.004

Anderson, R.P., 2003. Real vs. artefactual absences in species distributions: tests for Oryzomys albigularis (Rodentia: Muridae) in Venezuela. Journal of Biogeography 30, 591–605. https://doi.org/10.1046/j.1365-2699.2003.00867.x

Baker, D.J., Maclean, I.M.D., Goodall, M., Gaston, K.J., 2022. Correlations between spatial sampling biases and environmental niches affect species distribution models. Global Ecology and Biogeography 31, 1038–1050. https://doi.org/10.1111/geb.13491

Baker, D.J., Maclean, I.M.D., Goodall, M., Gaston, K.J., 2021. Species distribution modelling is needed to support ecological impact assessments. Journal of Applied Ecology 21–26. https://doi.org/10.1111/1365-2664.13782

Baker, M., Carter, J., Short, M., Jay, S., 2005. Strategic Environmental Assessment and Land Use Planning: An International Evaluation. Routledge.

Barber, R.A., Ball, S.G., Morris, R.K.A., Gilbert, F., 2022. Target-group backgrounds prove effective at correcting sampling bias in Maxent models. Diversity and Distributions 28, 128–141. https://doi.org/10.1111/ddi.13442

Beck, J., Ballesteros-Mejia, L., Nagel, P., Kitching, I.J., 2013. Online solutions and the ‘Wallacean shortfall’: what does GBIF contribute to our knowledge of species’ ranges? Diversity and Distributions 19, 1043–1050. https://doi.org/10.1111/ddi.12083

Bezombes, L., Kerbiriou, C., Spiegelberger, T., 2019. Do biodiversity offsets achieve No Net Loss? An evaluation of offsets in a French department. Biological Conservation 231, 24–29. https://doi.org/10.1016/j.biocon.2019.01.004

Bigard, C., Pioch, S., Thompson, J.D., 2017. The inclusion of biodiversity in environmental impact assessment: Policy-related progress limited by gaps and semantic confusion. Journal of Environmental Management 200, 35–45. https://doi.org/10.1016/j.jenvman.2017.05.057

Bigard, C., Thiriet, P., Pioch, S., Thompson, J.D., 2020. Strategic landscape-scale planning to improve mitigation hierarchy implementation: An empirical case study in Mediterranean France. Land Use Policy 90, 104286. https://doi.org/10.1016/j.landusepol.2019.104286

Boileau, J., Calvet, C., Pioch, S., Moulherat, S., 2022. Ecological equivalence assessment: The potential of genetic tools, remote sensing and metapopulation models to better apply the mitigation hierarchy. Journal of Environmental Management 305, 114415. https://doi.org/10.1016/j.jenvman.2021.114415

Boyce, M.S., Vernier, P.R., Nielsen, S.E., Schmiegelow, F.K.A., 2002. Evaluating resource selection functions. Ecological Modelling 157, 281–300. https://doi.org/10.1016/S0304-3800(02)00200-4

Boyd, R.J., Stewart, G., Pescott, O., 2023. Descriptive inference using large, unrepresentative nonprobability samples: An introduction for ecologists.

Brun, P., Thuiller, W., Chauvier, Y., Pellissier, L., Wüest, R.O., Wang, Z., Zimmermann, N.E., 2020. Model complexity affects species distribution projections under climate change. Journal of Biogeography 47, 130–142. https://doi.org/10.1111/jbi.13734

Bull, J.W., Gordon, A., Watson, J.E.M., Maron, M., 2016. Seeking convergence on the key concepts in 'no net loss' policy. Journal of Applied Ecology 53, 1686–1693. https://doi.org/10.1111/1365-2664.12726

Cadotte, M.W., Tucker, C.M., 2018. Difficult decisions: Strategies for conservation prioritization when taxonomic, phylogenetic and functional diversity are not spatially congruent. Biological Conservation 225, 128–133. https://doi.org/10.1016/j.biocon.2018.06.014

Canbek, G., 2022. Gaining insights in datasets in the shade of "garbage in, garbage out" rationale: Feature space distribution fitting. WIREs Data Mining and Knowledge Discovery 12, e1456. https://doi.org/10.1002/widm.1456

Carmona, C.P., Tamme, R., Pärtel, M., de Bello, F., Brosse, S., Capdevila, P., González-M., R., González-Suárez, M., Salguero-Gómez, R., Vásquez-Valderrama, M., Toussaint, A., 2021. Erosion of global functional diversity across the tree of life. Science Advances 7, eabf2675. https://doi.org/10.1126/sciadv.abf2675

CESBIO, 2022. Dynamic Habitat Indices of NDVI - Available data [WWW Document]. URL https://files.coinduf.eu/AFSfaJ2Kymt8iYWw/ (accessed 4.3.23).

CESBIO, 2021. Variables for Biodiversity – Theia [WWW Document]. URL https://www.theia-land.fr/en/ceslist/variables-for-biodiversity-sec/ (accessed 4.3.23).

Christensen, R., 2022. ordinal: Regression Models for Ordinal Data version 2022.11-16 from CRAN.

Church, R.L., Stoms, D.M., Davis, F.W., 1996. Reserve selection as a maximal covering location problem. Biological Conservation 76, 105–112. https://doi.org/10.1016/0006-3207(95)00102-6

Clare, J.D.J., Townsend, P.A., Anhalt-Depies, C., Locke, C., Stenglein, J.L., Frett, S., Martin, K.J., Singh, A., Van Deelen, T.R., Zuckerberg, B., 2019. Making inference with messy (citizen science) data: when are data accurate enough and how can they be improved? Ecological Applications 29, e01849. https://doi.org/10.1002/eap.1849

Costello, M.J., Wieczorek, J., 2014. Best practice for biodiversity data management and publication. Biological Conservation 173, 68–73. https://doi.org/10.1016/j.biocon.2013.10.018

Domisch, S., Friedrichs, M., Hein, T., Borgwardt, F., Wetzig, A., Jähnig, S.C., Langhans, S.D., 2019. Spatially explicit species distribution models: A missed opportunity in conservation planning? Diversity and Distributions 25, 758–769. https://doi.org/10.1111/ddi.12891

Dubos, N., Préau, C., Lenormand, M., Papuga, G., Monsarrat, S., Denelle, P., Louarn, M.L., Heremans, S., May, R., Roche, P., Luque, S., 2022. Assessing the effect of sample bias correction in species distribution models. Ecological Indicators 145, 109487. https://doi.org/10.1016/j.ecolind.2022.109487

Dupont, P., Demerges, D., Drouet, E., Luquet, G.Chr., 2013. Révision systématique, taxinomique et nomenclaturale des Rhopalocera et des Zygaenidae de France métropolitaine. Conséquences sur l'acquisition et la gestion des données d'inventaire (No. 19). MNHN-SPN.

Elvidge, C.D., Zhizhin, M., Ghosh, T., Hsu, F.-C., Taneja, J., 2021. Annual Time Series of Global VIIRS Nighttime Lights Derived from Monthly Averages: 2012 to 2019. Remote Sensing 13, 922. https://doi.org/10.3390/rs13050922

Essens, T., van Langevelde, F., Vos, R.A., Van Swaay, C.A.M., WallisDeVries, M.F., 2017. Ecological determinants of butterfly vulnerability across the European continent. Journal of Insect Conservation 21, 439–450. https://doi.org/10.1007/s10841-017-9972-4

Fahrig, L., 1997. Relative effects of habitat loss and fragmentation on population extinction. The Journal of wildlife management 61, 603–610. https://doi.org/10.2307/3802168

Ferrier, S., Pressey, R.L., Barrett, T.W., 2000. A new predictor of the irreplaceability of areas for achieving a conservation goal, its application to real-world planning, and a research agenda for further refinement. Biological Conservation 93, 303–325. https://doi.org/10.1016/S0006-3207(99)00149-4

Fletcher, R.J., Hefley, T.J., Robertson, E.P., Zuckerberg, B., McCleery, R.A., Dorazio, R.M., 2019. A practical guide for combining data to model species distributions. Ecology 100, e02710. https://doi.org/10.1002/ecy.2710

Fukuda, S., De Baets, B., 2016. Data prevalence matters when assessing species' responses using data-driven species distribution models. Ecological Informatics 32, 69–78. https://doi.org/10.1016/j.ecoinf.2016.01.005

Fundingsland Tetlow, M., Hanusch, M., 2012. Strategic environmental assessment: the state of the art. Impact Assessment and Project Appraisal 30, 15–24. https://doi.org/10.1080/14615517.2012.666400

Gargominy, O., Tercerie, S., Régnier, C., Ramage, T., Dupont, T., Daszkiewicz, P., Poncet, L., 2021. TAXREF v14, référentiel taxonomique pour la France : méthodologie, mise en oeuvre et diffusion, Rapport Patrinat. UMS PatriNat (OFB-CNRS-MNHN), Paris, France.

Gontier, M., Balfors, B., Mörtberg, U., 2006. Biodiversity in environmental assessment - current practice and tools for prediction. Environmental Impact Assessment Review 26, 268–286. https://doi.org/10.1016/j.eiar.2005.09.001

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R., Wintle, B.A., 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography 24, 276–292. https://doi.org/10.1111/geb.12268

Guillet, F., Le Floch, C., Julliard, R., 2019. Séquence Eviter-Réduire-Compenser : quelle biodiversité est visée par les mesures d'évitement ? Sciences Eaux et Territoires 1–8. https://doi.org/10.14758/SET-REVUE.2019.HS.07

Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat Suitability and Distribution Models: with Applications in R. Cambridge University Press.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P., Buckley, Y.M., 2013. Predicting species distributions for conservation decisions. Ecology Letters 16, 1424–1435. https://doi.org/10.1111/ele.12189

Hanson, J.O., Schuster, R., Morrell, N., Strimas-Mackey, M., Edwards, B.P.M., Watts, M.E., Arcese, P., Bennett, J., Possingham, H.P., 2021. prioritizr: Systematic Conservation Prioritization in R.

Hermoso, V., Kennard, M.J., Linke, S., 2015a. Assessing the risks and opportunities of presence-only data for conservation planning. Journal of Biogeography 42, 218–228. https://doi.org/10.1111/jbi.12393

Hermoso, V., Kennard, M.J., Linke, S., 2015b. Evaluating the costs and benefits of systematic data acquisition for conservation assessments. Ecography 38, 283–292. https://doi.org/10.1111/ecog.00792

Horváth, Z., Ptacnik, R., Vad, C.F., Chase, J.M., 2019. Habitat loss over six decades accelerates regional and local biodiversity loss via changing landscape connectance. Ecology Letters 22, 1019–1027. https://doi.org/10.1111/ele.13260

IGN, 2022. Géoservices - Téléchargement de données [WWW Document]. URL https://geoservices.ign.fr/telechargement (accessed 4.3.23).

Inglada, J., Vincent, A., Thierion, V., 2019. Theia OSO Land Cover Map 2019 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6538321

Iorio, É., Dusoulier, F., Soldati, F., Noël, F., Guilloton, J.-A., Doucet, G., Ponel, P., Dupont, P., Krieg-Jacquier, R., Chemin, S., Tillier, P., Touroult, J., 2022. Les Arthropodes terrestres dans les études d'impact : limites actuelles et propositions pour une meilleure prise en compte des enjeux de conservation. Naturae 43–99. https://doi.org/10.5852/naturae2022a4

IPBES, 2019. Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. IPBES secretariat. https://doi.org/10.5281/zenodo.6417333

Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K., Mooers, A.O., 2012. The global diversity of birds in space and time. Nature 491, 444–448. https://doi.org/10.1038/nature11631

Jomier, R., Solène, R., Dupont, P., 2018. National Scientific Validation for Data in the Information System on Nature and Landscapes (Système d'Information sur la Nature et les Paysages - SINP). Biodiversity Information Science and Standards e25881. https://doi.org/10.3897/biss.2.25881

Julliard, R., Jiguet, F., 2002. Un suivi intégré des populations d'oiseaux communs en France, in: Alauda. Presented at the Colloque Francophone Ornithologique, Limoges, France, pp. 137–147.

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M., 2021. Climatologies at high resolution for the earth's land surface areas. EnviDat 170122. https://doi.org/ 10.16904/envidat.228.v2.1

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M., 2017. Climatologies at high resolution for the earth's land surface areas. Scientific Data 4, 170122. https://doi.org/10.1038/sdata.2017.122

Kim, V., Ralphs, T., Guzelsoy, M., Mahajan, A., Reinhard, H., Hornik, K., Szymanski, C., Theussl, S., 2023. lpsymphony: Symphony integer linear programming solver in R.

Kujala, H., Moilanen, A., Gordon, A., 2018. Spatial characteristics of species distributions as drivers in conservation prioritization. Methods in Ecology and Evolution 9, 1121–1132. https://doi.org/10.1111/2041-210X.12939

Lagabrielle, E., Botta, A., Daré, W., David, D., Aubert, S., Fabricius, C., 2010. Modelling with stakeholders to integrate biodiversity into land-use planning - Lessons learned in Réunion Island (Western Indian Ocean). Environmental Modelling & Software 25, 1413–1427. https://doi.org/10.1016/j.envsoft.2010.01.011

Lees, C.M., Rutschmann, A., Santure, A.W., Beggs, J.R., 2021. Science-based, stakeholder-inclusive and participatory conservation planning helps reverse the decline of threatened species. Biological Conservation 260, 109194. https://doi.org/10.1016/j.biocon.2021.109194

Leroy, B., Delsol, R., Hugueny, B., Meynard, C.N., Barhoumi, C., Barbet-Massin, M., Bellard, C., 2018. Without quality presence–absence data, discrimination metrics such as TSS can be misleading measures of model performance. Journal of Biogeography 45, 1994–2002. https://doi.org/10.1111/jbi.13402

Lino, A., Fonseca, C., Rojas, D., Fischer, E., Pereira, M.J.R., 2019. A meta-analysis of the effects of habitat loss and fragmentation on genetic diversity in mammals. Mammalian Biology 94, 69–76. https://doi.org/10.1016/j.mambio.2018.09.006

Mac Nally, R., Duncan, R.P., Thomson, J.R., Yen, J.D.L., 2018. Model selection using information criteria, but is the "best" model any good? Journal of Applied Ecology 55, 1441–1444. https://doi.org/10.1111/1365-2664.13060

Margules, C.R., Pressey, R.L., 2000. Systematic conservation planning. Nature 405, 243–253. https://doi.org/10.1038/35012251

Mariton, L., Kerbiriou, C., Bas, Y., Zanda, B., Le Viol, I., 2022. Even low light pollution levels affect the spatial distribution and timing of activity of a "light tolerant" bat species. Environmental Pollution 305, 119267. https://doi.org/10.1016/j.envpol.2022.119267

Maron, M., Ives, C.D., Kujala, H., Bull, J.W., Maseyk, F.J.F., Bekessy, S., Gordon, A., Watson, J.E.M., Lentini, P.E., Gibbons, P., Possingham, H.P., Hobbs, R.J., Keith, D.A., Wintle, B.A., Evans, M.C., 2016. Taming a Wicked Problem: Resolving Controversies in Biodiversity Offsetting. BioScience 66, 489–498. https://doi.org/10.1093/biosci/biw038

Matsuki, K., Kuperman, V., Van Dyke, J.A., 2016. The Random Forests statistical technique: An examination of its value for the study of reading. Scientific Studies of Reading 20, 20–33. https://doi.org/10.1080/10888438.2015.1107073

Matutini, F., Baudry, J., Pain, G., Sineau, M., Pithon, J., 2021. How citizen science could improve species distribution models and their independent assessment. Ecology and Evolution 11, 3028–3039. https://doi.org/10.1101/2020.06.02.129536

Maxwell, S.L., Fuller, R.A., Brooks, T.M., Watson, J.E.M., 2016. Biodiversity: The ravages of guns, nets and bulldozers. Nature News 536, 143–145. https://doi.org/10.1038/536143a

MEDDE, GIS Sol, 2014. . Enveloppes des milieux potentiellement humides de la France métropolitaine. Notice d'accompagnement. Programme de modélisation des milieux potentiellement humides de France. Ministère d'Ecologie, du Développement Durable et de l'Energie, Groupement d'Intérêt Scientifique Sol.

Meyer, C., Kreft, H., Guralnick, R., Jetz, W., 2015. Global priorities for an effective information basis of biodiversity distributions. Nature Communications 6, 1–8. https://doi.org/10.1038/ncomms9221

Meyer, C.B., 2007. Does Scale Matter in Predicting Species Distributions? Case Study with the Marbled Murrelet. Ecological Applications 17, 1474–1483. https://doi.org/10.1890/06-1410.1

Middleton-Welling, J., Dapporto, L., García-Barros, E., Wiemers, M., Nowicki, P., Plazio, E., Bonelli, S., Zaccagno, M., Šašić, M., Liparova, J., Schweiger, O., Harpke, A., Musche, M., Settele, J., Schmucki, R., Shreeve, T., 2020. A new comprehensive trait database of European and Maghreb butterflies, Papilionoidea. Scientific Data 7, 351. https://doi.org/10.1038/s41597-020-00697-7

Morelli, F., Benedetti, Y., Callaghan, C.T., 2020. Ecological specialization and population trends in European breeding birds. Global Ecology and Conservation 22, e00996. https://doi.org/10.1016/j.gecco.2020.e00996

Moussus, J.-P., Lorin, T., Cooper, A., 2019. Guide pratique des papillons de France, Delachaux et Niestlé. ed. France.

Olden, J.D., Rooney, T.P., 2006. On defining and quantifying biotic homogenization. Global Ecology and Biogeography 15, 113–120. https://doi.org/10.1111/j.1466-822X.2006.00214.x

Phalan, B., Hayes, G., Brooks, S., Marsh, D., Howard, P., Costelloe, B., Vira, B., Kowalska, A., Whitaker, S., 2018. Avoiding impacts on biodiversity through strengthening the first stage of the mitigation hierarchy. Oryx 52, 316–324. https://doi.org/10.1017/S0030605316001034

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19, 181–197. https://doi.org/10.1890/07-2153.1

Ponder, W.F., Carter, G.A., Flemons, P., Chapman, R.R., 2001. Evaluation of Museum Collection Data for Use in Biodiversity Assessment. Conservation Biology 15, 648–657. https://doi.org/10.1046/j.1523-1739.2001.015003648.x

Pope, J., Bond, A., Morrison-Saunders, A., Retief, F., 2013. Advancing the theory and practice of impact assessment: Setting the research agenda. Environmental Impact Assessment Review 41, 1–9. https://doi.org/10.1016/j.eiar.2013.01.008

Pressey, R.L., Bottrill, M.C., 2008. Opportunism, Threats, and the Evolution of Systematic Conservation Planning. Conservation Biology 22, 1340–1345. https://doi.org/10.1111/j.1523-1739.2008.01032.x

Quétier, F., Regnery, B., Levrel, H., 2014. No net loss of biodiversity or paper offsets? A critical review of the French no net loss policy. Environmental Science & Policy 38, 120–131. https://doi.org/10.1016/j.envsci.2013.11.009

Rondinini, C., Wilson, K.A., Boitani, L., Grantham, H., Possingham, H.P., 2006. Tradeoffs of different types of species occurrence data for use in systematic conservation planning. Ecology Letters 9, 1136–1145. https://doi.org/10.1111/j.1461-0248.2006.00970.x

Sanders, H., Saxe, J., 2017. Garbage in, garbage out: how purportedly great ML models can be screwed up by bad data. Presented at the Blackhat, Las Vegas, USA, p. 6.

Schoener, T.W., 1968. The Anolis Lizards of Bimini: Resource Partitioning in a Complex Fauna. Ecology 49, 704–726. https://doi.org/10.2307/1935534

Sheard, C., Neate-Clegg, M.H.C., Alioravainen, N., Jones, S.E.I., Vincent, C., MacGregor, H.E.A., Bregman, T.P., Claramunt, S., Tobias, J.A., 2020. Ecological drivers of global gradients in avian dispersal inferred from wing morphology. Nature Communications 11, 2463. https://doi.org/10.1038/s41467-020-16313-6

Shirey, V., Belitz, M.W., Barve, V., Guralnick, R., 2021. A complete inventory of North American butterfly occurrence data: narrowing data gaps, but increasing bias. Ecography 44, 537–547. https://doi.org/10.1111/ecog.05396

Sofaer, H.R., Jarnevich, C.S., Pearse, I.S., Smyth, R.L., Auer, S., Cook, G.L., Edwards, T.C., Jr, Guala, G.F., Howard, T.G., Morisette, J.T., Hamilton, H., 2019. Development and Delivery of Species Distribution Models to Inform Decision-Making. BioScience 69, 544–557. https://doi.org/10.1093/biosci/biz045

Sousa-Baena, M.S., Garcia, L.C., Peterson, A.T., 2014. Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. Diversity and Distributions 20, 369–381. https://doi.org/10.1111/ddi.12136

Srisa-An, C., 2021. Guideline of Collinearity - Avoidable Regression Models on Time-series Analysis, in: 2021 2nd International Conference on Big Data Analytics and Practices (IBDAP). Presented at the 2021 2nd International Conference on Big Data Analytics and Practices (IBDAP), Bangkok, Thailand, pp. 28–32. https://doi.org/10.1109/IBDAP52511.2021.9552165

Stekhoven, D.J., Bühlmann, P., 2012. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112–118. https://doi.org/10.1093/bioinformatics/btr597

Støa, B., Halvorsen, R., Stokland, J.N., Gusarov, V.I., 2019. How much is enough? Influence of number of presence observations on the performance of species distribution models. Sommerfeltia 39, 1–28. https://doi.org/10.2478/som-2019-0001

Storchová, L., Hořák, D., 2018. Life-history characteristics of European birds. Global Ecology and Biogeography 27, 400–406. https://doi.org/10.1111/geb.12709

Tarabon, S., Bergès, L., Dutoit, T., Isselin-Nondedeu, F., 2019. Environmental impact assessment of development projects improved by merging species distribution and habitat connectivity modelling. Journal of Environmental Management 241, 439–449. https://doi.org/10.1016/j.jenvman.2019.02.031

Telenius, A., 2011. Biodiversity information goes public: GBIF at your service. Nordic Journal of Botany 29, 378–381. https://doi.org/10.1111/j.1756-1051.2011.01167.x

Tittensor, D.P., Walpole, M., Hill, S.L.L., Boyce, D.G., Britten, G.L., Burgess, N.D., Butchart, S.H.M., Leadley, P.W., Regan, E.C., Alkemade, R., Baumung, R., Bellard, C., Bouwman, L., Bowles-Newark, N.J., Chenery, A.M., Cheung, W.W.L., Christensen, V., Cooper, H.D., Crowther, A.R., Dixon, M.J.R., Galli, A., Gaveau, V., Gregory, R.D., Gutierrez, N.L., Hirsch, T.L., Höft, R., Januchowski-Hartley, S.R., Karmann, M., Krug, C.B., Leverington, F.J., Loh, J., Lojenga, R.K., Malsch, K., Marques, A., Morgan, D.H.W., Mumby, P.J., Newbold, T., Noonan-Mooney, K., Pagad, S.N., Parks, B.C., Pereira, H.M., Robertson, T., Rondinini, C., Santini, L., Scharlemann, J.P.W., Schindler, S., Sumaila, U.R., Teh, L.S.L., van Kolck, J., Visconti, P., Ye, Y., 2014. A mid-term analysis of progress toward international biodiversity targets. Science 346, 241–244. https://doi.org/10.1126/science.1257484

Toussaint, A., Brosse, S., Bueno, C.G., Pärtel, M., Tamme, R., Carmona, C.P., 2021. Extinction of threatened vertebrates will lead to idiosyncratic changes in functional diversity across the world. Nature Communications 12, 5162. https://doi.org/10.1038/s41467-021-25293-0

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., Legendre, F., 2017. Taxonomic bias in biodiversity data and societal preferences. Scientific Reports 7, 9132. https://doi.org/10.1038/s41598-017-09084-6

Tulloch, A.I.T., Gordon, A., Runge, C.A., Rhodes, J.R., 2019. Integrating spatially realistic infrastructure impacts into conservation planning to inform strategic environmental assessment. Conservation Letters 12, e12648. https://doi.org/10.1111/conl.12648

Tulloch, A.I.T., Sutcliffe, P., Naujokaitis-Lewis, I., Tingley, R., Brotons, L., Ferraz, K.M.P.M.B., Possingham, H., Guisan, A., Rhodes, J.R., 2016. Conservation planners tend to ignore improved accuracy of modelled species distributions to focus on multiple threats and ecological processes. Biological Conservation 199, 157–171. https://doi.org/10.1016/j.biocon.2016.04.023

Valavi, R., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G., 2021a. Modelling species presence-only data with random forests. Ecography 44, 1731–1742. https://doi.org/10.1111/ecog.05615

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., 2021b. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecological Monographs. https://doi.org/10.1002/ecm.1486

Velazco, S.J.E., Ribeiro, B.R., Laureto, L.M.O., De Marco Júnior, P., 2020. Overprediction of species distribution models in conservation planning: A still neglected issue with strong effects. Biological Conservation 252, 108822. https://doi.org/10.1016/j.biocon.2020.108822

Warren, D., Dinnage, R., 2022. ENMTools: Analysis of Niche Evolution using Niche and Distribution Models.

Warren, D.L., Glor, R.E., Turelli, M., 2008. Environmental Niche Equivalency Versus Conservatism: Quantitative Approaches to Niche Evolution. Evolution 62, 2868–2883. https://doi.org/10.1111/j.1558-5646.2008.00482.x

Whitehead, A.L., Kujala, H., Wintle, B.A., 2017. Dealing with Cumulative Biodiversity Impacts in Strategic Environmental Assessment: A New Frontier for Conservation Planning. Conservation Letters 10, 195–204. https://doi.org/10.1111/conl.12260

Wiemers, M., Chazot, N., Wheat, C.W., Schweiger, O., Wahlberg, N., 2020. A complete time-calibrated multi-gene phylogeny of the European butterflies. ZooKeys 938, 97–124. https://doi.org/10.3897/zookeys.938.50878

Wiemers, M., Chazot, N., Wheat, C.W., Schweiger, O., Wahlberg, N., 2019. A complete time-calibrated multi-gene phylogeny of the European butterflies - data [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3531555

Wilman, H., Belmaker, J., Simpson, J., de la Rosa, C., Rivadeneira, M.M., Jetz, W., 2014. EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. Ecology 95, 2027–2027. https://doi.org/10.1890/13-1917.1

Witté, I., Touroult, J., 2017. Identification et cartographie des zones de méconnaissance naturaliste à l'échelle nationale (métropole) à partir des données partagées (No. 2017–6). Muséum National d'Histoire Naturel (MNHN) - Service du Patrimoine Naturel (SPN), France.

Zambrano, J., Garzon-Lopez, C.X., Yeager, L., Fortunel, C., Cordeiro, N.J., Beckman, N.G., 2019. The effects of habitat loss and fragmentation on plant functional traits and functional diversity: what do we know so far? Oecologia 191, 505–518. https://doi.org/10.1007/s00442-019-04505-x

Zuckerberg, B., Huettmann, F., Frair, J., 2011. Proper Data Management as a Scientific Foundation for Reliable Species Distribution Modeling, in: Drew, C.A., Wiersma, Y.F., Huettmann, F. (Eds.), Predictive Species and Habitat Modeling in Landscape Ecology: Concepts and Applications. Springer, New York, NY, pp. 45–70. https://doi.org/10.1007/978-1-4419-7390-0_4

Zurell, D., Franklin, J., König, C., Bouchet, P., Dormann, C., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J.J., Leitão, P.J., Park, D., Peterson, A.T., Rapacciuolo, G., Schmatz, D., Schröder, B., Serra-Diaz, J.M., Thuiller, W., Yates, K., Zimmermann, N., Merow, C., 2020. A standard protocol for reporting species distribution models. https://doi.org/10.18452/22268

Appendix

Appendix A - Additional material and methods

A.1 - Collect and series of operation on species occurrences

Databases collection:

- Access to source databases:
    - French Natural and Landscape Information System (SINP, https://inpn.mnhn.fr), each of the three study sites has its own database: https://sinp-occitanie.fr/atlas/ for Lodévois-Larzac, https://obv-na.fr/ and https://observatoire-fauna.fr/ for Brocéliande, https://data.biodiversite-bretagne.fr for La Rochelle
    - GBIF: GBIF.org (04 October 2022) GBIF Occurrence Download https://doi.org/10.15468/dl.ry6uw7
    - French biodiversity monitoring scheme (Vigie Nature): https://www.vigienature.fr/, the database used are the STOC for Aves, the Vigie-Chiro for Chiroptera and Orthoptera, the Vigie-Flore for Flora, the STELI for Odonata, the STERF for Papilionidae
- Period of 10 years, from 01/01/2010 to 31/12/2020
- Extent: continental France (without island, Corsica and overseas regions)
- Spatial accuracy of 0 to 50m

Standardization of species names : French taxonomic reference TAXREF.V14 (Gargominy et al., 2021).

- Keep only the species and subspecies
- Groups data at the species level

Taxonomic filtering: from TAXREF.V14

- Keep only species in nine taxonomic groups selected: Amphibia, Aves, Chiroptera, Flora, Mammalia aptera (Mammalia), Orthoptera, Odonata, Papilionidae and Reptilia
- Delete species not present in France so keep: “Present”, “Endemic”, “Subendemic”, “Cryptogenic”, “Introduced”, “Invasive introduced"

Data transformation and filtering:

- Transformation data into occurrence, i.e. transform data into presence only
- Geographical filtering for reduce sampling biases and delete duplicate data: spatial thinning at a distance of 50m with “spThin” package (Aiello-Lammens et al., 2015)

Identify the presence of species into each study site: at least five observations from all databases in the study site to certify the presence of the species

Standardization of the observation period between databases: for Aves filtering observations between 1 March and 15 July corresponding to the STOC period.

A.2 - Environmental variables selected for Species Distribution Models (SDM) of Aves and Papilionidae species. Variables used for SDM are the variables with collinearity removed (Pearson >0.7).

Type of variables		Variables	Database	Year	Raw precision	Source	Variables used for SDM
Biogeography	Climatic	19 bioclimatic variables	Chelsa	1981-2010	1 km	Karger et al., 2021, 2017	Bio 2 (temp. range), Bio 5 (temp. warmest month), Bio 6 (min. temp. coldest month), Bio 9 (temp. driest quarter), Bio 13 (prec. wettest month), Bio 14 (prec. driest month), Bio 15(prec. seasonality)
	Elevation	Elevation, mean in a buffer of fifty meters	BD ALTI, IGN	2014	25 m	IGN, 2022	Elevation
		Slope, mean in a buffer of 50 meters					Slope
		Topographic Position Index (TPI), mean in a buffer of 50 meters					TPI
	Wetland	Area of wetland potential: very high in a buffer of 50, 500, 2000 meters	INRAE Wetland potential	2014	50 m	Berthier et al., 2014	Wetland in buffer of 50 and 500 m
	Waterway	Linear meters of waterway in a buffer of 50, 500, 2000 meters	BD TOPO, IGN	2022	Vector	IGN, 2022	Waterway in buffer of 50 and 500m
Human occupancy	Land-use	Area of artificial, conifer forest, crop, deciduous forest, land, lawn, mineral, prairie, water in buffer of 50, 500 and 2000 meters	OSO Land Cover	2018	25 m	Inglada et al., 2019	Artificial, conifer, crops, deciduous, fruit culture, land, mineral, prairie, water in a buffer of 50 and 500 m Lawn in a buffer of 500 m, Sand in a buffer of 50, 500, 2000 m
	Hedge line	Linear meters of hedge in a buffer of 50, 500, 2000 meters	BD TOPO, IGN	2022	Vector	IGN, 2022	Hedge line in buffer of 50, 500 m
	NDVI mean	NDVI mean in buffer of 50, 500, 2000 meters	DHI NDVI	2022	10 m	CESBIO, 2021, 2022	DHI NDVI in buffer of 50 and 500m
Pollution / Fragmentation	Light pollution	Mean value in buffer of 500 meters of night time light average masked of 2015	VIIRS night time lights	2015	500 m	Elvidge et al., 2021	Night time light
	Road fragmentation	Linear meters of fragmenting of main road in a buffer of 50, 500, 2000 meters	BD TOPO, IGN	2022	Vector	IGN, 2022	Main road in a buffer of 50, 500, 2000 m

A.3 - Species traits of Aves and Papilionidae. Missing traits were imputed by considering evolutionary relationships following the process of Carmona et al. (2021) with the help the R scripts of Toussaint et al. (2021).

	Traits	Description	Reference
Aves	Mass	Body mass using data from males, females, and/or unspecified adults in g	Storchová and Hořák, 2018
	HWI	Avian hand-wing index (HWI), an estimate of wing shape as a proxy for dispersal ability in birds	Sheard et al., 2020
	Nocturn	Nocturnal activity	Wilman et al., 2014
	Habitats (Deciduous Coniferous, Woodland, Shrub, Grassland, Mountain meadows, Reed, Swamps, Rocks, Urban)	Species occupies habitat in breeding area, 1: yes; 0: no	Storchová and Hořák, 2018
	Spe. Diet, Spe. Foraging behav., Spe. Diet strat, Spe. Hab., Spe. Nest, Spe. Mean	Specialization respectively of diet, foraging behaviour, foraging substrate, habitat, nesting site and mean of five specializations	Morelli et al., 2020
Papilionidae	Wingspan	Average length of male wingspan in mm	Middleton-Welling et al., 2020
	FMoMean	Average duration of yearly flight period	
	HostplantN	Hostplant specificity, number of hostplants	
	HostplantSpe	Hostplant specificity index	
	HPG	Hostplant growth form: Short herb/grass (<1 m) (Bi), tall herb/grass (>1 m) (Th), shrub (Sb), tree (Tr), liana (Li)	
	SSI	Species Specialization Index	Essens et al., 2017
	AltVeg	Altitudinal vegetation: A=Anthropogenic, Co=Foothill, TheMeMed=Thermo/Meso-Mediterranean, Med=Mediterranean, SupMed=Supra-Mediterranean, Mo=Montane, ASa=Alpine and Subalpine	Dupont et al., 2013
	Rarity	Rarity in France between 1 very rare to 3 common species	In Moussus et al., 2019 from oreina.org and faune.france.org

The phylogenetic trees used to consider evolutionary relationships were from Jetz et al. (2012) and Wiemers et al. (2019, 2020) for the Aves and Papilionidae, respectively.

Appendix B - Supplementary results

B.1 - Description of observation databases (Local SINP, GBIF, Vigie Nature) available nine taxonomic groups in the three study sites (T1: Lodévois-Larzac, T2: Brocéliande, T3: La Rochelle) for "local" (buffer zone 10 km), "Regional" and "France" (continental France) scales. Statistics: "nSp" is the number of species with at least three observations on the study site within a 10 km buffer; "pSp<15obs", "pSp15.50obs" and "pSp>50obs" are the percentage of species with <15, 15 to 50 and >50 observations respectively and "-" represents no data available.

B.1.A - Local scale (10km buffer zone).

Taxa	Statistc	Local											
		GBIF			SINP			Vigie Nature			All databases combined		
		T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3
Amphibia	nSp	11	1	11	13	2	0	No program			14	2	11
	pSp<15obs	82%	100%	100%	23%	50%	-				29%	50%	100%
	pSp15.50obs	18%	0%	0%	15%	50%	-				7%	50%	0%
	pSp.>50obs	0%	0%	0%	62%	0%	-				64%	0%	0%
Aves	nSp	100	73	77	197	94	82	114	155	101	205	171	132
	pSp<15obs	87%	93%	95%	37%	96%	85%	56%	50%	54%	37%	47%	59%
	pSp15.50obs	10%	6%	4%	23%	3%	15%	32%	17%	20%	22%	22%	13%
	pSp.>50obs	3%	1%	1%	40%	1%	0%	12%	33%	27%	41%	31%	28%
Chiroptera	nSp	4	0	5	21	6	18	25	22	24	26	22	26
	pSp<15obs	100%	-	100%	10%	83%	50%	12%	41%	21%	12%	41%	27%
	pSp15.50obs	0%	-	0%	38%	0%	44%	16%	36%	46%	19%	36%	34%
	pSp.>50obs	0%	-	0%	52%	17%	6%	72%	23%	33%	69%	23%	39%
Flora	nSp	786	476	639	1703	1335	843	286	177	136	1844	1460	1093
	pSp<15obs	81%	85%	81%	52%	48%	74%	100%	100%	100%	52%	50%	72%
	pSp15.50obs	16%	14%	17%	30%	25%	14%	0%	0%	0%	28%	24%	17%
	pSp.>50obs	3%	1%	3%	18%	27%	12%	0%	0%	0%	20%	25%	11%
Mammalia	nSp	10	1	14	23	7	30	No program			23	7	30
	pSp<15obs	100%	100%	100%	78%	86%	60%				74%	86%	57%
	pSp15.50obs	0%	0%	0%	17%	14%	20%				22%	14%	20%
	pSp.>50obs	0%	0%	0%	4%	0%	20%				4%	0%	23%
Odonata	nSp	38	5	23	57	1	0	0	1	24	57	7	30
	pSp<15obs	74%	100%	100%	28%	100%	-	-	100%	100%	28%	100%	97%
	pSp15.50obs	26%	0%	0%	37%	0%	-	-	0%	0%	30%	0%	3%
	pSp.>50obs	0%	0%	0%	35%	0%	-	-	0%	0%	42%	0%	0%
Orthoptera	nSp	53	3	24	91	8	0	25	19	19	98	25	34
	pSp<15obs	92%	100%	100%	71%	100%	-	24%	58%	16%	56%	68%	53%
	pSp15.50obs	8%	0%	0%	23%	0%	-	24%	32%	52%	18%	24%	29%
	pSp.>50obs	0%	0%	0%	6%	0%	-	52%	10%	32%	26%	8%	18%
Papilionidae	nSp	115	15	33	154	11	0	0	0	50	154	18	52
	pSp<15obs	60%	100%	94%	21%	100%	-	-	-	88	21%	100%	73%
	pSp15.50obs	30%	0%	6%	19%	0%	-	-	-	12	18%	0%	27%
	pSp.>50obs	10%	0%	0%	60%	0%	-	-	-	0	60%	0%	0%
Reptilia	nSp	18	2	7	21	4	0	No program			22	4	7
	pSp<15obs	67%	100%	100%	24%	100%	-				27%	75%	100%
	pSp15.50obs	28%	0%	0%	33%	0%	-				32%	25%	0%
	pSp.>50obs	6%	0%	0%	43%	0%	-				41%	0%	0%

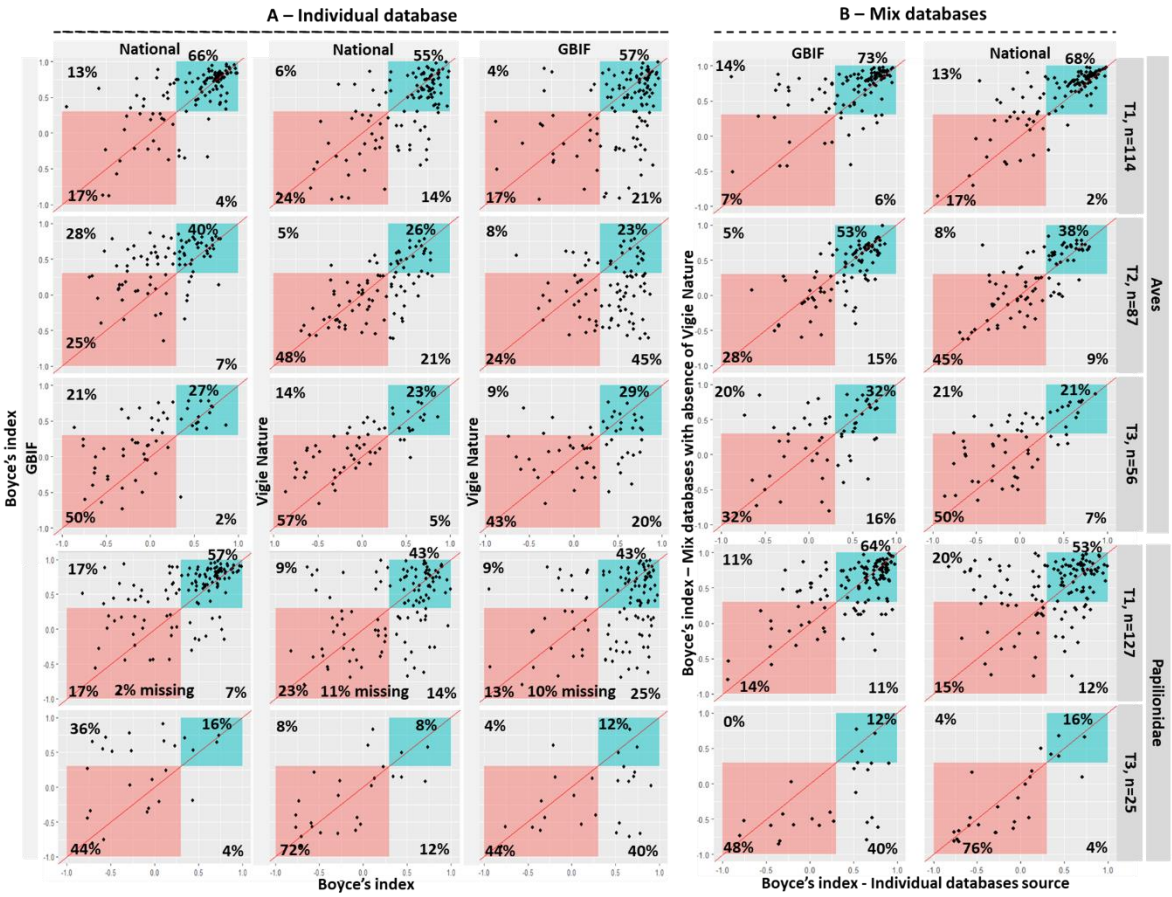
B.1.A - Regional scale.

Taxa	Statistic	Regional								
		GBIF			Vigie Nature			All databases combined		
		T1	T2	T3	T1	T2	T3	T1	T2	T3
Amphibia	nSp	13	2	11	No program			14	2	11
	pSp<15obs	8%	0%	73%				7%	0%	73%
	pSp15.50obs	23%	50%	18%				14%	50%	18%
	pSp.>50obs	69%	50%	9%				79%	50%	9%
Aves	nSp	198	159	120	182	143	120	205	166	132
	pSp<15obs	21%	21%	40%	31%	29%	32%	12%	16%	19%
	pSp15.50obs	12%	12%	22%	21%	15%	27%	11%	10%	28%
	pSp.>50obs	67%	67%	38%	48%	56%	41%	77%	74%	53%
Chiroptera	nSp	13	13	5	25	22	24	26	22	26
	pSp<15obs	100%	100%	100%	0%	0%	8%	0%	0%	15%
	pSp15.50obs	0%	0%	0%	4%	0%	17%	8%	0%	15%
	pSp.>50obs	0%	0%	0%	96%	100%	75%	92%	100%	69%
Flora	nSp	1428	1093	825	825	657	413	1844	1158	1093
	pSp<15obs	43%	38%	47%	95%	94%	94%	32%	40%	47%
	pSp15.50obs	19%	19%	20%	5%	6%	6%	26%	19%	23%
	pSp.>50obs	38%	49%	33%	0%	0%	0%	42%	41%	31%
Mammalia	nSp	23	6	21	No program			23	6	30
	pSp<15obs	39%	33%	62%				30%	33%	50%
	pSp15.50obs	26%	0%	33%				22%	0%	23%
	pSp.>50obs	35%	67%	5%				48%	67%	27%
Odonata	nSp	56	7	28	56	7	27	57	7	30
	pSp<15obs	18%	14%	64%	77%	71%	100%	9%	14%	53%
	pSp15.50obs	30%	14%	25%	23%	29%	0%	21%	14%	37%
	pSp.>50obs	52%	71%	11%	0%	0%	0%	70%	72%	10%
Orthoptera	nSp	93	25	26	25	19	19	98	25	34
	pSp<15obs	42%	24%	58%	4%	0%	11%	30%	4%	27%
	pSp15.50obs	39%	40%	42%	4%	0%	16%	22%	0%	29%
	pSp.>50obs	19%	36%	0%	92%	100%	74%	48%	96%	44%
Papilionidae	nSp	144	18	43	132	18	51	154	18	52
	pSp<15obs	12%	0%	37%	42%	11%	55%	7%	0%	34%
	pSp15.50obs	19%	11%	42%	30%	22%	41%	13%	6%	31%
	pSp.>50obs	69%	89%	21%	28%	67%	4%	81%	94%	35%
Reptilia	nSp	20	4	7	No program			22	4	7
	pSp<15obs	20%	0%	14%				18%	0%	14%
	pSp15.50obs	30%	25%	57%				28%	25%	57%
	pSp.>50obs	50%	75%	29%				59%	75%	29%

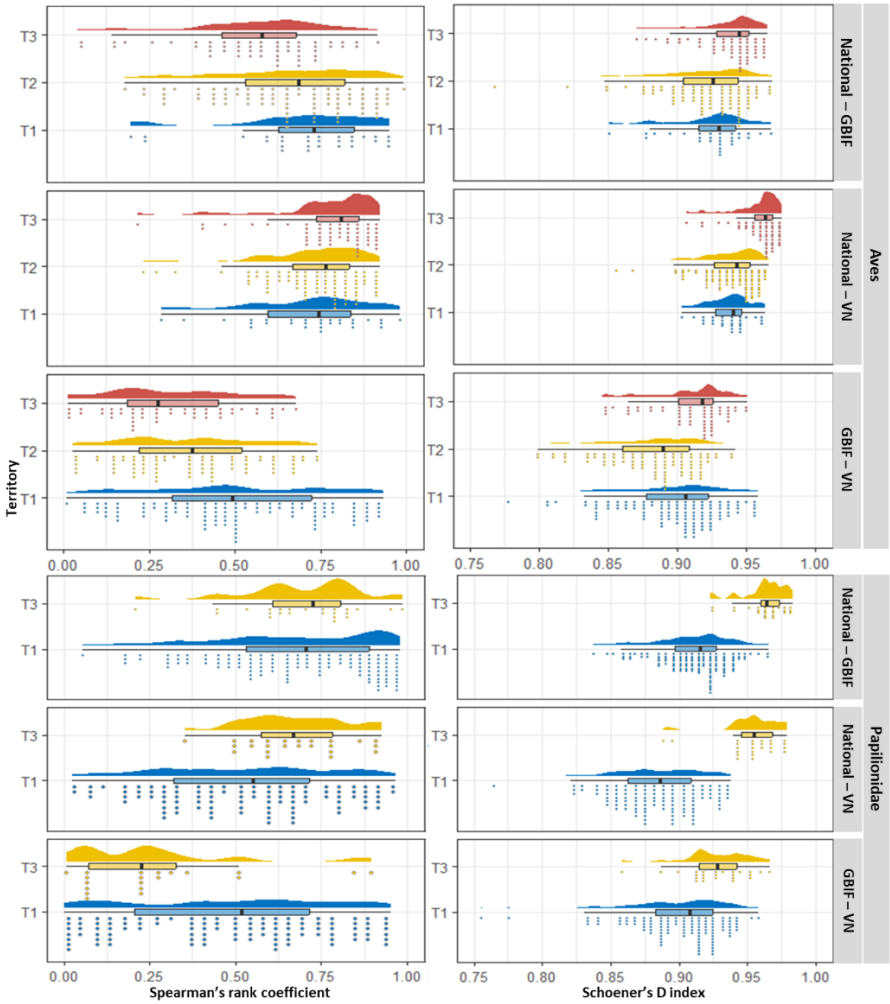
B.1.C - France scale.

Taxa	Statistic	GBIF T1	GBIF T2	GBIF T3	France Vigie Nature T1	France Vigie Nature T2	France Vigie Nature T3	All databases combined T1	All databases combined T2	All databases combined T3
Amphibia	nSp	14	2	11	No program			14	2	11
	pSp<15obs	7%	0%	0%				0%	0%	0%
	pSp15.50obs	0%	0%	9%				7%	0%	9%
	pSp.>50obs	93%	100%	91%				93%	100%	91%
Aves	nSp	204	170	131	201	170	132	205	171	132
	pSp<15obs	2%	1%	1%	8%	9%	4%	1%	0%	0%
	pSp15.50obs	3%	1%	1%	11%	9%	7%	1%	1%	0%
	pSp.>50obs	95%	98%	98%	81%	82%	89%	98%	99%	100%
Chiroptera	nSp	19	19	21	25	22	25	26	22	26
	pSp<15obs	90%	89%	90%	0%	0%	0%	0%	0%	4%
	pSp15.50obs	10%	11%	10%	0%	0%	0%	4%	0%	0%
	pSp.>50obs	0%	0%	0%	100%	100%	100%	96%	100%	96%
Flora	nSp	1573	1261	995	1336	1105	848	1844	1460	1093
	pSp<15obs	25%	20%	17%	68%	61%	53%	19%	14%	16%
	pSp15.50obs	18%	18%	16%	19%	23%	27%	21%	17%	16%
	pSp.>50obs	57%	62%	67%	13%	16%	20%	60%	69%	68%
Mammalia	nSp	23	6	28	No program			23	7	30
	pSp<15obs	13%	16%	18%				9%	14%	23%
	pSp15.50obs	17%	17%	28%				13%	29%	17%
	pSp.>50obs	70%	67%	54%				78%	57%	60%
Odonata	nSp	57	7	30	57	7	30	57	7	30
	pSp<15obs	2%	0%	0%	11%	0%	0%	0%	0%	0%
	pSp15.50obs	10%	14%	0%	26%	0%	13%	3%	0%	0%
	pSp.>50obs	88%	86%	100%	63%	100%	87%	97%	100%	100%
Orthoptera	nSp	95	25	34	25	19	19	98	25	34
	pSp<15obs	22%	12%	6%	4%	0%	0%	15%	0%	0%
	pSp15.50obs	28%	20%	21%	0%	0%	0%	25%	4%	9%
	pSp.>50obs	50%	68%	73%	96%	100%	100%	60%	96%	91%
Papilionidae	nSp	147	18	52	149	18	52	154	18	52
	pSp<15obs	4%	0%	0%	13%	5%	0%	1%	0%	0%
	pSp15.50obs	8%	0%	0%	18%	6%	0%	7%	0%	0%
	pSp.>50obs	88%	100%	100%	69%	89%	100%	92%	100%	100%
Reptilia	nSp	21	4	7	No program			22	4	7
	pSp<15obs	0%	0%	0%				4%	0%	0%
	pSp15.50obs	14%	0%	0%				5%	0%	0%
	pSp.>50obs	86%	100%	100%				91%	100%	100%

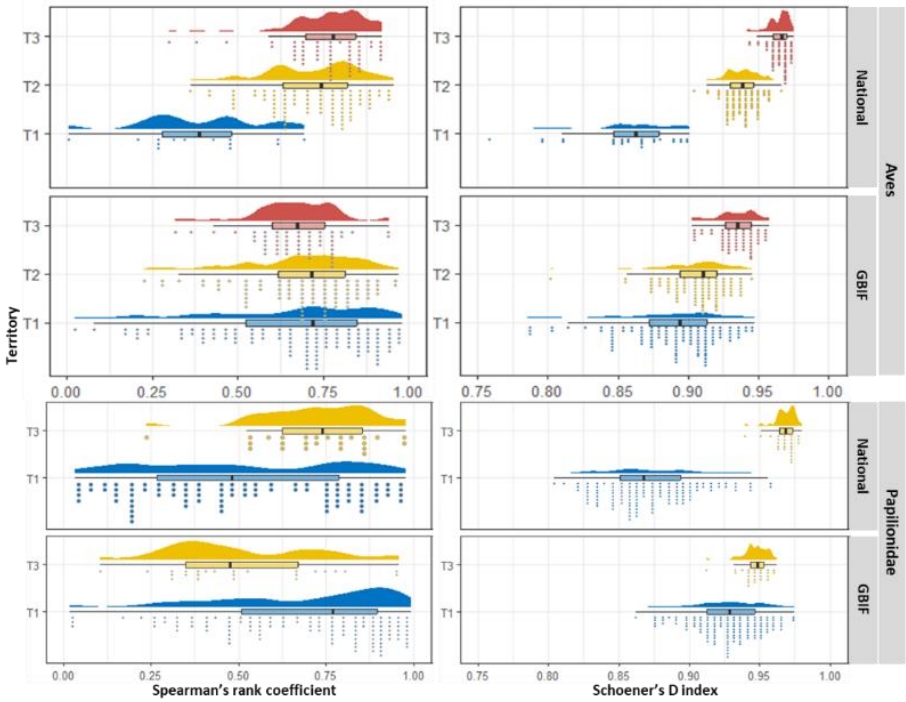
B.2 - SDM evaluation of Aves and Papilionidae taxonomic groups for (A) three individual database sources and (B) between SDM using different pseudo-absence sources. Each graphic is a pair of databases. Each point is an individual species that is listed in both databases, some species are assessed only in one database (missing). A Boyce's index threshold of 0.3 is used to assess the proportion of satisfactory and unsatisfactory models conjointly for each of the two database sources. The blue quadrant contains species whose SDMs are evaluated in a satisfactory manner in both data bases (Boyce index > 0.3 for both databases) ; the red quadrant contains the species for which SDMs are evaluated in an unsatisfactory manner in both databases (Boyce index < 0.3 for both databases). In the top-left quadrant and the bottom-right quadrants, SDM model evaluation was satisfactory for only one database (on the in y and x axes respectively).



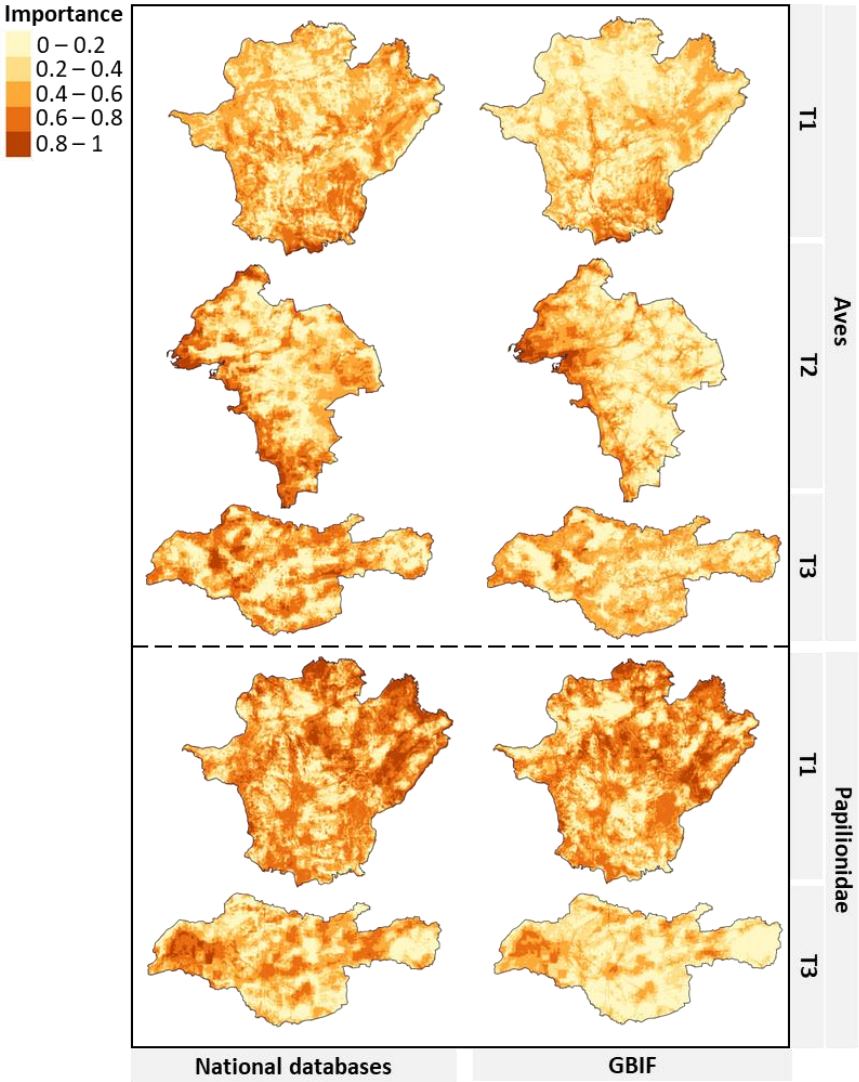
B.3 - Overlap between SDM from individual database sources. VN is Vigie Nature database.



B.4 - Overlap between SDM from individual database sources and SDM from mixed databases (i.e. using absence from Vigie Nature).



B.5 - Maps of conservation planning area for Aves and Papilionidae of tree study sites (T1, T2, T3) from different individual database source (National databases, GBIF, Vigie Nature).



B.6 - Evaluation of the quality of the full model (Table 5) compared to null model using AIC.

Model	Aves	Papilionidae
Full	3057	1322
Null	3357	1574