Potentialité de présence des espèces à partir de leurs relations avec les habitats - Rapport d'analyse

Tiphaine Ouisse, Florian Barnier

UMS PATRIMOINE NATUREL
CENTRE D'EXPERTISE ET DE DONNÉES
OFB • CNRS • MNHN

Potentialité de présence des espèces à partir de leurs relations avec les habitats

Rapport d'analyse

Tiphaine OUISSE & Florian BARNIER

Novembre 2020

UNITÉ MIXTE DE SERVICE

PATRIMOINE NATUREL

Nom du Programme/Projet : Relations espèces-habitats

Convention : Ce rapport est produit dans la cadre de la convention entre l'UMS PatriNat et la Direction des infrastructures de transports du Ministère de la transition écologique.

Chef de projet : Florian Barnier

Chargée de mission : Tiphaine Ouisse

Relecture : Isabelle Witté, Théo Duquesne

Référence du rapport conseillée : Ouisse, T. & Barnier, F. 2020. *Potentialité de présence des espèces à partir de leurs relations avec les habitats*. *Rapport d'analyse*. UMS PatriNat OFB-CNRS-MNHN, Paris. 47pp.

Crédits couverture (Images INPN)

Hêtraies nitrophiles médio-européennes (code EUNIS G1.63) © O. Roquinarc'h

Huîtrier-pie, *Haematopus ostralegus* Linnaeus, 1758 © P. Gourdain

Salamandre tachetée, *Salamandra salamandra* Linnaeus, 1758 © P. Gourdain

Gentiane jaune, *Gentiana lutea* Linnaeus, 1753 © J. Laignel

Boarmie du chêne, *Hypomecis roboraria* Denis & Schiffermüller, 1775 © D. Morel

L'UMS Patrimoine naturel - PatriNat

Centre d'expertise et de données sur la nature

Depuis janvier 2017, l'Unité Mixte de Service Patrimoine naturel assure des missions d'expertise et de gestion des connaissances pour ses trois tutelles, que sont le Muséum national d'Histoire naturelle (MNHN), l'Office français pour la biodiversité (OFB) et le Centre national de la recherche scientifique (CNRS).
Son objectif est de fournir une expertise fondée sur la collecte et l'analyse de données de la biodiversité et de la géodiversité présentes sur le territoire français, et sur la maitrise et l'apport de nouvelles connaissances en écologie, sciences de l'évolution et anthropologie. Cette expertise, établie sur une approche scientifique, doit contribuer à faire émerger les questions et à proposer les réponses permettant d'améliorer les politiques publiques portant sur la biodiversité, la géodiversité et leurs relations avec les sociétés et les humains.

En savoir plus : patrinat.fr

Co-directeurs :
Laurent PONCET, directeur en charge du centre de données
Julien TOUROULT, directeur en charge des rapportages et de la valorisation

Inventaire National du Patrimoine Naturel

Porté par l'UMS PatriNat, cet inventaire est l'aboutissement d'une démarche qui associe scientifiques, collectivités territoriales, naturalistes et associations de protection de la nature, en vue d'établir une synthèse sur le patrimoine naturel en France. Les données fournies par les partenaires sont organisées, gérées, validées et diffusées par le MNHN. Ce système est un dispositif clé du Système d'Information sur la Nature et les Paysages (SINP) et de l'Observatoire National de la Biodiversité (ONB).
Afin de gérer cette importante source d'informations, le Muséum a construit une base de données permettant d'unifier les données à l'aide de référentiels taxonomiques, géographiques et administratifs. Il est ainsi possible d'accéder à des listes d'espèces par commune, par espace protégé ou par maille de 10x10 km. Grâce à ces systèmes de référence, il est possible de produire des synthèses, quelle que soit la source d'information.
Ce système d'information permet de consolider des informations qui étaient jusqu'à présent dispersées. Il concerne la métropole et l'outre-mer, aussi bien sur la partie terrestre que marine. C'est une contribution majeure pour la connaissance naturaliste, l'expertise, la recherche en macroécologie et l'élaboration de stratégies de conservation efficaces du patrimoine naturel.

En savoir plus : inpn.mnhn.fr

SOMMAIRE

1 Introduction 5
1.1 Contexte de l'étude 5
1.2 Objectifs et démarche de l'étude 6
1.3 Organisation du rapport 7
2 Méthodes 7
2.1 Démarche d'analyse 7
2.2 Référentiels et données utilisées 9
2.2.1 Référentiels nationaux : TAXREF et HABREF 9
2.2.2 Base de connaissance des relations espèces – habitats 9
2.2.3 Autres jeux de données de référence sur la connaissance des espèces 10
2.2.4 Données d'inventaires pour la calibration du modèle 11
2.2.5 Bases de données de l'INPN et carroyage 11
2.3 Mise en forme du jeu de données utilisé pour l'analyse statistique 12
2.4 Analyse statistique 17
3 Résultats 21
3.1 Échelle spatiale d'analyse et résolution 21
3.2 Échelle temporelle d'analyse 22
3.3 Caractéristiques du modèle sélectionné 23
3.4 Détermination graphique des seuils de probabilité de présence 25
3.5 Diagnostic de la qualité de discrimination du modèle 26
3.5.1 Matrice de confusion et répartition des espèces selon les groupes taxonomiques 27
3.5.2 Critère de patrimonialité 28
3.5.3 Critère de méconnaissance locale du taxon 28
3.5.4 Critère de présence de l'espèce dans la maille 29
3.5.5 Conclusion sur le diagnostic 30
3.5.6 Cas des espèces mal attribuées 31
3.6 Test avec les données du Contournement Ouest de Nîmes 34
4 Limites et perspectives 39
4.1 Limites 39
4.2 Pistes d'amélioration 40
5 Bibliographie 42
6 Annexes 43

1 Introduction

1.1 Contexte de l'étude

Cette étude constitue un livrable de la convention passée entre la Direction des Infrastructures de Transport (DIT) du Ministère de la Transition Ecologique (MTE) et l'UMS PatriNat (OFB – CNRS – MNHN). Le premier objectif de cette convention était de construire une méthodologie permettant de caractériser les relations entre les espèces et les habitats qu'elles fréquentent. Les réflexions et la conception de cette méthodologie sont détaillées dans les rapports Jeanmougin et al. (2014) et Legros et al. (2016), et ont donné naissance à la base de connaissance sur les relations espèces – habitats. Le second objectif de la convention était de construire une application permettant de déterminer la potentialité de présence des espèces selon les habitats présents sur un site. Cette application a pour but premier de mieux cibler le choix des espèces à investiguer aux stades antérieurs des études d'impacts (études d'opportunité ou études préalables à l'enquête publique). Ce rapport s'intègre dans cette seconde étape et présente les analyses effectuées pour la construction d'un modèle prédictif de la potentialité de présence des espèces en fonction des habitats présents sur un site.

Afin de construire un modèle de potentialité de présence des espèces, des données d'inventaires effectués sur le terrain ont été utilisées. Ces données sont issues d'inventaires IPE et IQE (*cf.* paragraphe 2.2.4) réalisés sur des sites métropolitains dans le cadre de conventions avec des partenaires privés. En pratique, la comparaison des listes d'espèces potentiellement présentes d'après la base de connaissance des relations espèces – habitats et des listes d'espèces effectivement inventoriées permet d'évaluer la justesse de la prédiction. Le but de la modélisation est de maximiser le nombre d'espèces bien prédites tout en conservant un nombre total d'espèces prédites réaliste. Cette comparaison avec des données réelles permet aussi de définir un indice de qualité ou de confiance de la prédiction.

Pour résumer le travail mené dans le cadre de cette convention au cours des dernières années, une hiérarchisation des liens entre les espèces et les habitats qu'elles occupent (classifiés selon la typologie EUNIS[1]) a été établie, permettant de caractériser la force de la relation le long d'un gradient entre l'utilisation non sélective d'un habitat par une espèce à une espèce strictement spécialiste d'un habitat particulier (*cf.* figure 1 et tableau 1).

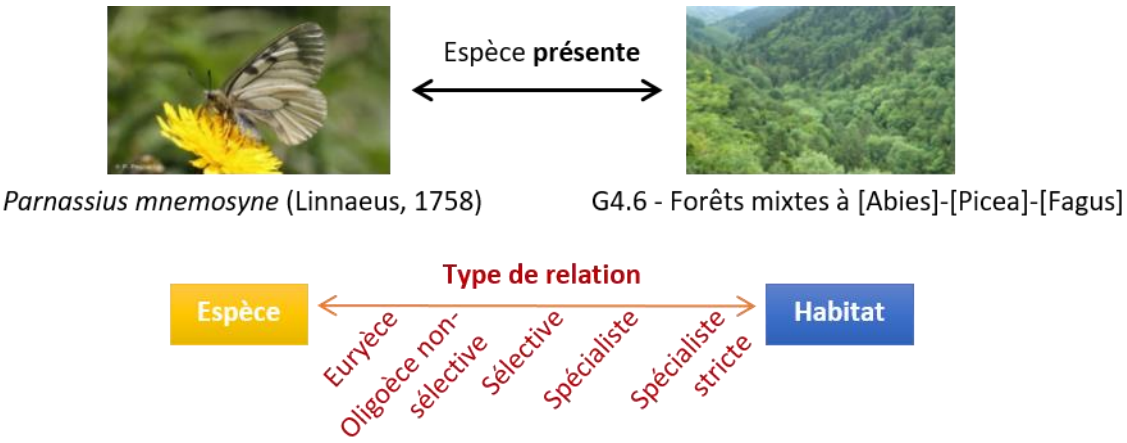


Fig. 1 Présentation de la hiérarchisation des liens existant entre les espèces et les habitats qu'elles occupent et exemple de la relation qui lie le Semi-Apollon à l'habitat G4.6 (Forêts mixtes à Sapins, Épicéas, Hêtres)

[1] La typologie EUNIS (Système d'Information Européen sur la Nature) est la référence européenne pour la classification des habitats. Voir https://inpn.mnhn.fr/habitat/cd_typo/7

Tab. 1 Définitions des différents types de relations entre les espèces et les habitats qu'elles occupent, et règles de rattachement

Type de relation	Définition	Règle de rattachement
Présente	Espèce dont la présence est documentée dans l'habitat sans information sur la force de la relation.	Ce statut est défini par défaut en l'absence d'information.
Euryèce	Espèce présente dans de nombreux types d'habitats sans préférence marquée. L'espèce est relativement indifférente vis-à-vis du type d'habitat (ubiquiste).	Espèce qui tolère d'importantes variations de facteurs écologiques. Elle est présente dans plusieurs grands types d'habitats (niveau 1 de la typologie EUNIS) sans préférence marquée pour un de ces niveaux 1, ni pour les niveaux inférieurs.
Oligoèce non-sélective	Espèce dont la présence est documentée dans l'habitat sans information sur la force de la relation ; Espèce présente dans un nombre restreint de types d'habitats sans préférence marquée pour l'un d'entre eux.	Espèce présente dans un seul grand type d'habitat (niveau 1 de la typologie EUNIS) sans préférence marquée aux niveaux inférieurs ; Espèce présente dans plusieurs types d'habitats de niveau inférieurs (niveaux 2 ou 3) appartenant à des niveaux 1 différents, sans préférence marquée pour aucun d'entre eux.
Sélective	Espèce oligoèce démontrant une préférence marquée pour un type d'habitat particulier	Habitat optimal de l'espèce, c'est-à-dire dont la part des effectifs ou le nombre de stations (la densité) est significativement plus important que dans les autres habitats. L'espèce sera notée présente pour les autres types d'habitats non optimaux.
Spécialiste	Espèce sténoèce à niche écologique étroite et présente dans un nombre très restreint de types d'habitats.	Espèce présente dans un seul type d'habitat de niveau 2 EUNIS (par exemple une espèce dont 90 % des effectifs ou des stations est dans cet habitat) ; Espèce présente dans un nombre très restreint de types d'habitats de niveau 3.
Spécialiste strict	Espèce sténoèce inféodée à un seul type d'habitat précis et lié à des conditions particulières.	Espèce associée à des niveaux fins de la typologie d'habitat (4 EUNIS et plus précis) et liée à des conditions particulières.

Différentes méthodes ont été utilisées pour caractériser les relations espèces – habitats, incluant la recherche bibliographique et l'expertise par des spécialistes. Outre le type de relation entre l'espèce et l'habitat, sont également renseignées, si possible, les conditions de validité de la relation selon la biogéographie, la structure de l'habitat, les stades de développement ou l'activité de l'espèce dans cet habitat.

1.2 Objectifs et démarche de l'étude

L'objectif de l'analyse est de construire un modèle qui permette de prédire la potentialité de présence d'espèces, en utilisant la base de connaissance des relations espèces – habitats d'une part, et des données d'observations d'espèces sur le terrain d'autre part. Ce modèle doit également permettre de déterminer des catégories de potentialité de présence des espèces, et donc d'attribuer un degré de confiance à la prédiction.

La construction du modèle doit répondre à la question suivante : quels paramètres généraux permettent de prédire la présence et l'absence des espèces dans des habitats où elles sont susceptibles d'être trouvées ?

Une des principales caractéristiques qui permette d'expliquer la répartition d'une espèce à l'échelle du paysage est, outre son histoire évolutive et l'historique du paysage en question, la présence d'un ou plusieurs habitats qu'elle

occupe pour accomplir son cycle de vie (notion de niche écologique). À une échelle biogéographique, les contraintes climatiques, la latitude, la longitude et l'altitude influencent les aires de répartitions de la plupart des espèces (Hortal et al., 2012). La cartographie des habitats de France, qui aurait pu permettre de mener à bien cette étude, est un projet en cours (projet CarHab), et devrait être accessible dans les années à venir. D'autres moyens sont donc employés pour réaliser l'analyse de prédiction de la potentialité de présence. Les données d'observations naturalistes de l'Inventaire national du patrimoine naturel (INPN), la base de données nationale du Système d'information de l'inventaire du patrimoine naturel (SINP), sont une source particulièrement intéressante pour étudier la distribution des espèces en France. Ainsi, couplées aux données géographiques et climatiques, les nombres d'observations des espèces cibles sur le territoire permettent d'affiner la prédiction. La comparaison de modèles permet de choisir les échelles temporelle et spatiale les plus pertinentes pour prédire le mieux possible la potentialité de présence d'un maximum d'espèces ciblées par l'analyse. Attention, cette étude relève de l'écologie, et les données utilisées ne sont en aucun cas exhaustives, qu'il s'agisse des données d'observations de l'INPN, de la base de connaissance des relations espèces – habitats ou des inventaires de terrain. Les prédictions à partir de ces bases de données, alimentées en permanence, seront donc améliorable au cours du temps.

La construction du modèle doit par ailleurs répondre à certaines contraintes, en particulier celle d'être le plus simple possible. En effet, l'utilisation de ce modèle pour l'application en ligne implique de prendre le moins de place possible sur les serveurs, et, afin que les demandes soient traitées rapidement, que les calculs ne demandent pas trop de données à requêter dans les bases de l'INPN.

Il est également nécessaire de rappeler que le modèle présenté ici n'a pas pour but de remplacer les inventaires sur le terrain, mais qu'il vise notamment à être un outil d'aide pour cibler certaines espèces lors de ces inventaires. Cet outil n'a pas non plus vocation à fournir la liste exhaustive des espèces présentes sur un site donné.

1.3 Organisation du rapport

Ce rapport est organisé à la manière d'un article scientifique. À la suite de cette introduction, une partie « Méthodes » s'attache à décrire les données utilisées, la procédure mise en œuvre pour la construction du modèle et la validation des données, puis les analyses statistiques réalisées. Une troisième partie compile les résultats pour chaque étape de la démarche d'analyse, ainsi que le test du modèle sélectionné sur d'autres données d'inventaire. La quatrième et dernière partie « Perspectives » permet de conclure sur les résultats de l'analyse et de présenter les limites et les pistes d'amélioration de l'outil de prédiction de présence des espèces.

2 Méthodes

2.1 Démarche d'analyse

L'enjeu de cette étude est de réussir à construire un modèle prédictif de qualité, pour de nombreuses espèces, avec des paramètres généraux. Les différentes étapes de la démarche d'analyse conduite pour cette étude sont résumées ci-après et détaillées dans les chapitres suivants.

1- La question initiale

La question primordiale qui guide la démarche d'analyse est la suivante : quels paramètres peuvent permettre de prédire la présence ou l'absence d'espèces sur un territoire donné ? Pour répondre à cette question, l'analyse teste, à l'aide d'un jeu de données d'observations « réelles » sur le terrain, quels critères prédisent le mieux la présence et l'absence des espèces.

Des critères très généraux ont été choisis pour cette étude, en raison de l'échelle importante à laquelle elle se situe (la France métropolitaine), et du besoin de simplicité pour la modélisation. Les critères choisis sont d'une part le climat, la latitude et la longitude du terrain inventorié, et d'autre part les données préexistantes sur l'observation des espèces potentiellement présentes. En effet, il est probable qu'une espèce observée fréquemment sur un territoire soit également présente à proximité. Ces données sont bancarisées dans l'INPN.

2- Les données de base

Les jeux de données qui permettent de tester le potentiel prédictif des critères choisis sont la base de connaissance des relations espèces – habitats et les données d'inventaires IQE/IPE qui compilent des inventaires d'espèces et d'habitats sur les mêmes sites. Ainsi, à partir de la liste des habitats inventoriés sur chaque site il est possible de connaître les espèces potentiellement présentes de par leur lien à ces habitats renseignés dans la base de connaissance.

D'autres données sont utilisées afin d'affiner la liste d'espèces potentielles, de renseigner les informations pour les différents critères explicatifs choisis, mais également sur la patrimonialité des espèces et leur groupe taxonomique. Les données spatiales sont renseignées à l'échelle du maillage 10x10 km de la France métropolitaine. C'est une échelle à laquelle sont renseignées plusieurs informations nécessaires à la construction du jeu de données telles que les données d'occurrences d'espèces dans l'INPN ou encore le taux de méconnaissance locale des groupes taxonomiques.

3- Le modèle statistique

La variable à expliquer est l'observation ou non dans les inventaires des espèces potentiellement présentes. Cette variable peut prendre deux valeurs : 1 ou 0, selon si l'espèce potentiellement présente a été observée sur le terrain ou non. Le modèle statistique privilégié dans ce contexte est une régression logistique. Chaque mention d'une espèce dans une maille (liée aux habitats qui s'y trouvent) est appelée dans ce rapport « **occurrence potentielle** ». Chaque espèce peut être potentiellement présente dans plusieurs mailles. La modélisation permet d'attribuer à chaque occurrence potentielle une probabilité de présence en fonction des variables explicatives (zone climatique, latitude et longitude du site d'inventaire, observations préexistantes). La spécificité des espèces est prise en compte dans le modèle sous la forme d'un facteur aléatoire. En effet, les multiples observations ne sont pas indépendantes, dans la mesure où la distribution de chaque espèce dépend de ses préférences écologiques.

4- La comparaison de modèles

Pour la variable explicative « données préexistantes », qui correspond aux données d'observations de l'INPN pour les espèces potentielles, la question se pose des échelles spatiales et temporelles les plus pertinentes. Différents modèles sont donc comparés selon la qualité de leurs prédictions, en modifiant d'abord l'échelle spatiale de la variable (données sur la maille, données locales et régionales) et leur résolution (nombre total d'observations ou proportions de mailles avec au moins une observation), puis l'échelle temporelle (données datant de moins de 10 ans, entre 10 et 20 ans ou plus de 20 ans). Les modèles sont calibrés sur 80 % du jeu de données et les prédictions sont calculées sur les 20 % restant, et cette procédure est répétée sur dix tirages aléatoires des données. Cette procédure permet de déterminer de manière robuste les variables explicatives à intégrer pour choisir le modèle le plus efficace pour discriminer les espèces potentielles selon leur observation ou non sur le terrain.

5- Diagnostic

Une fois les variables explicatives choisies, les prédictions du modèle le plus pertinent sont analysées selon différents facteurs non inclus dans les variables explicatives. Cette étape permet de tester si les prédictions sont similaires en fonction du groupe taxonomique, selon la patrimonialité de l'espèce, ou encore selon le critère de méconnaissance du groupe taxonomique à l'échelle locale.

6- Cas d'étude

Le modèle prédictif est ensuite appliqué sur d'autres données de terrain, issues de l'inventaire réalisé dans le cadre d'une étude d'impact. Cette étape est une étude de cas, simulant les conditions d'utilisation du modèle dans l'application, avec l'avantage de pouvoir vérifier la qualité de la prédiction pour ce cas précis, grâce à la liste des espèces inventoriées sur le site d'étude.

2.2 Référentiels et données utilisées

2.2.1 Référentiels nationaux : TAXREF et HABREF

La présente analyse s'appuie sur TAXREF v13 (Gargominy et al., 2019), référentiel national sur la faune, la flore et la fonge de France métropolitaine et ultramarine.
Pour les habitats, ce rapport se base sur HABREF v4.0 (Clair et al., 2017), référentiel national réunissant les versions officielles de référence des typologies d'habitats ou de végétation couvrant les milieux marins et/ou continentaux des territoires français de métropole et d'outre-mer.

Le référentiel HABREF inclut une table (*HABREF_CORRESP_TAXON*) qui compile les relations expertisées. La base de connaissance sur les relations espèces – habitats associe les espèces de France métropolitaine aux grands types de milieux qu'elles occupent en faisant le lien entre deux référentiels nationaux : le référentiel taxonomique (TAXREF) et le référentiel des habitats et des végétations (HABREF), en utilisant la typologie EUNIS (voir Louvel et al., 2013). L'objectif est de définir le type de relation existant entre un habitat et une espèce, mais également les conditions de validité de cette relation (conditions biogéographiques, stade de développement, structure de l'habitat, type d'activité de l'espèce dans l'habitat). L'UMS PatriNat gère la base de connaissance des relations espèces – habitats, régulièrement alimentée par de nouvelles relations caractérisées.

2.2.2 Base de connaissance des relations espèces – habitats

La base de connaissance des relations espèces – habitats étant en cours de construction (*cf.* tableau 2), les groupes taxonomiques de France métropolitaine qui ont été renseignés en priorité (les mieux connus) ont été sélectionnés pour la présente étude : amphibiens, lépidoptères, odonates, oiseaux nicheurs, orthoptères et reptiles. Pour les angiospermes, l'étude bénéficie des retours d'expertise des Conservatoires Botaniques Nationaux (CBN) de métropole sur les relations flore – habitats pour plus de 1 500 espèces. Ces données sont en cours de bancarisation dans la base de connaissance des relations espèces – habitats.

La base de connaissance des relations espèces – habitats est scindée en deux sous-parties. L'une d'elles est accessible au téléchargement au sein d'une table (*HABREF_CORRESP_TAXON_40* qui combine différentes typologies de description des habitats (CORINE biotopes, Prodrome des végétations de France, EUNIS, etc.). Elle combine également différents types de relations espèces – habitats (comme les espèces caractéristiques ou dominantes de groupements végétaux en phytosociologie par exemple). Une sélection a donc été réalisée pour ne conserver que les relations espèces – habitats expertisées dans le cadre du projet, présentées dans la figure 1 et pour lesquelles les habitats sont codés selon la typologie EUNIS. Cette sous-partie comprend uniquement des taxa animaux. L'autre sous-partie est issue de l'expertise des CBN, qui a été compilée fin 2019 mais n'est pas encore bancarisée dans le référentiel HABREF. Cette sous-partie comprend uniquement des espèces végétales.

Tab. 2 Degré d'exhaustivité, par groupe taxonomique, de la base de connaissance des relations espèces – habitats par rapport au référentiel TAXREF (v13) pour les espèces métropolitaines.

Groupe taxonomique	Nombre d'espèces		Taux d'exhaustivité
	TAXREF[2]	BdC espèces – habitats	
Amphibiens	43	33	77 %
Mammifères	129	57	44 %
Oiseaux nicheurs	307	278	91 %
Poissons	111	22	20 %
Reptiles	46	34	74 %
Insectes	40 366	2 010	5 %
Coléoptères	10 890	6	< 1 %
Hétérocères	5 247	1 455	28 %
Odonates	97	87	90 %
Orthoptères	237	203	86 %
Rhopalocères	303	248	82 %
Bivalves	42	3	7 %
Crustacés	946	2	< 1 %
Gastéropodes	676	4	< 1 %
Angiospermes	7 601	1 532	20 %

La partie « Faune » comprend 2 528 espèces et 354 habitats, avec en moyenne 3,6 ± 2,7 habitats par espèce (médiane = 3, maximum = 22, minimum = 1), et en moyenne 26 ± 48 espèces par habitat (médiane = 6, maximum = 363, minimum = 1). Le tableau ci-contre décline le nombre d'habitats dans chaque niveau de la typologie EUNIS.

Niveaux dans la typologie EUNIS des habitats de la partie Faune

Niveau EUNIS	Nombre d'habitats
N1	10
N2	66
N3	179
N4	95
N5	4

La partie « Flore » comptabilise 1 532 espèces et 209 habitats. Les ordres les plus représentés (en nombre d'espèces) sont les Asterales (229), les Asparagales (168), les Poales (167), les Lamiales (151), les Caryophyllales (124) et les Fabales (119). Cette partie de la base de connaissance compte en moyenne 5,7 habitats par espèce ± 3,3 (médiane = 5, maximum = 36, minimum = 1), et en moyenne 42 espèces par habitat ± 98 (médiane = 10, maximum = 1 039, minimum = 1). Le tableau ci-contre décline le nombre d'habitat dans chaque niveau de la typologie EUNIS.

Niveaux dans la typologie EUNIS des habitats de la partie Flore

Niveau EUNIS	Nombre d'habitats
N1	11
N2	45
N3	152
N4	1

2.2.3 Autres jeux de données de référence sur la connaissance des espèces

Les données de l'Atlas de Biodiversité Départementale et des Secteurs Marins (ABDSM) constituent un catalogue non exhaustif mais expertisé de la présence avérée des espèces par département. Ces données permettent de filtrer la liste d'espèces potentiellement présentes sur un territoire. Prenons un exemple fictif : si certains habitats

[2] Sont comptabilisées les espèces respectant à la fois les conditions d'habitats et de statut de présence suivantes : espèces vivant en habitat d'eau douce (2), terrestre (3), marin et d'eau douce (4), marin et terrestre (5), en eau saumâtre (6), terrestre et/ou eau douce (7), terrestre et eau douce (8) ; statut de présence des espèces en France métropolitaine : espèce présente (P), occasionnelle (B), endémique (E), subendémique (S), cryptogène (C), introduite (I), introduite envahissante (J).

pouvant être occupés par des chamois existaient dans le Finistère, cette espèce pourrait être considérée comme potentiellement présente. Or son statut dans l'ADBSM précisant qu'elle est absente de ce département permet de l'exclure automatiquement de la liste des espèces potentiellement présentes.

La Base de connaissance « Statuts » (V13, version au 05/02/2020, Régnier & Gargominy, 2018) récapitule les statuts de protection, de réglementation ou encore de conservation des espèces françaises.
Pour la présente analyse, sont considérées patrimoniales les espèces répondant à au moins un des critères suivants :

- statut Liste rouge nationale en catégories « VU » (Vulnérable), « EN » (En danger), « CR » (En danger critique) ou « CR* » (l'espèce est peut-être éteinte ou disparue) ;
- statut de protection nationale ;
- inscrites en annexe 2 ou 4 de la Directive 92/43/CEE dite « Habitats-Faune-Flore » ;
- inscrites en annexe 1 de la Directive 79/409/CEE dite directive « Oiseaux » ;
- définies comme une espèce déterminante ZNIEFF dans la région concernée.

2.2.4 Données d'inventaires pour la calibration du modèle

Des données « réelles » sont nécessaires à la calibration du modèle, afin de confronter les prédictions à des inventaires naturalistes réalisés sur le terrain et ainsi évaluer la qualité de prédiction du modèle.
L'Indice de Qualité Ecologique (IQE) est un protocole semi-standardisé élaboré par l'UMS PatriNat (Delzons et al., 2020). Il permet d'évaluer la biodiversité et la fonctionnalité écologique des sites dans une optique de mise en œuvre de mesures d'aménagements et de gestion écologiques, puis de leur suivi. Les inventaires de terrain s'attachent à décrire la patrimonialité des habitats naturels et des espèces végétales et animales présentes (en particulier des oiseaux, reptiles, amphibiens, papillons et libellules), la fonctionnalité des habitats naturels, et le degré de connectivité avec les réseaux écologiques. Les inventaires se déroulent sur 6 journées, incluant un passage crépusculaire et nocturne, échelonnées du début du printemps à la fin de l'été. Une variante de cet indice a également été élaborée sur la base d'un temps d'inventaire réduit à une journée (IPE : Indice de Potentialité Ecologique) avec pour objectif l'établissement d'un pré-diagnostic des enjeux de biodiversité du site (voir http://iqe-spn.mnhn.fr/accueil/liqe/).

L'analyse repose ainsi sur les données issues d'inventaires IQE et IPE bancarisées par l'UMS Patrinat et réalisés en interne par les collègues naturalistes de l'équipe Espaces et Partenariats de l'UMS PatriNat et/ou commandités par des entreprises partenaires (EDF, SUEZ, Fédération française de golf).
Pour chaque inventaire de terrain, les données suivantes ont été utilisées : liste des habitats (codés selon la typologie EUNIS), liste des espèces, date de l'inventaire, et le code de la maille 10x10 km où se situe chaque site inventorié.

2.2.5 Bases de données de l'INPN et carroyage

Pour toutes les espèces qui sont à la fois (i) potentiellement présentes d'après la liste des habitats et (ii) dans la liste des espèces observées dans les relevés IQE/IPE, le nombre d'observation dans chaque maille métropolitaine est compilé sur trois pas de temps : depuis 10 ans (sur la période 2009-2019), entre 10 et 20 ans (entre 1999 et 2009) et plus vieilles que 20 ans (antérieures à 1999).
Les données d'observation ont également été compilées sur trois échelles géographiques imbriquées : à l'échelle de la maille de 10x10 km où se trouve le site inventorié (maille focale), à l'échelle des mailles directement adjacentes à la maille focale et à l'échelle « régionale » (sans rapport avec les limites administratives), c'est-à-dire dans un rayon de 50 km autour du centroïde de la maille focale (*cf.* figure 2).

Ainsi, la construction du jeu de données a nécessité de compiler la liste des mailles de 10x10 km adjacentes à chaque maille métropolitaine, ainsi que la liste des mailles dont le centroïde est à une distance inférieure ou égale à 50 km du centroïde de chaque maille métropolitaine.

La zone climatique de chaque maille est connue grâce au croisement réalisé entre la grille des mailles métropolitaines de 10x10 km et les délimitations des enveloppes climatiques européennes (The Environmental Stratification of Europe, version 8, Metzger 2018, https://doi.org/10.7488/ds/2356, téléchargé depuis https://datashare.is.ed.ac.uk/handle/10283/3091). Dans les cas où une maille chevauche une limite entre deux zones climatiques, la zone qui occupe la plus grande partie de la maille a été attribuée. Ce croisement a été réalisée à l'aide du logiciel Qgis3.12 (QGIS, 2020).

D'autres informations sont prises en compte telles que le groupe taxonomique (information disponible dans TAXREF) et le critère de méconnaissance locale du groupe taxonomique. Ce critère provient de la comparaison entre le nombre d'espèces par maille de 10x10 km ayant une observation dans l'INPN, et l'expertise, par groupe taxonomique, du nombre de taxons minimum qui devraient être trouvés dans la maille de métropole ayant la plus faible diversité spécifique pour ce groupe. Un groupe taxonomique est considéré comme méconnu à l'échelle de la maille si le nombre de taxa renseignés dans les bases de données est inférieur au seuil expertisé (Witté & Touroult, 2017).

2.3 Mise en forme du jeu de données utilisé pour l'analyse statistique

Le processus de construction du jeu de données est présenté de façon schématique par la figure 4.

1- Compilation de la liste d'espèces potentielles

La première étape consiste à compiler la liste des espèces potentiellement présentes sur chaque site d'inventaire IQE/IPE. Cette liste est constituée par toutes les espèces liées, dans la base de connaissance des relations espèces – habitats, aux habitats inventoriés sur chaque site. Toutefois, les habitats décrits sur les sites et ceux utilisés pour décrire les relations avec les espèces ne sont pas nécessairement du même niveau dans la classification EUNIS. Par exemple, si l'habitat G4 a été mentionné dans l'inventaire de terrain, les espèces en relation avec des habitats « inférieurs » dans la hiérarchie, telles que *Parnassius mnemosyne* avec l'habitat G4.6, pourraient ne pas être pas incluses dans la liste des espèces potentiellement présentes. Il en va de même avec les espèces en relation avec des habitats « supérieurs » dans la hiérarchie. Il a donc été choisi de compiler toutes les espèces liées dans la base de connaissance avec tous les habitats de niveaux inférieurs et supérieurs dans la classification EUNIS par rapport à l'habitat mentionné dans l'inventaire de terrain.

2- Filtrage de la liste et ajout de paramètres

Cette liste est ensuite filtrée selon deux méthodes : premièrement en supprimant, pour les espèces dont la carte ABDSM est validée, les espèces absentes du département. Deuxièmement, seules sont conservées les espèces appartenant aux groupes taxonomiques expertisés avec un bon taux d'exhaustivité dans la base de connaissance des relations espèces – habitats (*cf.* tableau 4), autrement dit les oiseaux nicheurs, les reptiles, les amphibiens, les odonates, les orthoptères, les rhopalocères et les hétérocères. Les angiospermes ont également été prises en compte, parce que les espèces de flore patrimoniales ont été prioritaires pour l'expertise des relations espèces – habitats par les CBNs, et que ces espèces représentent une cible particulière pour l'outil de prédiction.

Les statuts de protection, de réglementation ou de conservation sont ajoutés pour chaque espèce, selon les critères décrits au paragraphe 2.1.2.
Lorsque sur une même maille sont renseignés plusieurs sites et plusieurs habitats, une seule mention de chaque espèce potentielle est conservée. Cette mesure permet d'éviter la pseudo-réplication, qui accorderait dans l'analyse un poids plus grand aux espèces souvent présentes par rapport aux espèces rares.

3- Ajout des données de l'INPN et zonage climatique

Ensuite sont compilés pour chaque espèce et pour chacun des trois pas de temps : le nombre d'observations bancarisées dans l'INPN dans chaque maille focale, dans les mailles adjacentes, dans les mailles dans un rayon de 50 km, ainsi que le nombre et la proportion de mailles dans un rayon de 50 km avec au moins une donnée d'observation pour l'espèce (*cf.* figure 2).

Le choix de la proportion de mailles avec au moins une donnée d'observation est issu d'une volonté d'homogénéisation inter-mailles, pour ne pas donner moins de poids dans l'analyse à des mailles frontalières ou littorales (qui ont moins de mailles adjacentes ou régionales par rapport aux mailles centrales du territoire métropolitain).

Fig. 2 Schématisation des 3 échelles géographiques prises en compte à partir du maillage 10x10 km de la France métropolitaine. En zone littorale, seules les mailles terrestres ont été prises en compte. Pour chaque échelle et pour chaque espèce du jeu de données, les observations répertoriées dans l'INPN ont été dénombrées. De plus, pour chaque espèce potentiellement présente, le nombre et la proportion de mailles adjacentes et de mailles régionales avec au moins une donnée ont été calculés.

La zone climatique de chaque maille focale est ajoutée dans le jeu de données (*cf.* tableau 3 et figure 3).

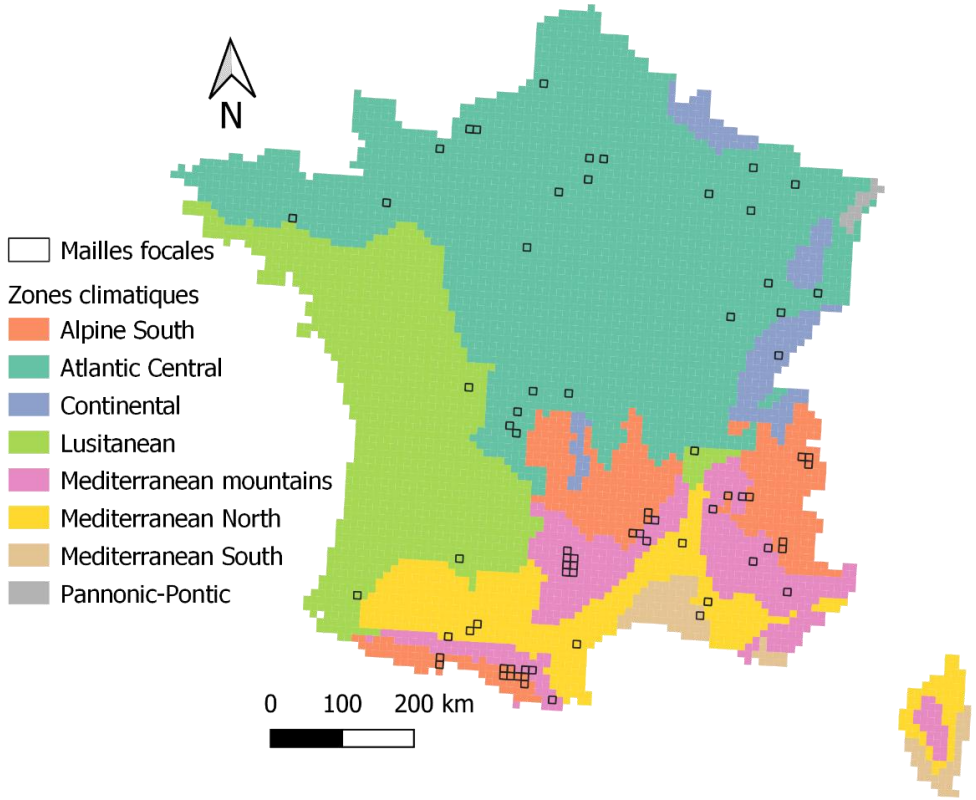


Fig. 3 Répartition des zones climatiques sur le maillage 10x10 km de France métropolitaine (d'après The Environmental Stratification of Europe, version 8, Metzger 2018). Les mailles focales de l'analyse sont mises en évidence par une bordure noire.

Tab. 3 Nombre de mailles focales par zone climatique

Zone climatique	Nb. mailles
Alpine South	19
Atlantic Central	24
Continental	1
Lusitanean	4
Mediterranean Mountains	19
Mediterranean North	6
Mediterranean South	1
Pannonic-Pontic	0

4- Ajout du paramètre d'observation ou non des espèces potentielles

La dernière étape de la construction du jeu de données est l'attribution du paramètre qui constituera la variable à expliquer du modèle statistique : les espèces potentielles effectivement observées dans la maille pendant les IQE/IPE se voient attribuer un 1, les autres un 0.

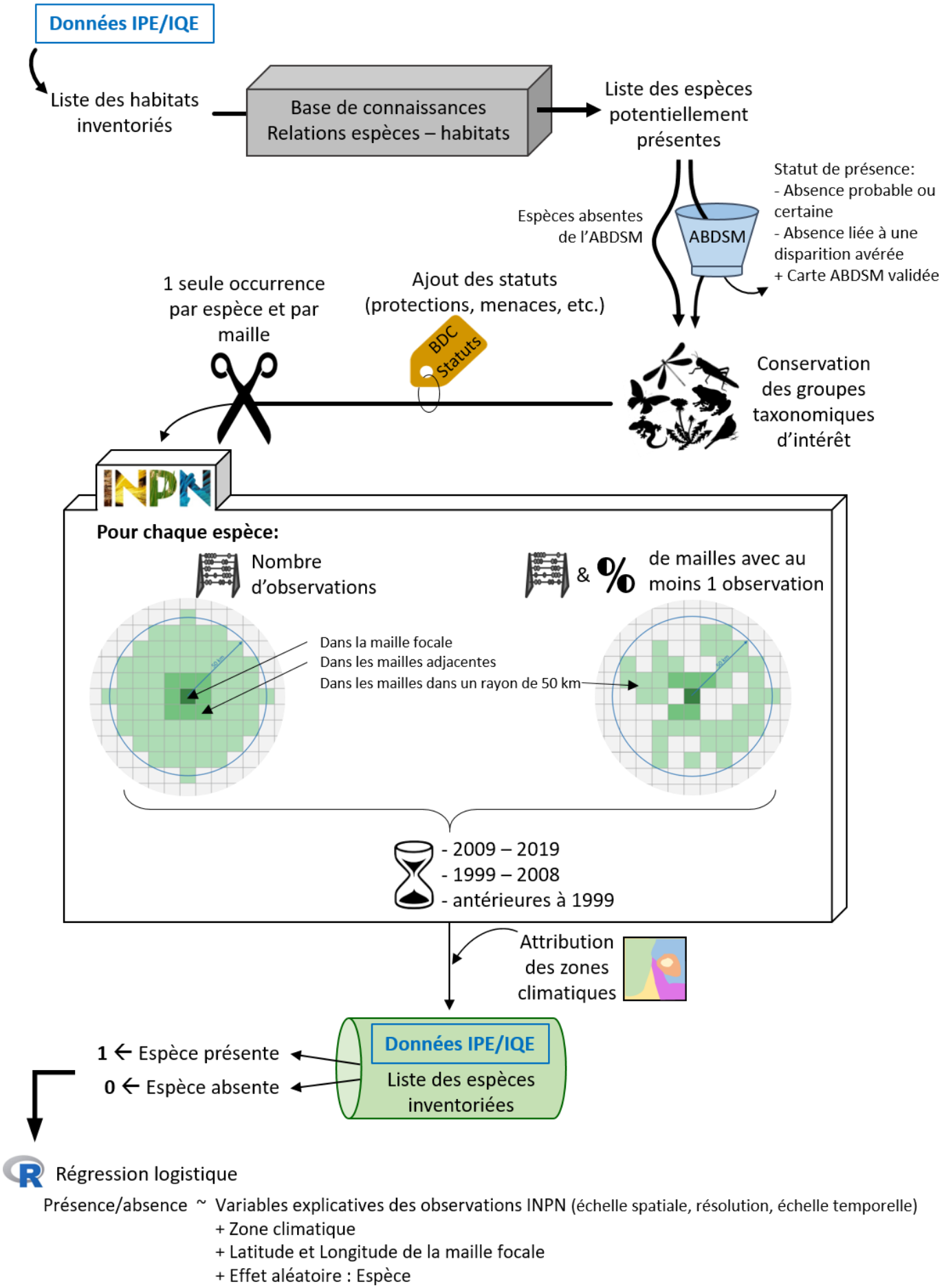


Fig. 4 Schéma de la production du jeu de données servant à la construction du modèle prédictif.

5- Caractéristiques du jeu de données

Le jeu de données comprend 187 498 occurrences potentielles (1 occurrence potentielle correspond à 1 espèce potentielle dans une maille). Les données sont réparties dans 73 mailles et sur 36 départements. Au total, 327 habitats sont représentés et 3 697 espèces, parmi lesquelles 2 512 sont patrimoniales (68 %). Un total de 34 151 occurrences potentielles (18 %) sont des espèces potentielles effectivement observées sur le terrain.

Les hétérocères sont finalement exclus de l'analyse parce que ce groupe n'est pas systématiquement échantillonné pendant les inventaires IQE/IPE, ce qui se traduit par une très faible proportion de données de présence par rapport aux autres groupes taxonomiques conservés dans l'analyse (*cf.* tab 4).

Le jeu de données final utilisé pour modéliser la potentialité de présence comprend donc un total de 2 287 espèces, dont 1 919 patrimoniales (84 %). Un quart des occurrences potentielles correspondent à des espèces potentielles effectivement observées sur le terrain (32 437 occurrences potentielles). Il y a en moyenne 55,7 ± 23 occurrences potentielles par espèce (médiane = 70, minimum = 1, maximum = 73).

Tab. 4 Caractéristiques du jeu de données par groupes taxonomiques

Groupe taxonomique	Nombre d'occurrences potentielles	Nombre d'espèces	Espèces patrimoniales		Pourcentage d'occurrences potentielles où l'espèce a été observée
			Nombre	Pourcentage	
Amphibiens	965	28	27	96 %	67 %
Angiospermes	965 250	1 478	1 246	84 %	18 %
Hétérocères	60 224	1 410	593	42 %	2,8 %
Odonates	2 686	83	70	84 %	65 %
Oiseaux	12 523	259	252	97 %	62 %
Orthoptères	9 187	191	127	66 %	18 %
Reptiles	1 032	30	29	97 %	61 %
Rhopalocères	4 361	218	168	77 %	68 %

Le nombre total d'espèces potentielles dans le jeu de données analysé représente presque la totalité des espèces présentes dans la base de connaissance des relations espèces – habitats (tableau 5). Les espèces ne faisant pas partie du jeu de données sont des espèces dont les habitats renseignés dans la base de connaissance n'ont pas été observés sur le terrain.

Tab. 5 Degré d'exhaustivité du jeu de données par rapport à la base de connaissance des relations espèces – habitats, selon les groupes taxonomiques d'intérêt.

Groupe taxonomique	Nombre d'espèces		Pourcentage
	BdC relations espèces – habitats	Jeu de données	
Amphibiens	33	28	85 %
Angiospermes	1 528	1 478	97 %
Hétérocères	1 455	1 410	97 %
Odonates	87	83	95 %
Oiseaux	278	259	93 %
Orthoptères	203	191	94 %
Reptiles	34	30	88 %
Rhopalocères	248	218	88 %

Les mailles focales sont réparties sur l'ensemble du territoire métropolitain continental (*cf.* figure 5). Le nombre d'observations INPN par maille (dans un rayon de 50 km autour du centroïde et toutes période confondues) est décuplé entre la maille ayant le moins d'observations et la maille en ayant le plus. Les mailles focales caractérisées par un grand nombre de données sont majoritairement dans la moitié sud de la métropole.

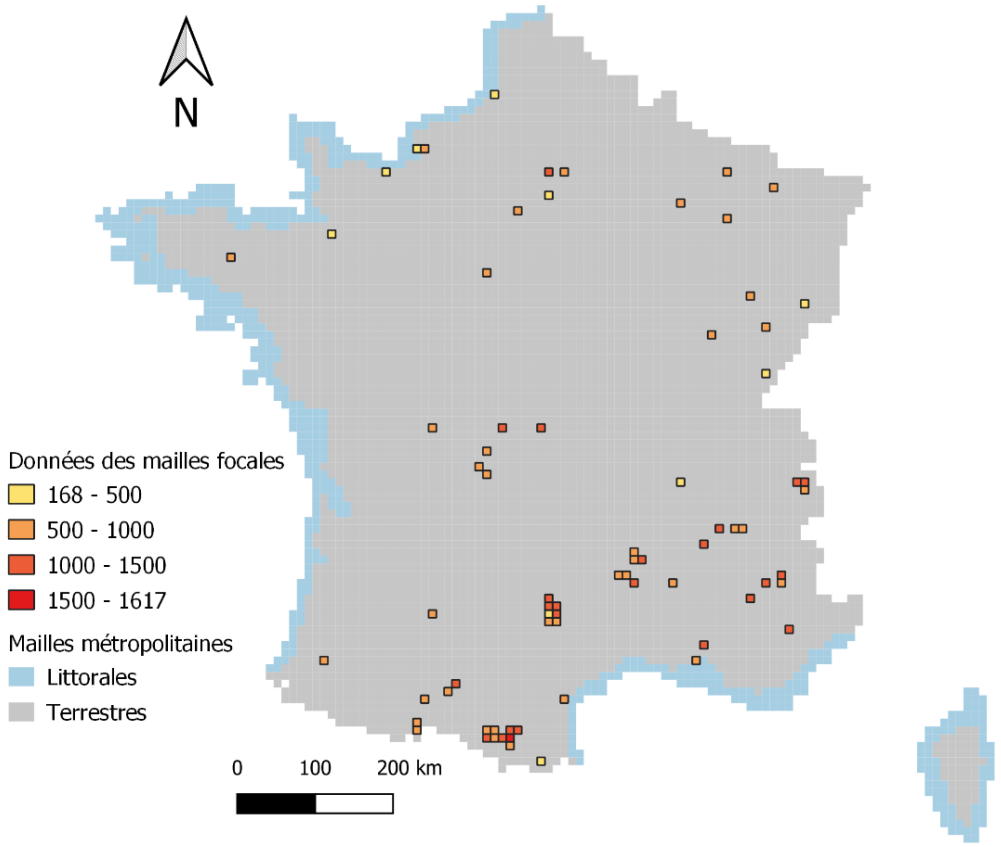


Fig. 5 Répartition et quantité de données par maille, en incluant les données d'observations INPN (toutes périodes de temps confondues) dans un rayon de 50 km autour du centroïde de chacune des mailles focales.

2.4 Analyse statistique

Il s'agit de tester les paramètres qui permettent de prédire de façon suffisamment fiable la présence ou l'absence des espèces potentiellement présentes d'après les données des inventaires IQE/IPE. Le modèle statistique utilisé est une régression logistique mixte dont la variable à expliquer suit une loi binomiale.
Le modèle final doit répondre aux contraintes suivantes :

- être le plus simple possible,
- rester très général, puisque l'analyse est réalisée à l'échelle de la France métropolitaine,
- permettre de prédire correctement les espèces observées ou non dans les données d'IQE/IPE.

Le choix du meilleur modèle se justifie plus par la justesse de la prédiction que par la significativité des variables. Le but est d'avoir un modèle qui discrimine effectivement les espèces observées des espèces non-observées, avec le moins possible de faux-positifs et surtout de faux-négatifs (figure 6). En effet, le modèle doit idéalement permettre de prédire comme étant présentes toutes les espèces observées sur le terrain.

		Modèle prédictif	
		Probablement présente	Probablement absente
IQE/IPE	Observée (1)	✓	Faux-négatif
	Non observée (0)	Faux-positif	✓

Fig. 6 Schéma d'une matrice de confusion, qui permet d'évaluer la qualité de discrimination du modèle.

Le modèle statistique a pour variable à expliquer l'observation (1) ou non (0) de chaque espèce potentielle sur les mailles focales pendant les inventaires de terrain.
Le modèle intègre les variables explicatives suivantes (voir également tableau 6) :

- le nombre brut d'observations (centré-réduit) ou la proportion de mailles avec au moins une observation dans l'INPN, à une certaine échelle spatiale et un certain pas de temps (à comparer) ;
- la zone climatique des mailles focales ;
- les coordonnées spatiales des centroïdes des mailles focales : latitude et longitude exprimées en degrés décimaux divisé par 10^9 ;
- l'espèce, en effet aléatoire.

Le modèle général est donc de la forme suivante :

$$f(observation) = \alpha \times données\ INPN + \beta + \gamma \times latitude + \delta \times longitude + \varepsilon + C$$

Avec :

- f la fonction *logit* ($\log \frac{x}{1-x}$)
- α le coefficient du paramètre données INPN,
- β le coefficient de la zone climatique de la maille focale,
- γ le coefficient du paramètre latitude,
- δ le coefficient du paramètre longitude,
- ε le coefficient aléatoire de l'espèce cible,
- C l'intercept du modèle.

Afin de comparer l'effet spatial, l'effet de la résolution (nombre brut ou proportion) et l'effet temporel des données d'observations INPN, différents modèles sont construits, en changeant à chaque fois la variable explicative associée aux observations INPN. Le premier critère de comparaison est double, en incluant l'échelle spatiale et la résolution des observations INPN. Pour s'affranchir de l'effet temporel lors de cette première étape de choix, les données d'observation sur les trois pas de temps ont été combinées.

Une fois que la variable spatiale permettant la meilleure prédiction est sélectionnée, la dimension temporelle des données d'observation INPN est choisie. Pour chaque pas de temps un modèle est créé et les prédictions sont comparées afin de choisir la période la plus pertinente.

Tab. 6 Description des différentes variables utilisées pour l'analyse statistique et le diagnostic permettant d'évaluer la qualité de discrimination du modèle.

Variable	Usage
Espèce observée dans la maille focale pendant les inventaires IPE/IQE (1= OUI, 0= NON)	Modèle (variable à expliquer)
Nombre d'observations dans l'INPN de l'espèce dans la maille focale (3 pas de temps : données datant de moins de 10 ans, entre 10 et 20 ans, plus de 20 ans)	Modèle (effet fixe)
Nombre d'observations dans l'INPN de l'espèce dans les mailles adjacentes à la maille focale (3 pas de temps)	Modèle (effet fixe)
Nombre d'observations dans l'INPN de l'espèce dans les mailles dans un rayon de 50 km autour du centroïde de la maille focale (3 pas de temps)	Modèle (effet fixe)
Nombre de mailles dans un rayon de 50 km autour du centroïde de la maille focale avec au moins une observation dans l'INPN de l'espèce (3 pas de temps)	Modèle (effet fixe)
Proportion de mailles dans un rayon de 50 km autour du centroïde de la maille focale avec au moins une observation dans l'INPN de l'espèce (3 pas de temps)	Modèle (effet fixe)
Zone climatique de la maille focale (7 possibilités)	Modèle (effet fixe)
Longitude du centroïde de la maille focale	Modèle (effet fixe)
Latitude du centroïde de la maille focale	Modèle (effet fixe)
Nom de l'espèce	Modèle (effet aléatoire)
Groupe taxonomique (7 possibilités)	Diagnostic
Patrimonialité de l'espèce (OUI ou NON)	Diagnostic
L'espèce fait partie d'un taxon considéré comme connu sur cette maille (1 = oui, 0 = non)	Diagnostic
L'espèce a des données d'observation dans la maille focale (OUI ou NON)	Diagnostic

Chaque modèle est calibré sur 10 tirages aléatoires de 80 % du jeu de données (sous-échantillons d'entrainements comprenant chacun 101 819 occurrences potentielles), et ses prédictions sont testées sur les 20 % de données restantes (sous-échantillons de test, 25 455 occurrences potentielles chacun). Cette procédure permet d'une part d'éviter un biais de sous-estimation du taux d'erreurs, en testant les prédictions sur de nouvelles données. D'autre part la répétition des tests pour chaque modèle permet d'obtenir une qualité de prédiction robuste.
En revanche, une fois le modèle le plus satisfaisant sélectionné, il est recalibré sur le jeu de données complet, afin de récupérer pour toutes les espèces potentielles leur coefficient aléatoire, qui servira à tester sur d'autres données réelles la pertinence du modèle prédictif.

Pour chaque modèle créé, les valeurs de prédiction de présence sont attribuées pour chaque espèce potentielle dans chaque maille focale du sous-échantillon de test.
L'efficacité de discrimination des modèles est ensuite évaluée en comparant la distribution des valeurs prédites des occurrences potentielles selon leur observation ou non sur le terrain. Dans l'idéal, tous les cas d'espèces non-observées sur le terrain sont dans une gamme de valeurs de prédiction de présence différente de ceux pour lesquels les espèces ont été observées pendant les IQE/IPE (*cf.* figure 7).

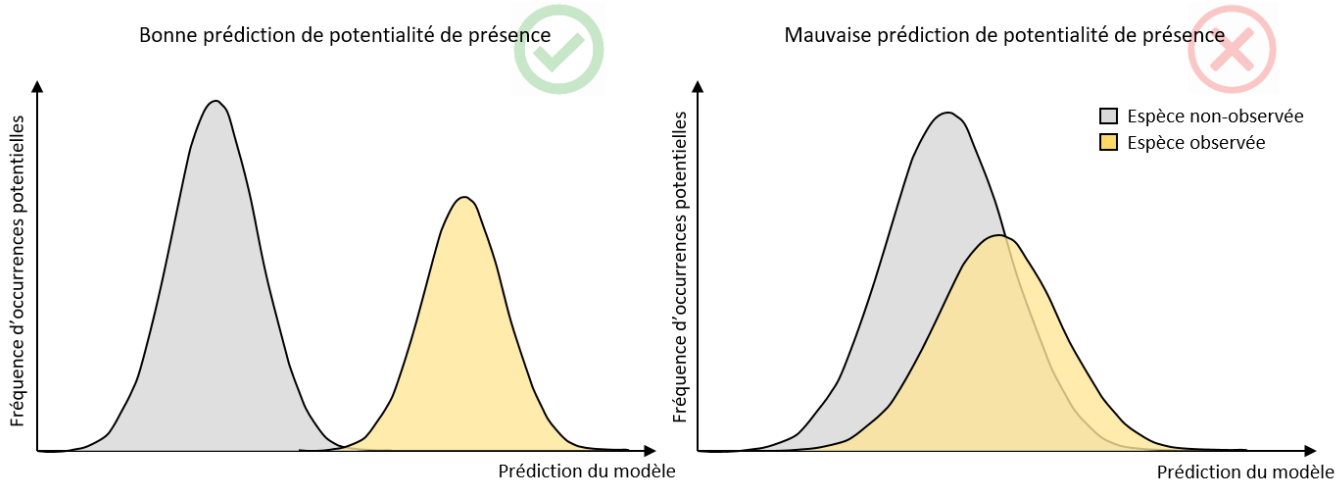


Fig. 7 Schéma illustrant les prédictions de potentialité de présence attendues si le modèle permet une bonne discrimination des catégories d'observations des espèces (à gauche) ou une mauvaise discrimination (à droite). Les courbes représentent la fréquence d'occurrences potentielles en fonction de la prédiction du modèle et selon la catégorie d'observation sur le terrain : en gris les occurrences potentielles correspondant à des espèces potentielles non-observées sur le terrain, en jaune les occurrences potentielles correspondant à des espèces observées sur le terrain.

Le meilleur modèle est choisi selon la proportion moyenne du jeu de données en dehors de la zone de chevauchement des deux catégories de données : plus la zone est réduite plus le modèle a un grand pouvoir de discrimination (*cf.* figure 8). Cette méthode de comparaison des modèles permet de sélectionner les variables explicatives les plus pertinentes pour la prédiction. Une fois les variables explicatives sélectionnées, le modèle est appliqué à l'ensemble du jeu de données et, pour une meilleure perception du résultat prédictif, les valeurs de prédiction du modèle final sont circonscrites entre 0 et 1.

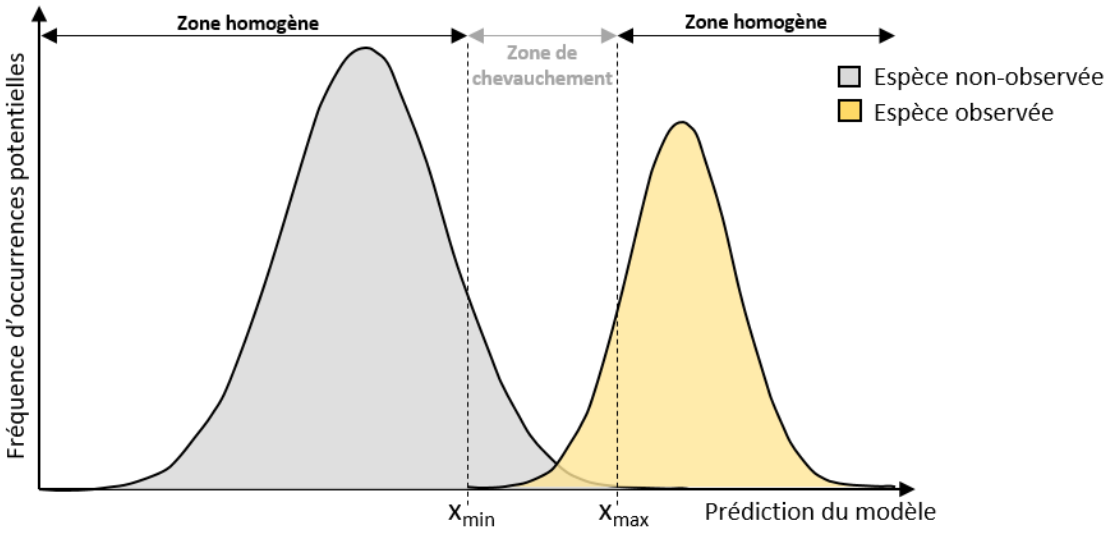


Fig. 8 Schéma illustrant la proportion d'occurrences potentielles prédites dans gammes de valeurs d'attribution homogènes (hors de la zone de chevauchement des distributions), utilisée pour comparer les modèles selon la qualité de discrimination des occurrences potentielles selon leur catégorie d'observation. La proportion du jeu de données en dehors de la gamme de valeurs prédites entre x_{min} (la valeur de prédiction la plus basse de l'aire jaune) et x_{max} (la valeur la plus haute de l'aire grise) est calculée pour chaque modèle testé.

Les analyses statistiques sont réalisés à l'aide du logiciel R version 4.0.3 (R Core Team, 2020), le traitement statistique utilise en particulier la fonction glmer du package *lme4* (Bates et al., 2015)

3 Résultats

3.1 Échelle spatiale d'analyse et résolution

La première étape de l'analyse consiste à choisir l'échelle spatiale des données INPN pour expliquer la présence ou l'absence des espèces potentielles. Les trois échelles testées sont la maille focale, les mailles adjacentes à la maille focale et les mailles régionales (dans un rayon de 50 km). La résolution des données d'observation INPN fait référence au type de données, entre l'utilisation brute du nombre de données ou du nombre de mailles ayant au moins une donnée d'observation, ou bien l'utilisation de la proportion de ces mailles selon l'échelle spatiale étudiée.

Pour cela, une comparaison de modèles en ne changeant qu'une variable est réalisée. Les variables suivantes sont toujours incluses : zone climatique, latitude et longitude de la maille focale, et l'espèce en effet aléatoire. Les variables comparées sont, pour chaque espèce potentiellement présente :

1. le nombre d'observations dans la maille focale ;
2. le nombre d'observations dans les mailles adjacentes à la maille focale ;
3. le nombre d'observations dans les mailles dans un rayon de 50 km autour du centroïde de la maille focale (échelle régionale) ;
4. le nombre de mailles dans un rayon de 50 km autour du centroïde de la maille focale avec au moins une donnée d'observation dans l'INPN ;
5. la proportion de mailles dans un rayon de 50 km autour du centroïde de la maille focale avec au moins une donnée d'observation dans l'INPN ;

Pour cette comparaison de variables spatiales, les données INPN sont collectées toutes périodes confondues.

L'annexe 1 présente pour chaque modèle les valeurs moyennes des paramètres (modèles construits sur les sous-échantillons d'entrainement) et l'annexe 2 présente la distribution moyenne des prédictions de potentialités de présence, calculées sur les sous-échantillons de test. Les graphiques de densité montrent que tous les modèles présentent deux distributions, avec une discrimination satisfaisante du jeu de données entre les données de présences/absences des espèces potentiellement présentes. Les coefficients des différents paramètres restent dans les mêmes ordres de grandeur entre les différents modèles.

Le tableau 7 présente pour chaque modèle le pourcentage moyen du jeu de donnée hors de la zone de chevauchement des distributions des deux catégories d'observation des espèces potentielles (voir figure 8). Les modèles qui permettent la meilleure discrimination des données sont les modèles 4 et 5, avec 48 % du jeu de données se situant dans une gamme de valeurs prédites hors de la zone de chevauchement. La zone de chevauchement est plus large que la zone visible sur les graphiques en annexe du fait de quelques valeurs x_{min} et x_{max} extrêmes, en trop faibles proportions pour être vues graphiquement.

Tab. 7 Pourcentage moyen du jeu de données hors de la zone de chevauchement des distributions des deux catégories d'observation des espèces potentielles pour chacun des cinq modèles comparant l'échelle et la résolution spatiale.

Modèle	Variable spatiale testée	Pourcentage moyen (± erreur standard) du jeu de données hors de la zone de chevauchement
1	Nombre d'observations dans la maille focale	32 % ± 20
2	Nombre d'observations dans les mailles adjacentes à la maille focale	26 % ± 17
3	Nombre d'observations dans les mailles dans un rayon de 50 km autour du centroïde de la maille focale	35 % ± 18
4	Nombre de mailles dans un rayon de 50 km autour du centroïde de la maille focale avec au moins une donnée d'observation dans l'INPN	48 % ± 20
5	Proportion de mailles dans un rayon de 50 km autour du centroïde de la maille focale avec au moins une donnée d'observation dans l'INPN	48 % ± 20

Entre les deux modèles avec le meilleur pouvoir de discrimination, le modèle 5 paraît être le plus pertinent parce que la variable relative à l'échelle spatiale est une proportion. En effet, cela permet de limiter les effets de bordure en donnant aux mailles frontalières et littorales le même poids dans l'analyse que les autres mailles métropolitaines.

3.2 Échelle temporelle d'analyse

Une fois choisie l'échelle spatiale d'analyse des données d'observation INPN, la composante temporelle est à déterminer pour la proportion de mailles dans un rayon de 50 km autour du centroïde de la maille focale avec au moins une donnée d'observation INPN des espèces potentielles. Le choix se porte sur (i) la totalité de ces données disponibles dans les bases de données (modèle 5), (ii) les données de moins de 10 ans, (iii) les données entre 10 et 20 ans et (iv) les données de plus de 20 ans.
Le modèle 5 ayant été construit à partir de toutes les données, il suffit de modifier ce modèle en testant les différents pas de temps.

L'annexe 3 présente les valeurs moyennes des coefficients pour les différentes variables explicatives de chaque modèle, et l'annexe 4 un exemple de la distribution des probabilités de présence pour un des sous-échantillons de test. Les trois pas de temps comparés présentent des résultats semblables, ainsi qu'avec le modèle 5.

Le tableau 8 présente, pour chaque modèle comparant les variables temporelles, la proportion moyenne du jeu de données hors de la zone de chevauchement des aires des catégories d'observation des espèces potentielles. Le modèle avec la plus grande proportion du jeu de donnée hors de la zone de chevauchement des prédictions, et donc la meilleure discrimination, est le modèle incluant les données de moins de 10 ans.

Tab. 8 Pourcentage moyen du jeu de données de test hors de la zone de chevauchement des distributions des deux catégorie d'observation des espèces potentielles pour chacun des 3 modèles comparant l'échelle temporelle des données d'observation INPN.

Modèle	Variable temporelle testée	Pourcentage moyen (± erreur standard) du jeu de données hors de la zone de chevauchement
8	Données INPN < 10 ans	52 % ± 21
9	Données INPN entre 10 et 20 ans	42 % ± 21
10	Données INPN > 20 ans	47 % ± 20

Les modèles testés sur les trois pas de temps sont assez équivalents. Les résultats de la modélisation s'accordent avec la logique écologique : les données les plus récentes sont les plus pertinentes, afin de produire une prédiction la plus à jour possible.

3.3 Caractéristiques du modèle sélectionné

Le modèle choisi présente donc comme variable spatiale des données d'observations INPN la proportion de mailles dans un rayon de 50 km autour du centroïde de la maille focale avec au moins une donnée d'observation, et comme variable temporelle les données d'observation INPN de moins de 10 ans (période 2009 -2019).

Le modèle est donc appliqué à l'ensemble du jeu de données, afin de calculer les prédictions pour toutes les espèces. Le tableau 9 résume les paramètres des variables à effet fixe du modèle construit à partir de l'ensemble du jeu de données (sans les hétérocères). Les coefficients de chaque variable (colonne « Estimation ») sont extraits pour confronter le modèle à d'autres données réelles (sous-partie 3.6). Les coefficients aléatoires de chacune des 2287 espèces potentielles sont également extraits.

Afin de rendre le résultat de la prédiction plus perceptible, la gamme de valeurs prédites est conscrite dans l'intervalle [0,1]. Cette transformation modifie le rendu graphique de la fréquence d'occurrences potentielles en fonction de la prédiction (*cf.* figure 9), puisqu'une majorité des occurrences reçoivent des valeurs de prédiction très proches de 0 ou très proches de 1. La distribution des occurrences potentielles de l'ensemble du jeu de données selon le critère « observée ou non sur le terrain » présentée par la figure 9 montre la discrimination efficace du modèle. Les espèces non observées sur le terrain reçoivent majoritairement des valeurs prédites proches de 0, alors que les espèces observées sur le terrain reçoivent des valeurs prédites proches de 1 (*cf.* figure 9.A). La zone de chevauchement, où l'attribution est incertaine, concerne 18 % des occurrences potentielles (22 794 occurrences qui ne sont comprises ni dans la première barre ni dans la dernière barre de l'histogramme de la figure 9.A).

Tab. 9 Sommaire des variables à effet fixe du modèle sélectionné. Les variables sont, de haut en bas, l'intercept du modèle, la proportion de mailles dans un rayon de 50 km autour du centroïde des mailles focales avec au moins une observation de moins de 10 ans des espèces potentielles, les 6 zones climatiques qui recoupent les mailles focales (ATC = Atlantic Central, CON = Continental, LUS = Lusitanean, MDM = Mediterranean Mountains, MDN = Mediterranean North, MDS = Mediterranean South) (la zone climatique ALS (Alpine South) est la référence pour cette variable discrète, elle a un coefficient de 0 et les paramètres des autres zones sont calculés par rapport à celle-ci), la latitude et la longitude des mailles focales.

Variable	Estimation	Erreur standard	Valeur z	Valeur p	
(Intercept)	7,80191	0,90481	8,623	$< 2x10^{-16}$	***
Prop. mailles régionales 10 ans	2,74587	0,10698	25,666	$< 2x10^{-16}$	***
Zone climatique					
ATC	-0,48874	0,08044	-6,076	$1,23x10^{-9}$	***
CON	-1,74601	0,16704	-10,453	$< 2x10^{-16}$	***
LUS	-0,34978	0,08317	-4,206	$2,6x10^{-5}$	***
MDM	0,03864	0,04902	0,788	0,431	
MDN	-1,09205	0,06453	-16,923	$< 2x10^{-16}$	***
MDS	-0,0376	0,15759	-0,239	0,811	
Latitude	-3,13946	0,14448	-21,73	$< 2x10^{-16}$	***
Longitude	-0,98073	0,10988	-8,925	$< 2x10^{-16}$	***

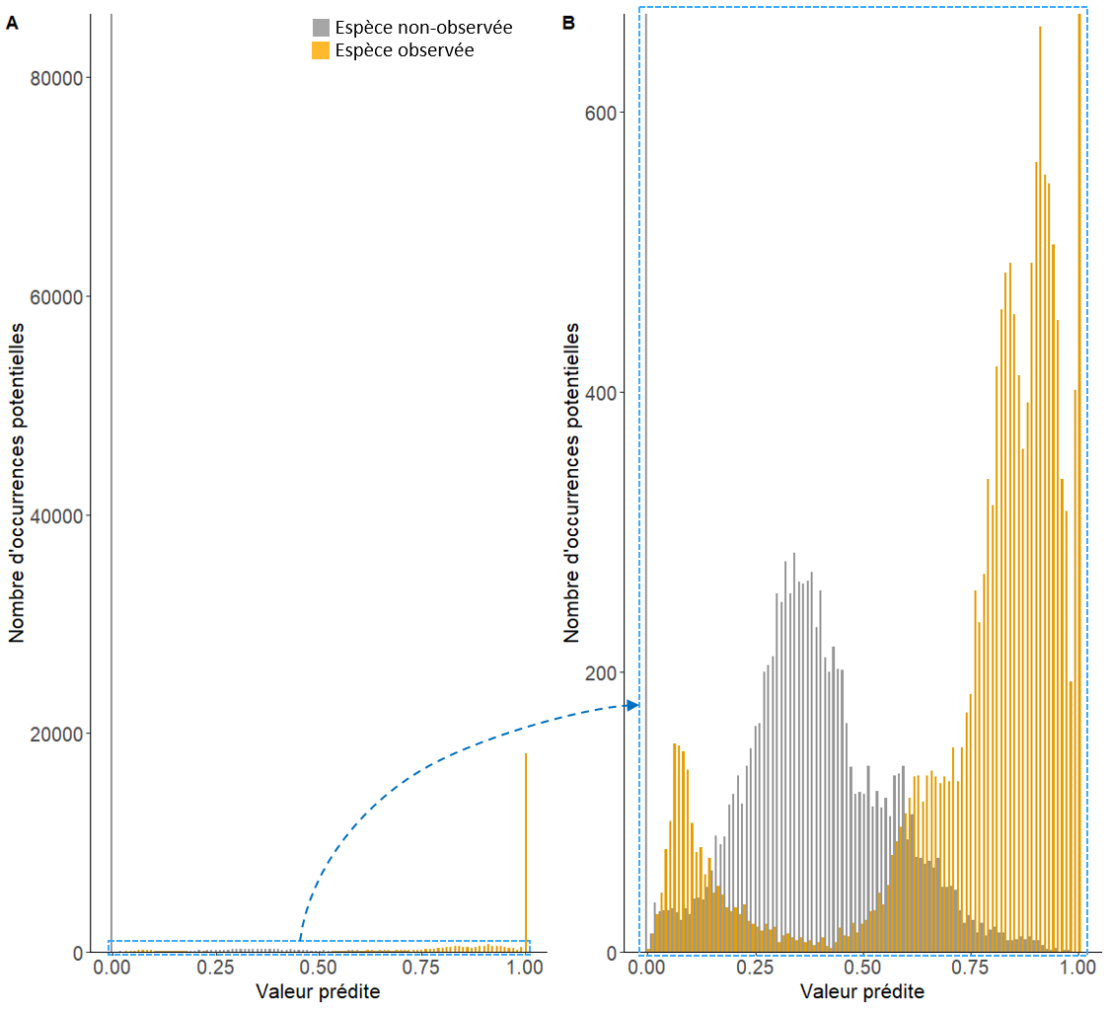


Fig. 9 Histogrammes de la répartition des occurrences potentielles de l'ensemble du jeu de données en fonction de la prédiction calculée et selon la catégorie d'espèces observée ou non pendant les inventaires sur le terrain. A : graphique réalisé avec l'ensemble des données ; B : zoom sur l'axe des ordonnées pour agrandir les données dans l'encart bleu du graphique A. Le rectangle bleu contient 18 % des occurrences potentielles.

3.4 Détermination graphique des seuils de probabilité de présence

La représentation du nombre d'occurrences potentielles en fonction des valeurs de prédiction (figure 9. A et B) permet de discriminer graphiquement différentes catégories : les occurrences potentielles ayant des valeurs très proches de 0 sont des espèces probablement absentes, les occurrences potentielles ayant des valeurs prédites très proche de 1 sont des espèces probablement présentes, et les occurrences potentielles ayant des valeurs intermédiaires sont en chevauchement.

Ces seuils permettent de scinder le jeu de données, avec aux extrémités du gradient de prédiction deux groupes homogènes et correctement prédits, qui rassemblent 82 % des occurrences potentielles du jeu de données complet. Entre les deux, la zone de chevauchement des distributions peut être divisée en deux parties, afin de séparer un groupe de valeurs prédites pour lesquelles il y a plus de chances que l'espèce n'ait pas été observée sur le terrain (zone 2 de la figure 10), plus proche de 0, et un groupe de valeurs prédites pour lesquelles il y a plus de chances que l'espèce ait été observée sur le terrain (zone 3 de la figure 10). La zone 2 présente une incohérence sur l'intervalle [0,005 ; 0,15], où des espèces observées sur le terrain reçoivent de faibles valeurs de prédiction.

Cette lecture graphique donne lieu à la proposition de catégorisation présentée en figure 10. Les valeurs seuils ont été choisies en utilisant la fonction « ggplot_build » (package R *ggplot2*, Wickham, 2016) qui permet d'extraire les valeurs de fréquence pour chaque barre de l'histogramme. Ces seuils correspondent aux coordonnées en abscisse des barres qui se trouvent en limite de proportion majoritaire de chacune des catégories « observées ou non sur le terrain ». Les seuils des deux zones extrêmes ont été choisis de manière à avoir des groupes homogènes, *i.e.* composés uniquement d'occurrences observées sur le terrain ou bien non-observées sur le terrain (deux occurrences potentielles de deux espèces de rhopalocères sont cependant observées sur le terrain mais classées en catégorie « très probablement absente ». Ces choix sont corroborés par la représentation de la proportion d'occurrences potentielles dans chaque catégorie (*cf.* figure 11).

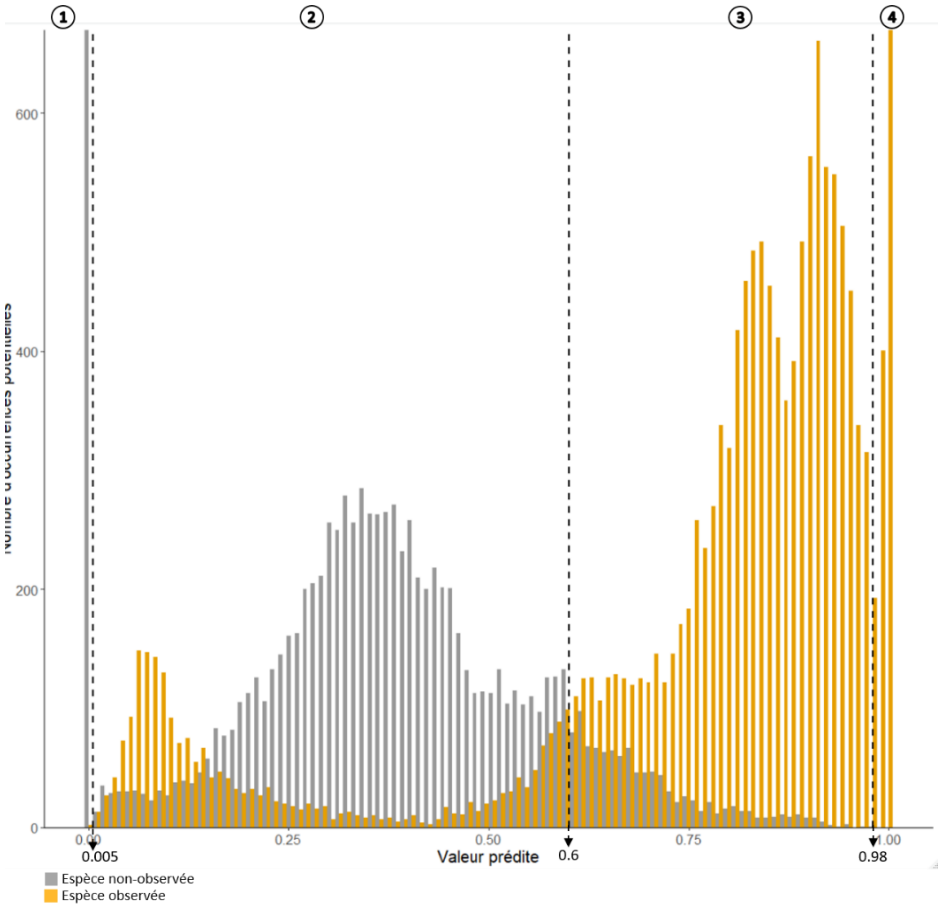


Fig. 10 Graphique de la répartition des occurrences potentielles en fonction de la valeur prédite selon l'état de l'espèce sur le terrain (présente ou non dans les inventaires des sites) (graphique identique à la figure 9.B). Les lignes pointillées verticales correspondent aux trois valeurs seuils qui délimitent quatre catégories de probabilités de présence : (1) espèce très probablement absente pour des valeurs prédites comprises dans l'intervalle [0 ; 0,005[, (2) espèce probablement absente : [0,005 ; 0,6[, (3) espèce probablement présente : [0,6 ; 0,98[, (4) espèce très probablement présente : [0,98 ;1].

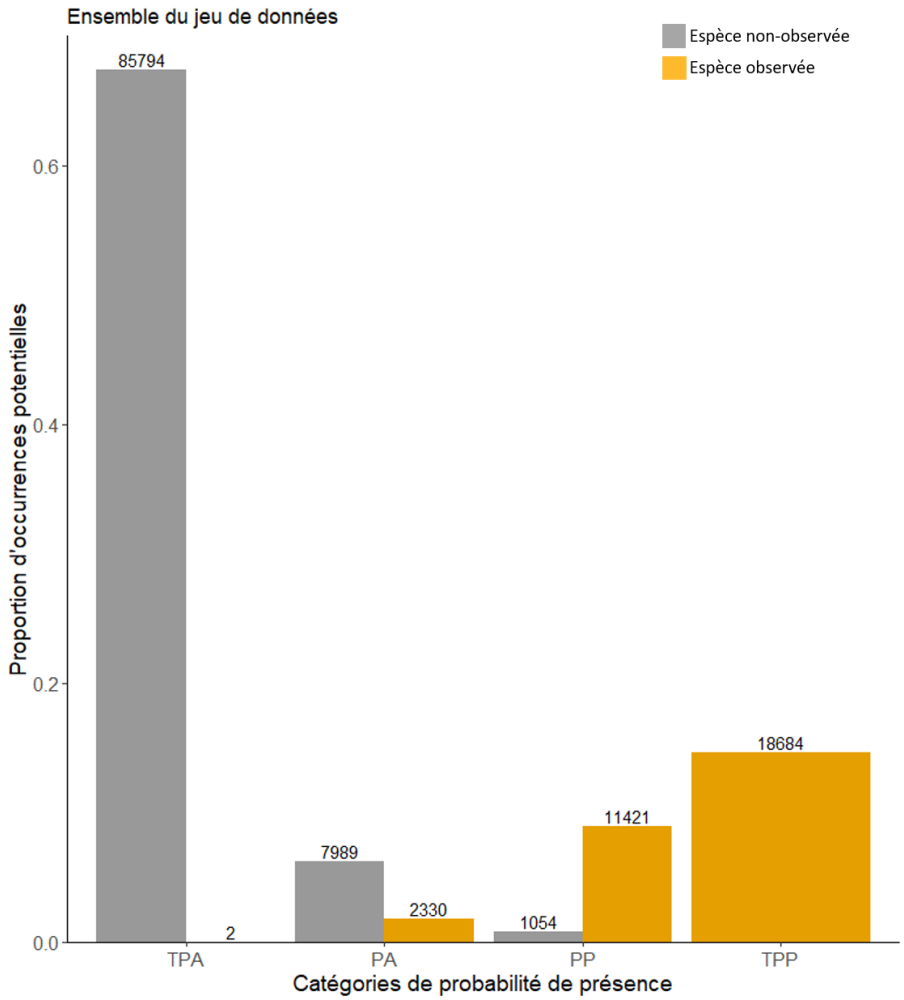


Fig. 11 Proportions d'occurrences potentielles attribuées à chaque catégorie de probabilité de présence (valeurs seuils présentées dans la figure 10). Les chiffres au-dessus des barres indiquent le nombre d'occurrences potentielles dans chaque catégorie. Signification des sigles de probabilité de présence : TPA : espèce très probablement absente ; PA : espèce probablement absente ; PP : espèce probablement présente ; TPP : espèce très probablement présente dans les inventaires de terrain.

3.5 Diagnostic de la qualité de discrimination du modèle

Le diagnostic de la qualité de discrimination examine si le modèle discrimine différemment les espèces selon les critères suivant : groupe taxonomique de l'espèce, patrimonialité[3], groupe taxonomique considéré comme connu ou non dans la maille focale, existence de données d'observation pour l'espèce dans la maille focale depuis moins de 10 ans ou non. Il peut en effet être attendu des divergences d'attribution selon ces différents facteurs : plus d'erreurs d'attributions pour des espèces non patrimoniales, habituellement moins bien connues, ou pour des taxons localement moins échantillonnés par exemple. Ou encore, un biais d'attribution pour des espèces avec des données d'observation datant de moins de 10 ans dans la maille focale comparé à des espèces sans observation locale dans les bases de données INPN.

La première étape consiste à examiner si les espèces potentielles ont été observées ou non pendant les inventaires IQE/IPE, afin de connaître la quantité de faux-positifs (espèces prédites présentes et non observées sur le terrain) et de faux-négatifs (espèces prédites absentes mais observées sur le terrain). Les sous-parties qui suivent s'attachent à décrire les attributions selon les différents critères de discrimination.

[3] Les critères choisis dans cette étude pour définir une espèce patrimoniale sont précisés au chapitre 2.2.3.

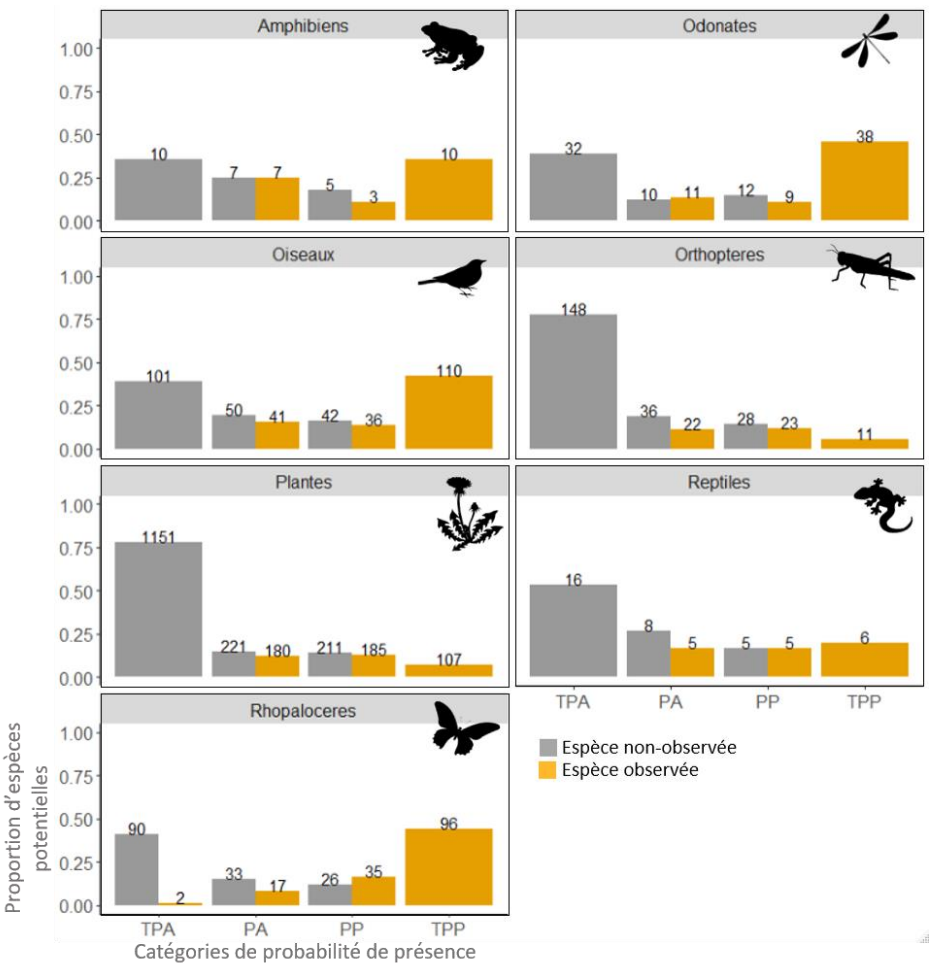
3.5.1 Matrice de confusion et répartition des espèces selon les groupes taxonomiques

Dans un premier temps, une matrice de confusion permet de trier les attributions des occurrences potentielles aux différentes catégories de probabilité de présence selon l'observation ou non des espèces pendant les inventaires IPE/IQE, afin de mettre en évidence la part de bonnes et de mauvaises attributions (tableau 10).

Tab. 10 Matrice de confusion du modèle prédictif construite à partir du pourcentage d'occurrences potentielles dans chaque catégorie de probabilité de présence. Les colonnes « probablement présente » cumulent les occurrences potentielles classées en catégories « très probablement présente » et « probablement présente », et les colonnes « probablement absente » cumulent les occurrences potentielles des catégories « très probablement absente » et « probablement absente ». Les cases grisées correspondent aux bonnes attributions, les cases blanches correspondent aux mauvaises attributions.

		Modèle prédictif	
		Probablement présente	Probablement absente
IPE/IQE	Observée	23,7 %	0,8 %
	Non-observée	1,8 %	73,7 %

Le tableau 10 montre que plus de 97 % des occurrences potentielles sont correctement attribuées. Le nombre d'espèces par catégorie est moins informatif parce qu'une espèce peut être comptée dans plusieurs catégories, selon le contexte régional des mailles où elle a été observée par exemple. Ainsi, si 85 % des espèces sont attribuées à une seule catégorie de probabilité de présence, respectivement 14 % et 0,8 % sont attribuées à deux ou trois catégories de probabilité de présence.



La répartition du nombre d'espèces dans chaque catégorie selon le groupe taxonomique (*cf.* figure 12) donne de précieuses informations sur les qualités de discrimination du modèle. Les espèces de certains groupes sont mieux discriminées, avec une plus grande proportion d'espèces observées sur le terrain dans la catégorie « très probablement présente » que dans les catégories incertaines, comme les rhopalocères, les oiseaux ou les odonates, comparé aux espèces de reptiles, de plantes ou d'orthoptères.

Fig. 12 Répartition de la proportion d'espèces dans chaque catégorie de probabilité de présence selon le groupe taxonomique. Les proportions sont calculées selon le nombre d'espèces de chacun des 7 groupes. Les valeurs au-dessus des barres représentent le nombre d'espèces dans chaque catégorie. À noter qu'une espèce peut être comptabilisée dans plusieurs catégories.

3.5.2 Critère de patrimonialité

La répartition des occurrences potentielles dans les différentes catégories de probabilité de présence selon le critère de patrimonialité de l'espèce suit la même logique pour les espèces patrimoniales et non patrimoniales. En terme de proportions d'occurrences potentielles bien ou mal attribuées, les proportions sont similaires à celles du jeu de données complet (tableau 11).

Il n'y a *a priori* pas d'effet du critère de patrimonialité sur la prédiction de présence. La connaissance plus approfondie de ces espèces ne permettrait donc pas une meilleure prédiction de présence.

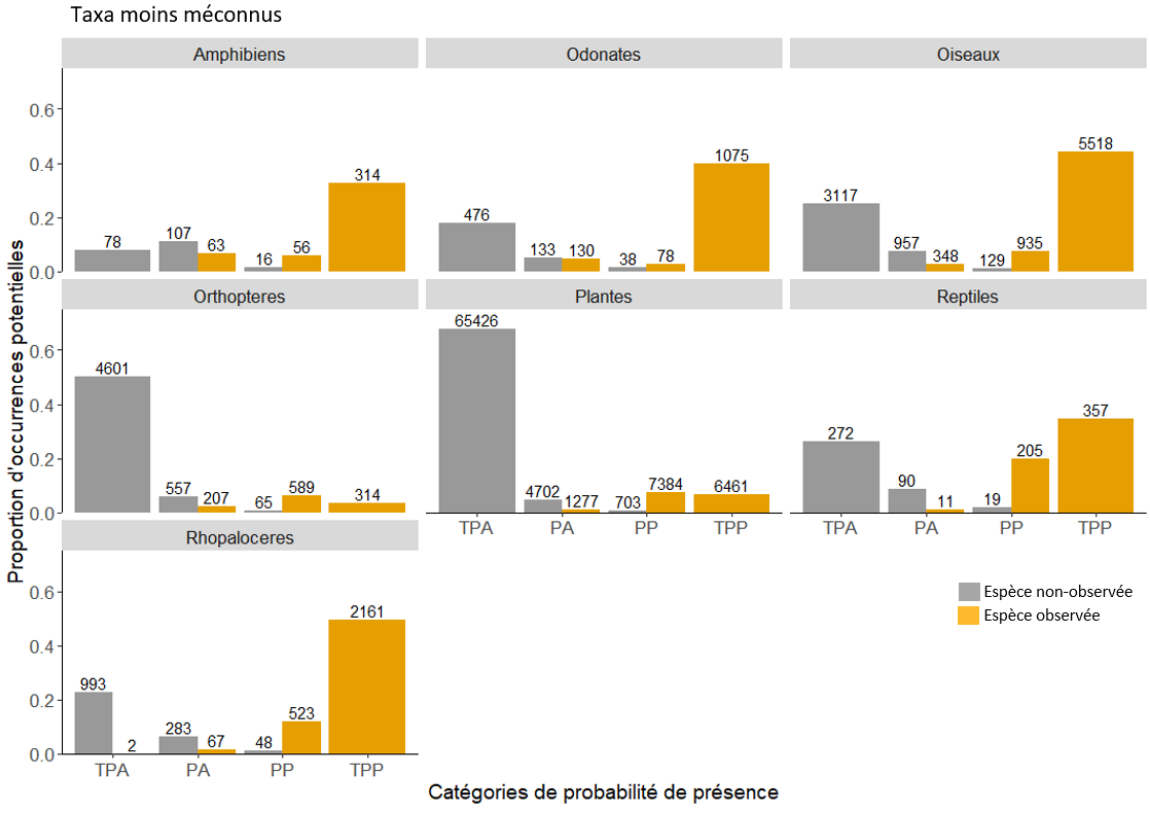
Tab. 11 Matrice de confusion selon le facteur de patrimonialité de l'espèce. Les proportions sont calculées par rapport au nombre d'occurrences potentielles correspondant à des espèces patrimoniales (n= 105 756 occurrences) ou non patrimoniales (n= 21 518 occurrences), respectivement.

Espèces patrimoniales	Probablement présente	Probablement absente
Observée (1)	23,6 %	1,8 %
Non observée (0)	0,8 %	73,6 %

Espèces non patrimoniales	Probablement présente	Probablement absente
Observée (1)	23,7 %	1,9 %
Non observée (0)	0,7 %	73,6 %

3.5.3 Critère de méconnaissance locale du taxon

La distribution des occurrences potentielles selon le critère de méconnaissance du groupe taxonomique à l'échelle de la maille focale est présentée figure 13. Cette représentation permet de montrer que tous les groupe sont majoritairement considérés comme ayant au moins une connaissance minimale à l'échelle de la maille focale. Concernant la répartition des occurrences potentielles entre les différentes catégories de probabilité de présence, elle est similaire avec celle du jeu de données complet, ainsi qu'entre les groupes. Les orthoptères et les plantes semblent moins bien discriminés que les autres groupes, que ces taxa soient considérés méconnus ou non dans les mailles focales.



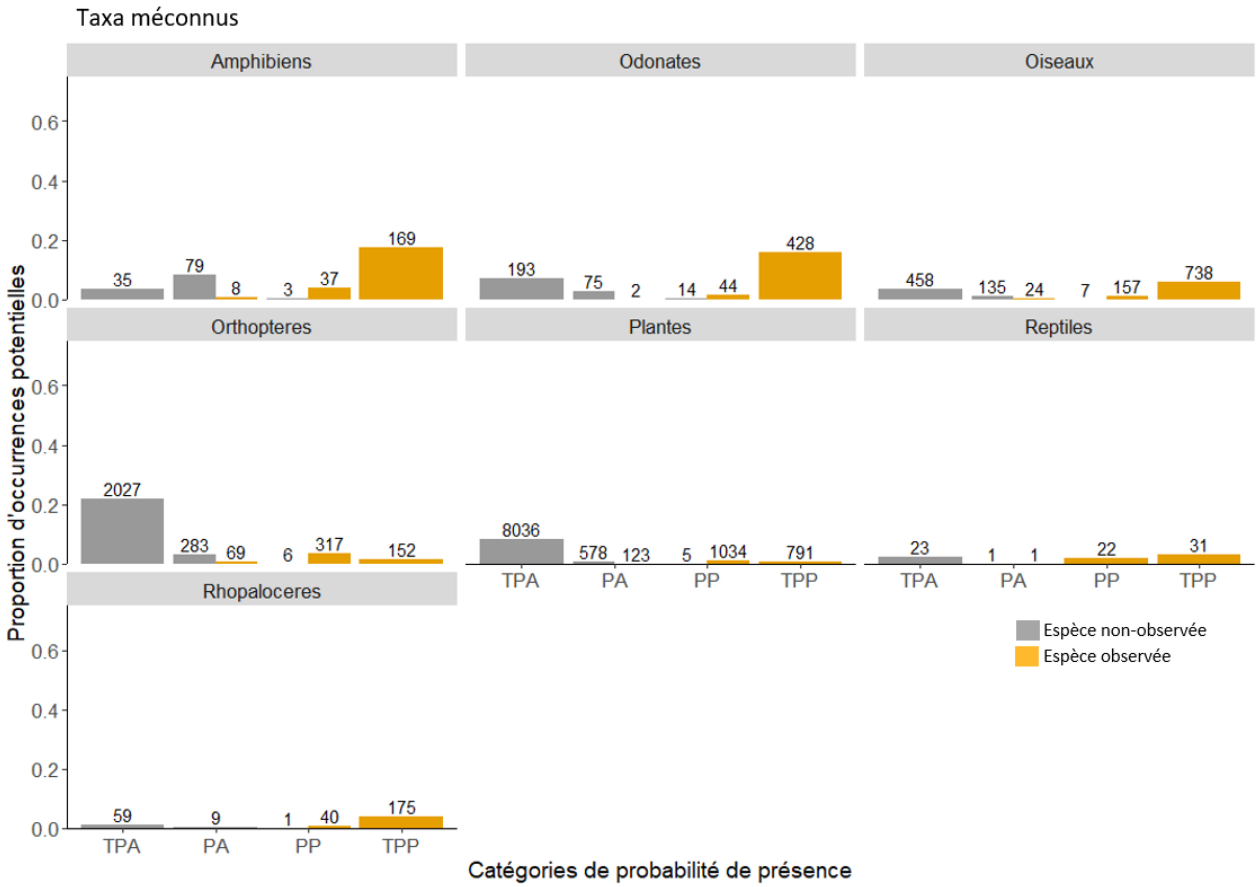


Fig. 13 Répartition des occurrences potentielles attribuées à chaque catégorie de probabilité de présence par groupe taxonomique et selon le critère de méconnaissance du groupe par maille. Les proportions sont calculées par rapport au nombre d'occurrences potentielles par groupe taxonomique, en sachant qu'un taxon peut être moins méconnu dans certaines mailles et méconnu dans d'autres.

3.5.4 Critère de présence de l'espèce dans la maille

La figure 14 présente la distribution des occurrences potentielles en fonction de l'existence de données d'observations INPN de moins de 10 ans pour l'espèce cible dans la maille focale, pour chaque catégorie de probabilité de présence. La majorité des occurrences potentielles correspondent à des espèces sans données sur la maille focale et non observées sur le terrain. La logique d'attribution des occurrences potentielles à chaque catégorie de probabilité de présence est similaire entre les deux groupes, et avec celle du jeu de données au complet. Les espèces observées sur le terrain sans données d'observations INPN sur la maille sont plus nombreuses en catégories incertaines (probablement absente ou probablement présente) qu'en catégorie « très probablement présente », par rapport aux espèces avec des données d'observation INPN sur la maille et observées sur le terrain.

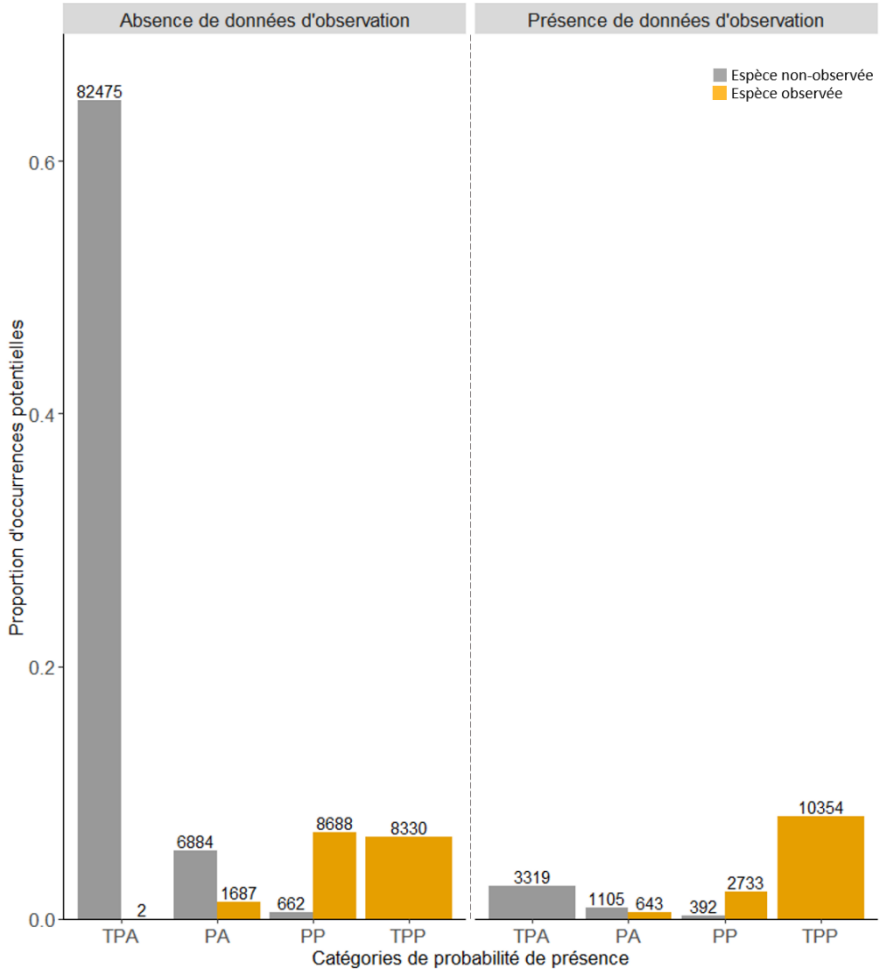


Fig. 14 Proportion des occurrences potentielles dans les différentes catégories de probabilité de présence selon le critère de présence ou d'absence de données d'observation de l'espèce dans la maille focale dans les données INPN datant de moins de 10 ans. Les proportions sont calculées à partir du nombre total d'occurrences potentielles du jeu de données. Les valeurs au-dessus des barres correspond au nombre d'occurrences potentielles dans chaque catégorie.

3.5.5 Conclusion sur le diagnostic

Quatre critères ont été testés pour mettre en évidence d'éventuelles disparités dans les prédictions de présence. La patrimonialité de l'espèce ne semble pas affecter les prédictions. Certains groupes taxonomiques (orthoptères, reptiles et plantes) sont moins bien discriminés que d'autres, avec une plus grande proportion d'occurrences potentielles d'espèces observées sur le terrain attribuées en catégories intermédiaires de probabilité de présence. Ce résultat se retrouve pour les orthoptères et les plantes dans l'investigation du critère de méconnaissance locale des groupes taxonomiques. Ces deux groupes semblent moins bien discriminés par rapport aux autres groupes taxonomiques, qu'ils soient considérés comme connus ou méconnus localement. Enfin, le critère d'existence de données d'observations INPN des espèces potentielles dans la maille focale montre que les occurrences potentielles sont attribuées aux catégories intermédiaires de probabilité de présence si les espèces potentielles n'ont pas d'observations dans la maille focale. A l'inverse, les espèces avec des données d'observation INPN dans la maille focale ont très peu d'occurrences potentielles attribuées en catégories « très probablement absente » et une plus grande proportion classée en catégorie « très probablement présente ».

Pour conclure, les critères utilisés pour évaluer la qualité de discrimination du modèle prédictif donnent des résultats généralement concordants. Cela conforte l'idée que cette première ébauche de modèle prédictif est relativement performante compte tenu de sa simplicité.

3.5.6 Cas des espèces mal attribuées

Malgré la constance des résultats apportés par les différents facteurs de discrimination disponibles, il reste à éclaircir les raisons de la mauvaise attribution de certaines occurrences potentielles. Les espèces prédites présentes mais non-observées sur le terrain sont peu préoccupantes, car il n'est pas possible de savoir si l'espèce est réellement absente du site ou si elle n'y a pas été observée pendant l'inventaire. Le cas des espèces observées sur le terrain mais prédites comme probablement absentes par le modèle doit en revanche être examiné.

Cette sous-partie du jeu de données concerne 2 332 occurrences potentielles et 283 espèces, qui se distribuent ainsi : 7 amphibiens, 11 odonates, 41 oiseaux, 22 orthoptères, 17 rhopalocères, 5 reptiles et 180 angiospermes. Ces occurrences sont réparties sur 42 mailles différentes et dans 6 zones climatiques (toutes sauf Mediterranean South). Les paragraphes suivants s'attachent à explorer certaines hypothèses pour comprendre les causes de ces mauvaises attributions.

1- La proportion de mailles régionales avec des données d'observation INPN pour l'espèce

Les attributions erronées pourraient provenir d'un biais du facteur ayant le plus fort pouvoir explicatif, à savoir, pour chaque occurrence potentielle, la proportion de maille régionales avec au moins une donnée d'observation INPN. Une première hypothèse serait donc que les occurrences potentielles mal attribuées (espèce observée sur le terrain mais prédite absente) pourraient avoir de plus faibles proportions de mailles régionales avec des données d'observation INPN, par rapport aux occurrences correctement attribuées.
L'annexe 5 présente cette proportion en fonction des catégories de prédiction de présence et selon la bonne ou mauvaise attribution des occurrences. Les espèces observées mais prédites absentes n'ont pas de plus faibles proportions de mailles régionales avec des données d'observation INPN par rapport aux espèces correctement attribuées. Cette hypothèse n'est donc pas vérifiée et ce facteur ne semble pas impliqué dans l'attribution erronée des occurrences potentielles.

2- Indice de prédiction spécifique

Un indice permet d'identifier les espèces avec le plus d'erreurs d'attributions (comprenant à la fois les espèces observées mais prédites absentes et les espèces non-observées mais prédites présentes), sachant que 1 902 espèces, soit 83 %, n'ont pas d'erreurs d'attribution. Pour les 385 espèces concernées, cet indice de prédiction a été calculé de la manière suivante :

$$indice\ de\ prédiction\ spécifique = 1 - \frac{\Sigma\ \text{occurrences potentielles mal attribuées}}{\Sigma\ \text{occurrences potentielles de l'espèce}}$$

Plus l'indice de prédiction spécifique est proche de 1 plus les occurrences potentielles de l'espèce sont correctement attribuées (en absence ou en présence).

La figure 15 présente les résultats de cet indice par groupes taxonomiques, pour les 385 espèces pour lesquelles l'indice est calculable. Les indices de prédiction spécifique des espèces ayant été observées au moins une fois sur le terrain mais dont la probabilité de présence est dans l'intervalle [0,005 ; 0,15] sont mis en évidence en rouge. Ce sont les occurrences incohérentes identifiées sur la figure 10 (espèces observées sur le terrain mais prédites absentes). Il apparaît que les espèces présentant ces occurrences ont en majorité des indices de prédiction plus faibles que les autres espèces présentant des erreurs d'attributions. Pour les espèces de flore, d'oiseaux, de rhopalocères et d'orthoptères, ces occurrences proviennent d'espèces avec des valeurs extrêmes par rapport à la distribution de l'indice de prédiction pour leur groupe taxonomique.

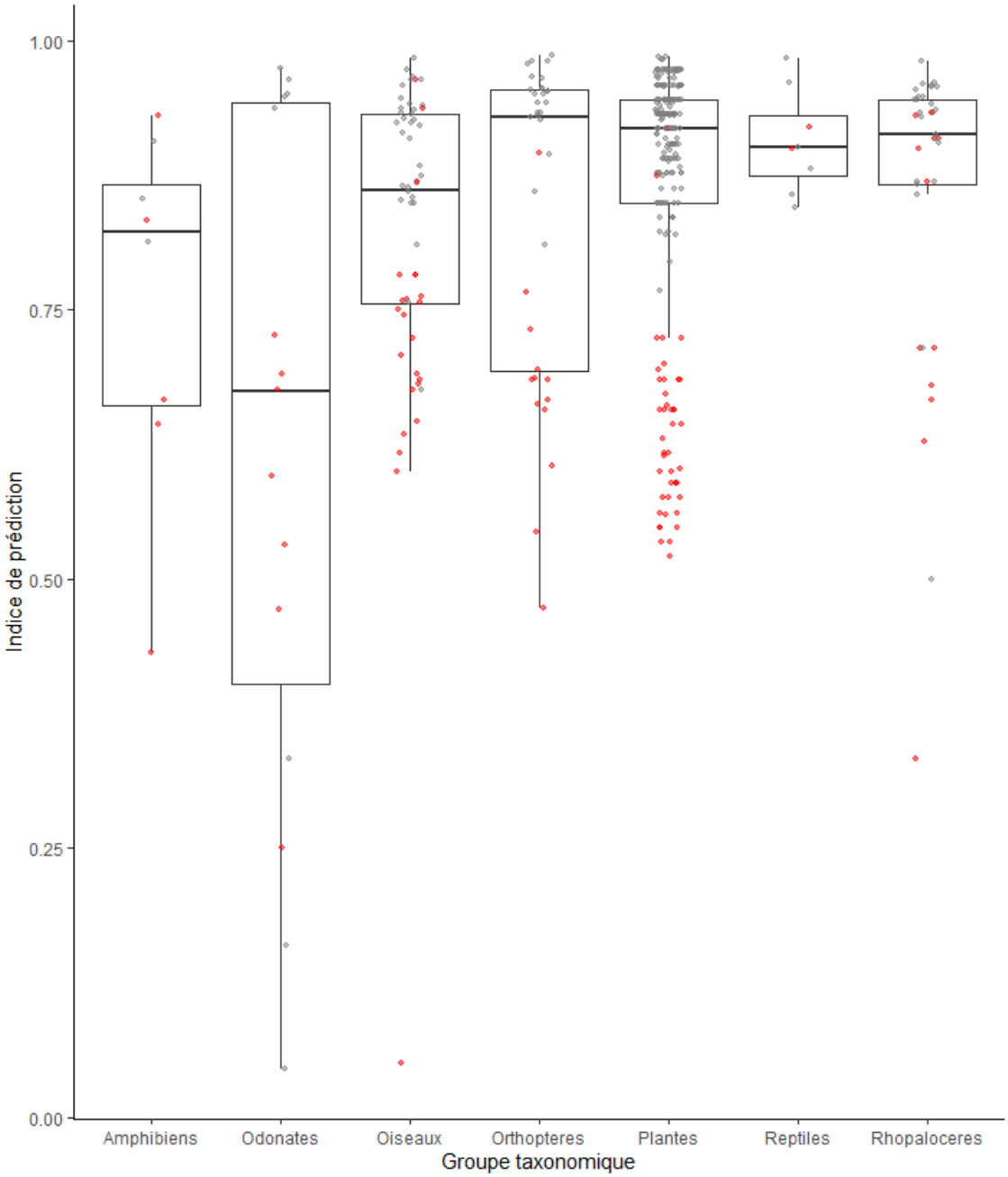


Fig.15 Distribution de l'indice de prédiction spécifique par groupe taxonomique. Les boites à moustaches figurent la médiane (barre noire), le premier et le dernier quartile de la distribution au sein de chaque groupe. Les points correspondent aux valeurs spécifiques de chacune des 385 espèces présentant des erreurs d'attributions. En rouge sont figurées les valeurs de l'indice des espèces dont au moins une occurrence potentielle a été observée sur le terrain mais prédite dans l'intervalle [0,005 ; 0,15].

3- Corrélation entre l'indice de prédiction et la proportion d'occurrences potentielles observées

Une autre hypothèse pourrait être que, à l'instar des hétérocères, les espèces dotées de très peu d'observations sur le terrain sont moins bien prédites par le modèle. Ainsi, la proportion d'occurrences potentielles pour lesquelles l'espèce a été observée sur le terrain a été calculée pour chaque espèce en faisant le rapport du nombre d'occurrences potentielles pour lesquelles l'espèce a été observée sur le terrain sur le nombre total d'occurrences potentielles pour l'espèce. Cette valeur sera appelée proportion d'occurrences potentielles observées par la suite.

La corrélation potentielle entre l'indice de prédiction et la proportion d'occurrences potentielles observées par espèce est représenté par la figure 16 (toujours pour les espèces qui ont des occurrences potentielles mal

attribuées). Il n'y a pas de corrélation entre les l'indice de prédiction et la proportion d'observations, mais les espèces sont réparties en deux groupes.

Le groupe du haut est caractérisé par des espèces avec un fort indice de prédiction et une forte proportion d'occurrences potentielles observées. Le groupe du bas est caractérisé par des espèces avec un indice de prédiction variable et de faibles proportions d'occurrences potentielles observées. C'est dans ce groupe que se trouvent les espèces dont au moins une occurrence potentielle est observée sur le terrain mais prédite dans l'intervalle [0,005 ; 0,15].

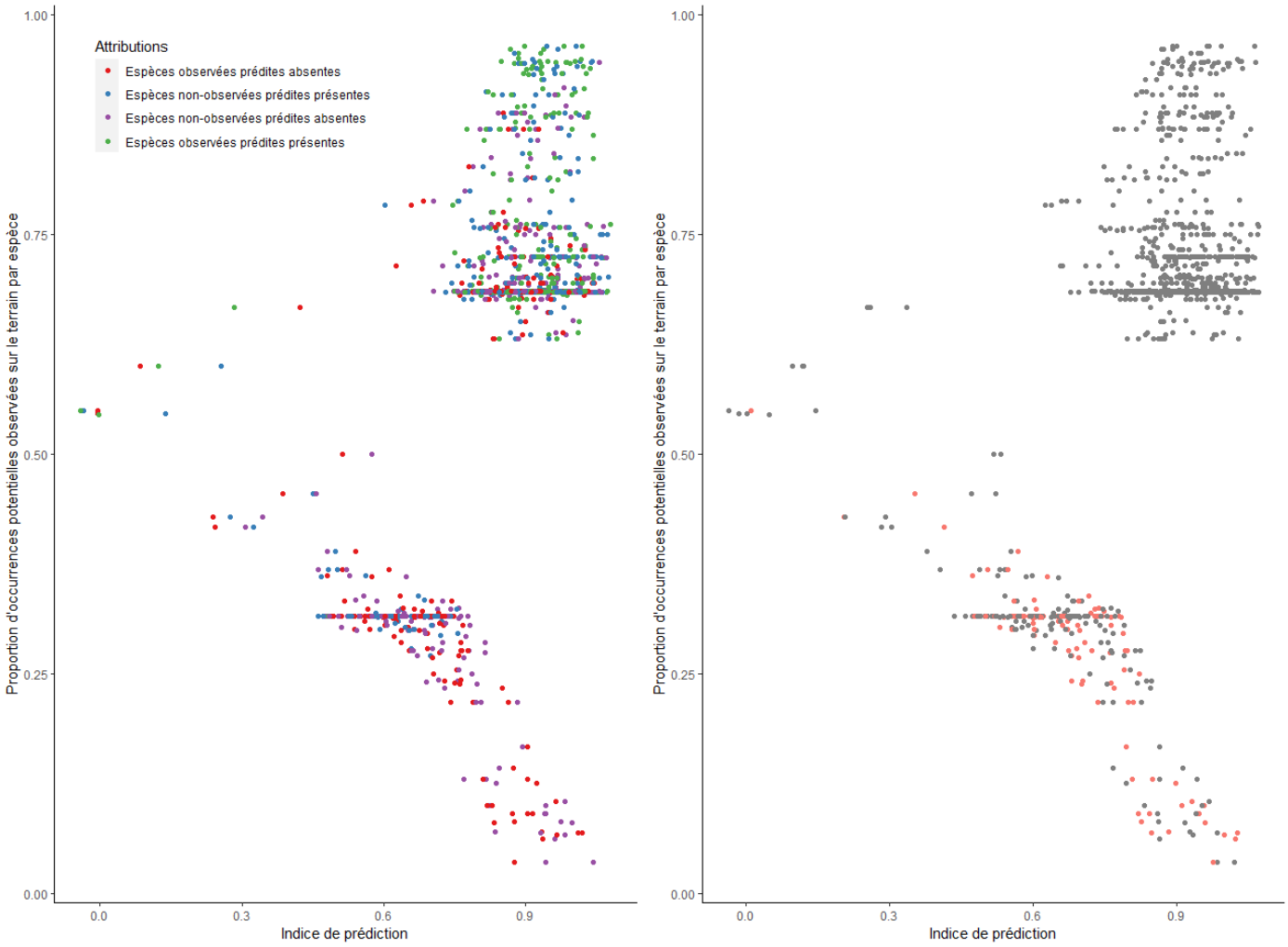


Fig. 16 Répartition des espèces (n = 385) selon l'indice de prédiction spécifique et la proportion spécifique d'observations sur le terrain. Le graphique de gauche présente les données en fonction du type d'attribution des occurrences potentielles. Le graphique de droite figure en rouge les espèces dont au moins une occurrence potentielle a été observée sur le terrain mais prédite dans l'intervalle [0,005 ; 0,15]. Chaque espèce est représentée par plusieurs points (un point par type d'attribution).

Ce résultat permet de corroborer l'hypothèse selon laquelle les espèces peu observées sur le terrain ont plus de chances d'être mal assignées. Même si cela ne concerne qu'une petite partie du jeu de données, ce n'est pas une bonne nouvelle pour les espèces rares, qui sont dans ce cas de figure. Les prochaines versions du modèle devront s'attarder sur cette problématique afin d'améliorer les prédictions.

3.6 Test avec les données du Contournement Ouest de Nîmes

Le Contournement Ouest de Nîmes, dont le maître d'ouvrage est la Direction des Infrastructures de Transport, est un projet de construction d'une 2x2 voies à l'Ouest de Nîmes pour désengorger le trafic routier sur la RN106, autour de laquelle l'urbanisation s'est fortement accrue depuis la fin des années 1990.

Le rapport d'étude d'impact a été diffusé en octobre 2017. L'état écologique initial a été défini par le bureau d'étude Naturalia à partir d'analyses bibliographiques et d'inventaires faunistiques et floristique. La surface d'emprise du projet totalise 1 230 ha dans la banlieue ouest de Nîmes dans le Gard (30), et l'étude d'impact considère les habitats et les espèces dans un tampon de 5 km autour du tracé du projet. Cette surface est répartie sur la maille 10x10 km de France métropolitaine référencée « 10kmL93E080N630 ».

Les inventaires réalisés lors de l'étude d'impact se sont focalisés sur les espèces patrimoniales ou à enjeux, en prenant en compte plus de critères que la présente étude pour la définition de patrimonialité des espèces. La patrimonialité inclut dans cette étude d'impact l'inscription des espèces à la liste des espèces végétales protégées régionalement (Languedoc-Roussillon), les espèces endémiques ou sub-endémiques de France métropolitaine, les espèces en limite d'aire de répartition, les espèces présentant une aire de répartition disjointe et certaines espèces bio-indicatrices. Certains de ces critères sont inaccessibles pour la présente analyse, en particulier ceux qui concernent les aires de répartition des espèces ou la bio-indication. Pour la comparaison ont donc été considérées comme patrimoniales les espèces répondant à au moins un des critères suivants :

- Inscrite à l'annexe I de la Directive Oiseaux ;
- Inscrite à l'annexe II ou IV de la Directive Habitat Faune Flore ;
- Inscrite sur la Liste Rouge Nationale avec un statut « EN », « VU », « CR » ou « CR* » ;
- Inscrite sur la Liste Rouge Régionale de Languedoc-Roussillon avec un statut « EN », « VU », « CR » ou « CR* » ;
- Inscrite sur les listes de Protection Nationales et/ou de Protection Régionale ;
- Considérée comme déterminante ZNIEFF dans la région Languedoc-Roussillon.

Le test consiste à comparer les listes d'espèces inventoriées par l'étude d'impact avec la liste d'espèces potentiellement présentes. La liste d'espèces potentielles est compilée à partir de la liste des habitats inventoriés sur le terrain et listés dans le rapport de l'étude d'impact.

Les habitats décrits en mosaïque dans la typologie EUNIS ont été séparés en chacune de leurs composantes (tableau 12). La base de connaissance des relations espèces – habitats est ensuite filtrée selon cette liste d'habitats inventoriés afin d'extraire la liste d'espèces potentiellement présentes, qui totalise alors 2 866 espèces. En ne conservant que les groupes taxonomiques étudiés pour le présent rapport (pour rappel : amphibiens, angiospermes, odonates, oiseaux, orthoptères, reptiles, rhopalocères), le chiffre descend à 1 827 espèces.

La liste d'espèces potentielles est ensuite affinée en utilisant l'ABDSM : sont exclues les espèces ayant une carte validée par un expert considérées absentes dans le département du Gard (statuts « Absence probable ou certaine » ou « Absence liée à une disparition avérée »). Cela représente un total de 60 espèces : 11 amphibiens, 7 rhopalocères, 7 odonates, 4 orthoptères, 7 reptiles et 24 plantes.
Pour information et pour le diagnostic ultérieur de ce test, 83 % des espèces potentielles sont patrimoniales (1 467 espèces).

Tab. 12 Liste des habitats mentionnés dans l'étude d'impact, après avoir isolé les habitats composant les mosaïques paysagères.

Code EUNIS	Description de l'habitat
I1	Cultures et jardins maraîchers
C1.3	Lacs, étangs et mares eutrophes permanents
C1.6	Lacs, étangs et mares temporaires
F3.1	Fourrés tempérés
E2.1	Pâturages permanents mésotrophes et prairies de post-pâturage
F6.1	Garrigues occidentales
FB.4	Vignobles
G3.F	Plantations très artificielles de conifères
J1.1	Bâtiments résidentiels des villes et des centres-villes
J1.4	Sites industriels et commerciaux en activité des zones urbaines et périphériques
E5.12	Communautés d'espèces rudérales des constructions urbaines et suburbaines récemment abandonnées
G3.74	Pinèdes à *Pinus halepensis*
F6.11	Garrigues occidentales à *Quercus coccifera*
F6.1D	Garrigues occidentales à *Helianthemum* et *Fumana*
E1.31	Pelouses xériques ouest-méditerranéennes
F5.13	Matorrals à *Juniperus*
G2.91	Oliveraies à *Olea europaea*
G1.D4	Vergers d'arbres fruitiers
J5.31	étangs et lacs à substrat entièrement artificiel
I1.52	Jachères non inondées avec communautés rudérales annuelles
G2.121	Chênaies à *Quercus ilex* mésoméditerranéennes

L'étape suivante consiste à récupérer pour chaque espèce potentielle les données d'observation dans l'INPN, et la proportion de mailles avec au moins une observation sur la période 2007-2017 (10 ans avant l'inventaire) dans un rayon de 50 km autour du centroïde de la maille focale. La zone bioclimatique est également ajoutée au jeu de données (Mediterranean South), ainsi que les valeurs de longitude et de latitude du centroïde de la maille focale. Les coefficients aléatoires de chaque espèce, conservés lors du calcul par le modèle général sélectionné dans la partie précédente (*cf.* tableau 9), sont attribués.

La potentialité de présence de chaque espèce est ensuite calculée en utilisant les coefficients du modèle général selon la formule suivante :

$$Potentialité\ de\ présence = \frac{e^{prédicteur}}{1 + e^{prédicteur}}$$

Sachant que :

$$prédicteur = \alpha \times données\ INPN + \beta + \gamma \times latitude + \delta \times longitude + \varepsilon + C$$

Avec :

- α le coefficient du paramètre données INPN,
- β le coefficient de la zone climatique de la maille focale,
- γ le coefficient du paramètre latitude,
- δ le coefficient du paramètre longitude,
- ε le coefficient aléatoire de l'espèce cible,
- C l'intercept du modèle.

Après application de la formule, les espèces potentielles sont réparties dans les différentes catégories de probabilité de présence ainsi : 67 % en « très probablement absente » (n = 1 154 espèces potentielles), 4,5 % en « probablement absente » (n = 79), 14 % en « probablement présente » (n = 248) et 14,5 % en « très probablement présente » (n = 256). Comme ce jeu de données est construit sur une seule maille, il n'y a qu'une occurrence potentielle par espèce et chaque espèce est attribuée à une seule catégorie de probabilité de présence.

Le jeu de données est ensuite comparé à la liste des espèces observées pendant l'étude d'impact. Cette liste compte 394 espèces, parmi lesquelles 212 (54 %) sont prédictibles par le modèle (elles sont renseignées dans la base de connaissance des relations espèces – habitats). Les espèces non prédictibles rassemblent un oiseau, un orthoptère, des lépidoptères (9, dont 8 hétérocères) et des plantes (162). Parmi les 212 espèces prédictibles, 173 (82 %) sont patrimoniales d'après les critères décrits en début de chapitre. Toutefois 30 de ces espèces (dont 25 patrimoniales) sont absentes de la liste finale des espèces potentiellement présentes (2 oiseaux, 1 odonate, 2 orthoptères et 25 rhopalocères) : ceci est dû au fait que leurs habitats renseignés dans la base de connaissance des relations espèces – habitats n'ont pas été mentionnés dans l'étude d'impact.

Un bilan numérique des espèces potentielles prédites présentes (catégories « probablement présente » et « très probablement présente ») et observées est présenté dans le tableau 13.

Tab. 13 Nombre d'espèces par groupes taxonomiques dans la liste d'espèces potentielles prédites présentes et dans la liste des espèces observées pendant l'étude d'impact. La première partie du tableau concerne toutes les espèces et la seconde ne prend en compte que les espèces patrimoniales. Les proportions sont calculées par rapport au nombre d'espèces observées prédictibles [(a)], ou par rapport au nombre d'espèces potentielles prédites présentes [(b)].

	Groupe taxonomique	Nombre d'espèces observées prédictibles	Nombre d'espèces potentielles observées	Pourcentage d'espèces observées potentielles [a]	Nombre total d'espèces potentielles	Pourcentage d'espèces potentielles observées [b]
Toutes les espèces	Amphibiens	6	5	83 %	8	62 %
	Odonates	10	6	60 %	29	21 %
	Oiseaux	64	58	91 %	117	50 %
	Orthoptères	27	6	22 %	26	23 %
	Plantes	48	35	73 %	290	12 %
	Reptiles	8	2	25 %	9	22 %
	Rhopalocères	49	17	35 %	25	68 %
	Total	212	129	61 %	504	26 %
Espèces patrimoniales	Amphibiens	6	5	83 %	8	62 %
	Odonates	3	1	33 %	20	5 %
	Oiseaux	62	54	87 %	110	49 %
	Orthoptères	21	4	19 %	19	21 %
	Plantes	39	27	69 %	241	11 %
	Reptiles	8	2	25 %	9	22 %
	Rhopalocères	34	9	26 %	13	69 %
	Total	173	102	59 %	420	24 %

L'analyse diagnostique des attributions des différentes catégories de probabilité de présence aux espèces selon leur présence ou non dans la liste d'espèces observées lors de l'étude d'impact est présentée dans le tableau 14. Un total de 76 % des espèces potentielles sont correctement prédites par le calcul de potentialité de présence (somme des pourcentages des cases grisées). Les proportions sont identiques si elles sont calculées sur les espèces patrimoniales seules.

Tab. 14 Matrice de confusion des attributions de catégories de probabilité de présence aux espèces potentielles. Les assignations correctes sont dans les cases grisées. Les pourcentages sont calculés par rapport au nombre total d'espèces potentielles.

	Probablement présente	Probablement absente
Observée (1)	129 (7 %)	53 (3 %)
Non observée (0)	375 (21 %)	1210 (69 %)

La répartition des espèces potentiellement présentes dans les différentes catégories de probabilité de présence selon la présence ou l'absence de l'espèce dans la liste d'espèces observées pendant l'étude d'impact est présentée figure 17.

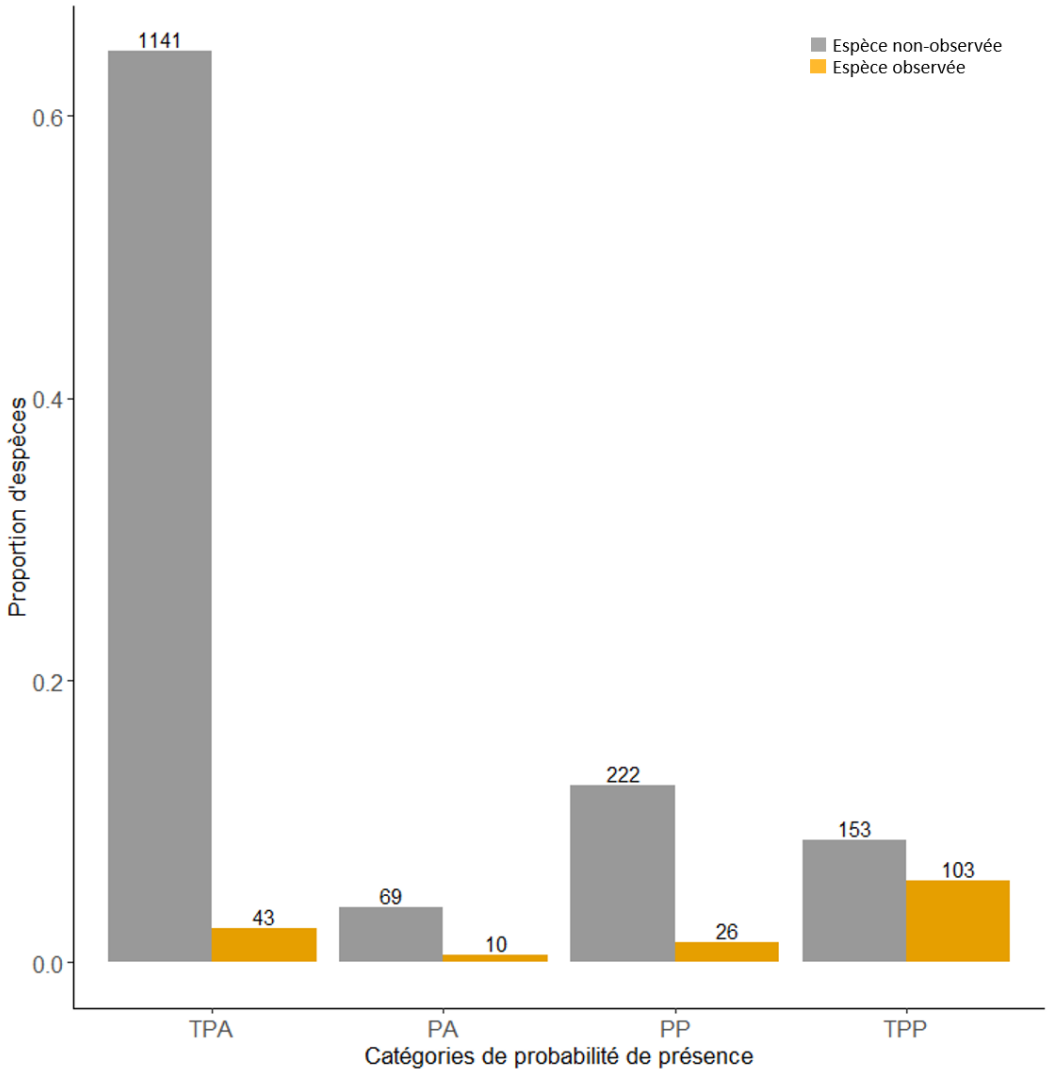


Fig. 17 Distribution des espèces potentielles dans les différentes catégories de probabilité de présence selon leur observation ou non lors de l'étude d'impact. Les proportions sont calculées par rapport au nombre total d'espèces potentielles (1 767). Les valeurs au-dessus des barres représentent le nombre d'espèces dans chaque catégorie. Pour rappel TPA : Très probablement absente, PA : Probablement absente, PP : Probablement présente, TPP : Très probablement présente.

Les résultats sont moins satisfaisants qu'avec le modèle général, avec des espèces observées pendant l'étude d'impact catégorisées comme très probablement absentes, et une plus grande proportion d'espèces non observées pendant l'étude d'impact en catégorie probablement présente.

La figure 18 décline le résultat précédent par groupes taxonomiques. Cette représentation permet de mettre en évidence que les espèces de certains groupes sont un peu mieux prédites. C'est le cas des rhopalocères, amphibiens et oiseaux par rapport aux orthoptères, plantes et reptiles.

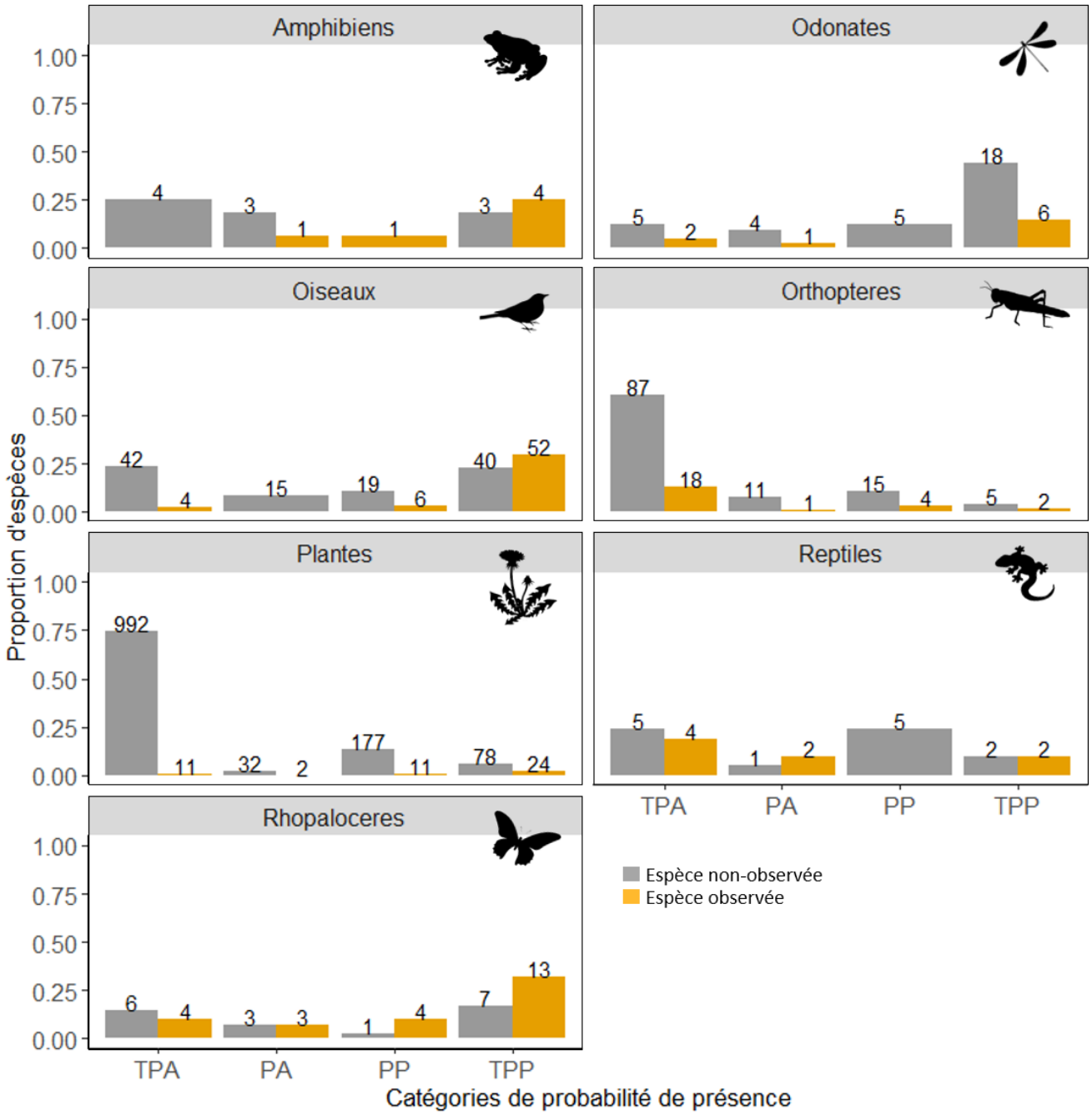


Fig. 18 Proportions d'espèces dans chaque catégorie de probabilité de présence selon l'observation ou non de l'espèce lors de l'étude d'impact, par groupe taxonomique. Les proportions sont calculées par rapport au nombre total d'espèces potentielles dans chaque groupe taxonomique. Les valeurs au-dessus des barres correspondent au nombre d'espèces dans chaque catégorie.

Ce test permet de mettre en évidence un point de vigilance quant à l'utilisation de ce modèle pour des études initiales : la liste des habitats sur le site d'étude doit être la plus complète possible, au risque de ne pas prédire correctement certains groupes inféodés à des milieux particuliers. En effet, un premier test a d'abord été réalisé sans les habitats C1.3 et C1.6 (mentionnés dans le texte et sur les cartes mais qui n'étaient pas la liste globale des habitats de l'étude d'impact). Le nombre d'espèces potentielles d'amphibiens et d'odonates était alors beaucoup plus faible. Le corollaire est cependant peu attrayant, car une longue liste d'habitats produit une liste d'espèces

potentielles très fournie, et dont la majeure partie sera très probablement absente du terrain d'étude. De plus, 50 espèces observées sur le terrain et prédictibles d'après la base de connaissance des relations espèces – habitats n'ont pas pu être prédites parce que l'habitat avec lesquelles elles sont liées n'apparaissent pas dans la liste des habitats inventoriés.

Un parallèle avec les données d'IPE/IQE a été réalisé. Parmi les 53 espèces mal attribuées (observées sur le terrain mais prédites absentes), 43 n'ont aucune observation sur le terrain dans les inventaires IPE/IQE et les 10 espèces qui ont des occurrences potentielles où l'espèce est observée ont des proportions d'observation faibles (inférieurs à 35 %). Il se trouve également que ces 10 espèces ont des occurrences potentielles mal attribuées dans l'intervalle [0,005 ; 0,15]. Il apparait donc que les mauvaises attributions du modèle général se répercutent lors de l'utilisation de la formule du modèle sur d'autres données d'inventaire.

D'autres tests de ce type nécessitent d'être réalisés, en prenant en compte des projets répartis le long d'un gradient d'échelle géographique, mais aussi le long d'un gradient d'artificialisation des paysages inventoriés. La répétition des tests permettra de mieux comprendre l'importance relative des facteurs choisis sur les prédictions de potentialité de présence et de mettre en lumière des points de vigilance quant à l'interprétation des résultats et à l'utilisation de ce modèle prédictif.

4 Limites et perspectives

Les données à disposition pour cette première version d'un outil de prédiction de la potentialité de présence des espèces permettent de vérifier qu'il est possible, avec un modèle simple et l'importante base de données de l'INPN, de dresser une liste d'espèces potentiellement présentes ou absentes pour différents groupes taxonomiques et à grande échelle géographique.

Cette étude permet de pointer du doigt des points de vigilance quant à l'utilisation de cet outil prédictif, en particulier l'importance de bien définir les habitats sur chaque site soumis à l'outil. Une certaine variabilité dans la prédiction existe également selon les groupes taxonomiques. Le critère de patrimonialité de l'espèce n'est pas discriminant, ce qui représente un résultat plutôt encourageant vu l'échelle d'étude et la simplicité de la méthode. Ainsi, même si la portée première de l'outil est de déterminer la potentialité de présence d'espèces patrimoniales, les analyses pourront être réalisées sur des espèces qui ne répondent pas à ce critère sans avoir besoin de les traiter séparément ou d'interpréter les résultats avec encore plus de précautions.

4.1 Limites

Le nombre d'espèces pour lesquelles il est possible de calculer la potentialité de présence est de 2 287, ce qui correspond à 2,6 % des espèces terrestres et/ou dulcicoles de métropole. Bien que cette proportion puisse paraître faible par rapport au nombre d'espèces décrites en France, il s'agit principalement d'espèces patrimoniales, qui sont particulièrement importantes à prendre en compte lors des projets d'infrastructures ou d'autres aménagements. Cette proportion est limitée d'une part par la liste des espèces pour lesquelles les relations espèces – habitats ont été expertisées (liste qui devrait augmenter au fur et à mesure des expertises), et d'autre part par les données des bases utilisées pour la modélisation. En effet pour l'instant seules les espèces observées sur le terrain pendant les inventaires IQE/IPE des sites conservés dans l'analyse sont prédictible, via l'obtention d'une valeur de facteur aléatoire. Cette liste est donc tributaire des mailles focales étudiées, et de la liste d'habitats qui a été dressée pendant l'inventaire (88 % des habitats renseignés dans la base de connaissance sont inclus dans l'analyse, et 56 % des espèces de la base de connaissance des relations espèces – habitats, 95 % si seuls les

amphibiens, odonates, oiseaux, orthoptères, plantes, reptiles et rhopalocères sont comptés). Il serait intéressant de tester si les prédictions restent correctes en incluant des espèces non-observées lors d'inventaires et de leur attribuer un coefficient aléatoire soit nul, soit équivalent à la médiane des coefficients aléatoire du groupe taxonomique par exemple. L'ensemble des espèces de la base de connaissance des relations espèces – habitats seraient alors prédictibles.

L'objectif de l'outil n'est pas d'aboutir à une prédiction exhaustive des espèces présentes ou absentes. Premièrement parce que l'échelle d'étude (mailles 10x10 km de France métropolitaine) est trop grande mais surtout parce que la répartition des espèces et des individus n'est pas seulement tributaire de la présence ou de l'absence des habitats. Elle dépend aussi de la phénologie, des interactions biotiques, des sources de dérangement, et de bien d'autres facteurs difficiles à appréhender et à prédire. Si les espèces les plus communes, rencontrées partout sur le territoire, ont de fortes chances d'être présentes quelle que soit la maille, les difficultés de prédictions sont particulièrement prégnantes pour les espèces rares. En effet, l'étude montre que pour les espèces qui ont été peu observées lors des inventaires IQE/IPE, le modèle a plus de chances d'aboutir à une prédiction d'absence.

D'autre part, la base de connaissance des relations espèces – habitats, qui sert de référence pour cette analyse, est en cours de construction et est majoritairement constituée de dires d'experts. Si cette source d'information est la plus fiable qui ait pu être consultée, elle peut présenter des lacunes (des relations espèces – habitats non-renseignées par exemple, en particulier dans les habitats anthropisés).

4.2 Pistes d'amélioration

Le jeu de données utilisé dans cette analyse est principalement composé d'espèces végétales. Or, parmi les groupes taxonomiques considérés par l'étude, c'est le plus grand clade phylogénétique et les espèces qui le constituent ont de très grande différences écologiques. Il serait probablement judicieux à l'avenir de subdiviser ce grand groupe en familles, ou bien selon certaines préférences ou traits écologiques.

La base de connaissance des relations espèces – habitats est en cours de construction. L'expertise des angiospermes se poursuit, notamment pour les espèces forestières qui font l'objet d'une convention avec l'ONF. Il serait possible d'ajouter à l'outil prédictif certains groupes taxonomiques avec beaucoup de données d'observations dans les bases de données de l'INPN et pour lesquelles les relations espèces – habitats ont été expertisées. C'est le cas par exemple des chiroptères. Les chauves-souris sont en effet le groupe de mammifères le mieux complété dans la base de connaissance de relations espèces – habitats avec 34 espèces dont les relations ont été expertisées sur les 35 espèces présentes en France métropolitaine (il ne manque que *Myotis crypticus*, distingué en 2019 du *Myotis nattereri* sur la base de données génétiques). Il serait également intéressant de pouvoir ajouter les hétérocères, qui ne sont que très peu inventoriés au sein de l'UMS. L'implémentation de ces deux groupes dans les modèles nécessiterait toutefois d'obtenir des jeux de données d'observation à grande échelle géographique.

Les modèles testés lors de cette première ébauche étaient volontairement les plus simples possibles, sans interaction entre les variables explicatives. Des modèles plus complexes pourraient tester ces interactions, par exemple entre la position géographique et la zone climatique. L'altitude peut également avoir une influence sur la distribution de certaines espèces. Ce facteur n'a pas été testé pour cette première ébauche car à l'échelle de la maille 10x10 km les variations d'altitudes peuvent être très grandes, en particulier en région montagneuse. De futures versions de l'outil pourraient tester l'influence de l'altitude sur la prédiction.

Pour cette première version du modèle prédictif, trois échelles spatiales ont été testées pour la variable concernant les données d'observations INPN. Il serait intéressant de tester l'effet de différentes tailles de buffer autour de la

maille focale. L'effet de la taille du buffer sur la prédiction est peut-être lié au groupe taxonomique des espèces, selon la taille moyenne de leur domaine vital et leurs capacités de dispersion.

Une des limitations pour la prédiction de présence des espèces, mais qui ne pourra pas être contournée, est le manque de données d'absence dans les bases de données d'observation. Les absences considérées dans les modèles utilisés lors de cette première ébauche sont des non-observations. Une possibilité d'amélioration du modèle peut être la simulation de pseudo-absences, comme testé par Engler et al. (2004), en couplant une analyse de facteur de niche écologique et un modèle linéaire généralisé avec des pseudo-absences simulées. Cette approche pourrait être particulièrement intéressante pour des espèces avec très peu de données comme les espèces rares.
Etant donné la large échelle géographique et la très grande distance phylogénétique entre les espèces considérées, des incertitudes entourent les prédictions de cette première ébauche. Afin de contrer ce problème, une approche bayésienne plutôt qu'inférentielle pourrait être à privilégier (Golding & Purse, 2016). Il serait intéressant de tester de telles approches pour modéliser la distribution de groupes avec peu d'espèces dans un premier temps, ou bien sur une échelle géographique plus fine (départementale par exemple). La littérature scientifique récente se tourne vers ces méthodes, désormais accessibles techniquement même pour de gros jeux de données, des données fragmentaires et des données de présence seulement (Fukaya et al., 2020; Redding et al., 2017; Villar-Hernández & Pérez-Elizalde, 2015).

Afin de progresser vers un fonctionnement de l'outil prédictif tel qu'imaginé lors de la conception de la base de connaissance des relations espèces – habitats, l'amélioration de la base de connaissance nécessite la réalisation de nombreux inventaires couplés espèces/habitats, pour lesquels il est possible de savoir quelles espèces ont été observées dans quels habitats. Cette information est relativement simple à renseigner pour la flore et la faune non volante. Ces inventaires pourraient notamment servir de référence pour des validations *in situ* des expertises des relations espèces – habitats réalisées, sur la partie « présence de l'espèce dans l'habitat ». Ainsi, des espèces inventoriées dans des habitats pour lesquels aucune relation n'est renseignée dans la base de connaissance pourra faire l'objet d'une demande d'expertise.
De la même façon, une fois que l'application sera mise à disposition des maîtres d'ouvrages et des bureaux d'études, les résultats des inventaires réalisés sur le terrain à partir de la liste d'espèces potentielles pourraient servir de tests de validité afin de compléter et d'améliorer la base de connaissance des relations espèces – habitats. De façon concrète, cela suppose la mise à disposition d'un formulaire de retour qui permette de renseigner, outre la liste des espèces potentiellement présentes et réellement observées, la liste des autres espèces observées, les coordonnées du site, la date d'inventaire et la météo, la méthode d'inventaire utilisée (liste non-exhaustive).

Idéalement, ce modèle aurait été construit à partir de la liste des habitats sur chaque site d'IPE/IQE et de la liste des espèces inventoriées sur chaque habitat pour valider le modèle prédictif. Il sera possible de mettre en œuvre un outil prédictif basé sur ces données dans quelques années avec l'aboutissement du projet de cartographie des habitats de métropole (CarHab).

L'amélioration de cette première ébauche permettra d'élargir les possibilités d'études, avec par exemple la mise en évidence des cortèges d'espèces retrouvées ensemble dans des habitats clés. Cela permettra d'étudier à grande échelle les réseaux trophiques, afin de mieux comprendre les interactions biotiques et abiotiques.

5 Bibliographie

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1648. https://doi.org/doi:10.18637/jss.v067.i01

Clair, M., Gaudillat, V., Michez, N., Poncet, R., & Poncet, L. (2017). *HABREF v4.0, référentiel des typologies d'habitats et de végétation pour la France Guide méthodologique* (p. 64). AFB/CNRS/MNHN.

Delzons, O., Cima, V., Fournier, C., Gourdain, P., Hérard, K., Lacoeuilhe, A., Laignel, J., Roquinarc'h, O., & Thierry, C. (2020). *Indice de Qualité Ecologique (IQE), Indice de Potentialité Ecologique (IPE) - Guide méthodologique – Version 2.0* (p. 118). UMS Patrimoine Naturel – Centre d'expertise et de données (OFB-CNRS-Muséum national d'Histoire naturelle).

Engler, R., Guisan, A., & Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, *41*(2), 263-274. https://doi.org/10.1111/j.0021-8901.2004.00881.x

Fukaya, K., Kusumoto, B., Shiono, T., Fujinuma, J., & Kubota, Y. (2020). Integrating multiple sources of ecological data to unveil macroscale species abundance. *Nature Communications*, *11*(1), 1695. https://doi.org/10.1038/s41467-020-15407-5

Gargominy, O., Tercerie, S., Régnier, C., Ramage, T., Dupont, P., Daszkiewicz, P., & Poncet, L. (2019). *TAXREF v13, référentiel taxonomique pour la France : Méthodologie, mise en oeuvre et diffusion.* (p. 63). Patrinat, Muséum national d'Histoire naturelle.

Golding, N., & Purse, B. V. (2016). Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology and Evolution*, *7*(5), 598-608. https://doi.org/10.1111/2041-210X.12523

Hortal, J., Lobo, J. M., & Jiménez-Valverde, A. (2012). Basic questions in biogeography and the (lack of) simplicity of species distributions : Putting species distribution models in the right place. *Natureza & Conservação*, *10*(2), 106-116. https://doi.org/10.4322/natcon.2012.029

Jeanmougin, M., Plattner, G., Porcher, E., Julliard, R., Touroult, J., & Poncet, L. (2014). *Synthèse bibliographique des changements d'échelles cartographiques et des relations écologiques entre les espèces et leurs habitats* (p. 83) [SPN-CESCO-MNHN]. MEDDE.

Legros, B., Ichter, J., Cellier, P., Houard, X., Louboutin, B., Poncet, L., Puissauve, R., & Touroult, J. (2016). *Caractérisation des relations Espèce-Habitat naturel et gestion de l'information* (Guide méthodologique N° 2016-01; p. 38). Service du patrimoine naturel, Muséum national d'Histoire naturelle.

Louvel, J., Gaudillat, V., & Poncet, L. (2013). *EUNIS, European Nature Information System, Système d'information européen sur la nature. Classification des habitats. Traduction française. Habitats terrestres et d'eau douce.* (p. 289). MNHN-DIREV-SPN, MEDDE.

QGIS. (2020). *QGIS Geographic Information System*. Open Source Geospatial Foundation Project. http://qgis.org

R Core Team. (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*.

Redding, D. W., Lucas, T. C. D., Blackburn, T. M., & Jones, K. E. (2017). Evaluating Bayesian spatial methods for modelling species distributions with clumped and restricted occurrence data. *PLOS ONE*, *12*(11), e0187602. https://doi.org/10.1371/journal.pone.0187602

Régnier, C., & Gargominy, O. (2018). *Diffusion des statuts des espèces : Principes et objectifs* (Patrinat N° 109; p. 8). UMS 2006 Patrimoine Naturel (AFB / CNRS / MNHN).

Villar-Hernández, B. de J., & Pérez-Elizalde, S. (2015). Bayesian models for species distribution modelling with only-presence records. *Tropical and Subtropical Agroecosystems*, *18*(2), Article 2. http://www.revista.ccba.uady.mx/ojs/index.php/TSA/article/view/1989

Wickham, H. (2016). *ggplot2 : Elegant Graphics for Data Analysis* (Springer-Verlag).

Witté, I., & Touroult, J. (2017). *Identification et cartographie des zones de méconnaissance naturaliste à l'échelle nationale (métropole) à partir des données partagées* (SPN N° 2017-6; p. 48). MNHN.

6 Annexes

Annexe 1 Résumés statistiques des différents modèles comparés pour choisir l'échelle et la résolution les plus pertinentes pour la variable spatiale liée aux données d'observations INPN des espèces. Les valeurs de paramètres sont moyennées par rapport aux résultats de chaque sous-échantillon d'entrainement du jeu de données (80 %). Pour chaque modèle, la variable testée figure au-dessus du tableau. Noms des zones bioclimatiques : ATC = Atlantic Central, CON = Continental, LUS = Lusitanean, MDM = Mediterranean Mountains, MDN = Mediterranean North, MDS = Mediterranean South.

Modèle 1 : Nombre centré-réduit d'observations INPN dans la maille focale

Paramètre	Coefficient	Erreur standard	Valeur-z	Valeur-p
(Intercept)	9,96	1,07	9,24	<0,001
Variable spatiale	0,51	0,04	13,18	<0,001
ATC	-0,38	0,09	-4,10	0,081
CON	-1,52	0,18	-8,32	<0,001
LUS	-0,31	0,09	-3,27	0,018
MDM	0,07	0,05	1,21	0,246
MDN	-1,16	0,07	-16,09	<0,001
MDS	-0,16	0,18	-0,91	0,409
Latitude	-3,50	0,17	-20,81	<0,001
Longitude	-0,43	0,12	-3,59	0,011

Modèle 2 : Nombre centré-réduit d'observations INPN dans les mailles adjacentes à la maille focale

Paramètre	Coefficient	Erreur standard	Valeur-z	Valeur-p
(Intercept)	9,89	1,05	9,39	<0,001
Variable spatiale	0,63	0,04	15,41	<0,001
ATC	-0,36	0,09	-3,94	<0,001
CON	-1,51	0,18	-8,23	<0,001
LUS	-0,28	0,09	-3,02	0.006
MDM	0,08	0,05	1,56	0.127
MDN	-1,10	0,07	-15,24	<0,001
MDS	-0,13	0,18	-0,77	0.467
Latitude	-3,48	0,17	-20,64	<0,001
Longitude	-0,51	0,12	-4,27	<0,001

Modèle 3 : Nombre centré-réduit d'observations INPN dans les mailles dans un rayon de 50 km autour du centroïde de la maille focale

Paramètre	Coefficient	Erreur standard	Valeur-z	Valeur-p
(Intercept)	10,41	1,07	9,67	<0,001
Variable spatiale	0,79	0,05	16,16	<0,001
ATC	-0,35	0,09	-3,78	0.070
CON	-1,36	0,18	-7,53	<0,001
LUS	-0,31	0,09	-3,30	0.015
MDM	0,05	0,05	0,88	0.400
MDN	-1,14	0,07	-15,85	<0,001
MDS	-0,14	0,18	-0,80	0.448
Latitude	-3,51	0,17	-20,92	<0,001
Longitude	-0,67	0,12	-5,52	<0,001

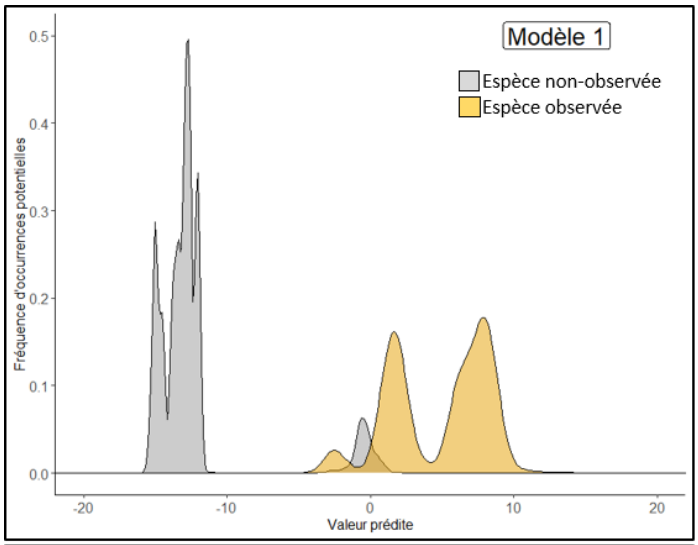
Modèle 4 : Nombre centré-réduit de mailles ayant au moins une donnée d'observations INPN dans un rayon de 50 km autour du centroïde de la maille focale

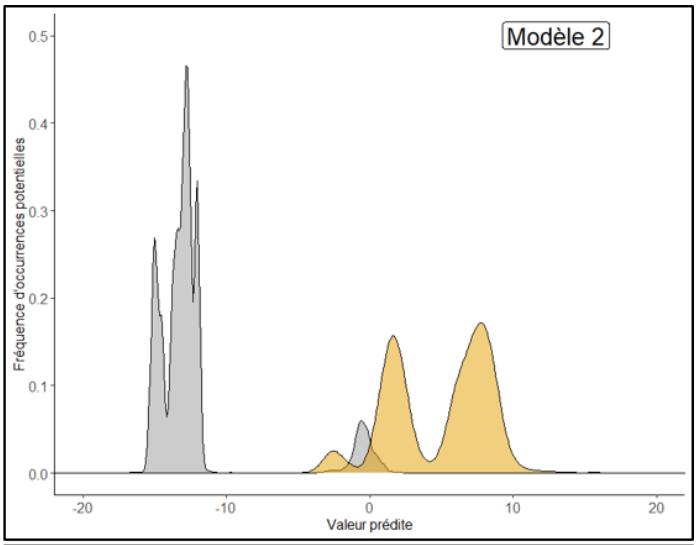
Paramètre	Coefficient	Erreur standard	Valeur-z	Valeur-p
(Intercept)	8,92	1,05	8,48	<0,001
Variable spatiale	0,68	0,03	24,24	<0,001
ATC	-0,55	0,09	-5,96	<0,001
CON	-1,50	0,18	-8,22	<0,001
LUS	-0,38	0,09	-4,09	<0,001
MDM	-0,07	0,06	-1,23	0.242
MDN	-1,21	0,07	-16,48	<0,001
MDS	-0,05	0,18	-0,31	0.719
Latitude	-3,20	0,17	-19,04	<0,001
Longitude	-1,01	0,12	-8,14	<0,001

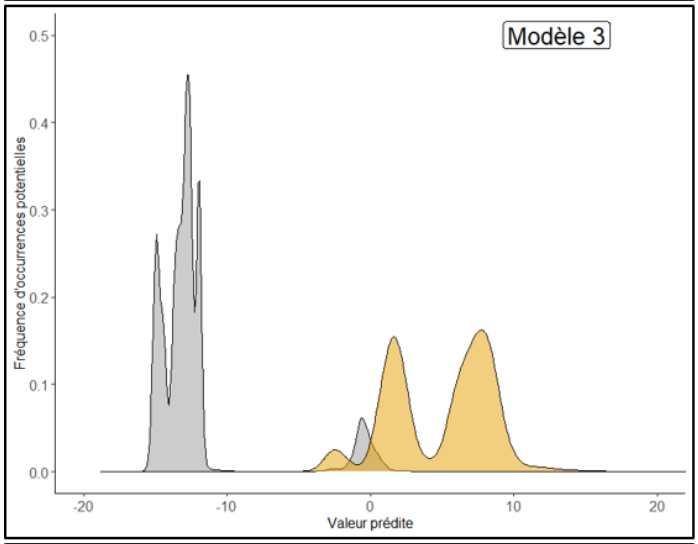
Modèle 5 : Proportion de mailles ayant au moins une donnée d'observations INPN dans un rayon de 50 km autour du centroïde de la maille focale

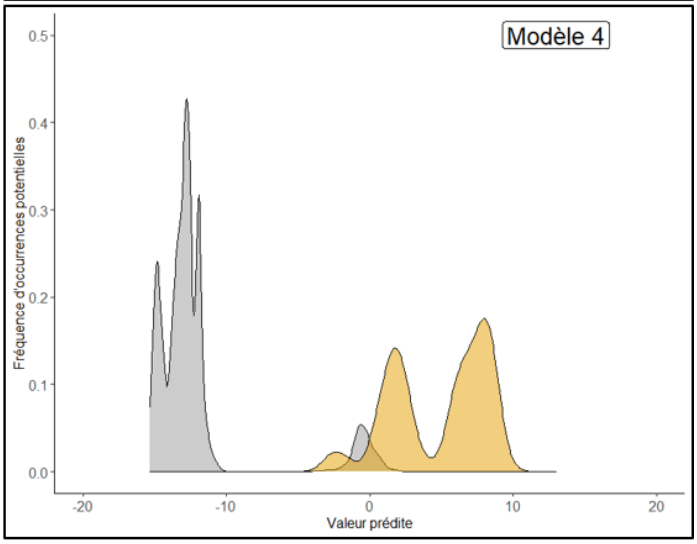
Paramètre	Coefficient	Erreur standard	Valeur-z	Valeur-p
(Intercept)	7,94	1,05	7,54	<0,001
Variable spatiale	2,25	0,09	24,18	<0,001
ATC	-0,50	0,09	-5,43	<0,001
CON	-1,70	0,19	-9,15	<0,001
LUS	-0,30	0,09	-3,24	0.003
MDM	0,01	0,06	0,22	0.725
MDN	-1,11	0,07	-15,20	<0,001
MDS	0,04	0,18	0,23	0.795
Latitude	-3,12	0,17	-18,53	<0,001
Longitude	-1,01	0,12	-8,12	<0,001

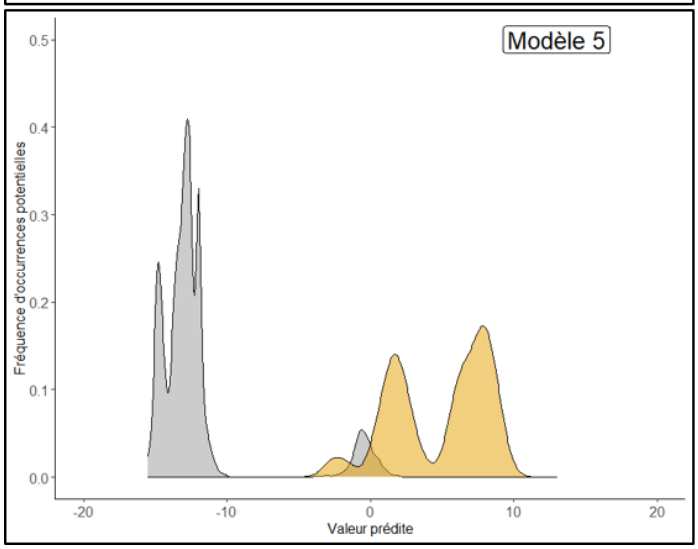
Annexe 2 Comparaison des modèles prédictifs sur différentes échelles spatiales et résolutions des données d'observations INPN. Les graphiques sont réalisés à partir des prédictions calculées sur un des sous-échantillons de test (20 % du jeu de données). Chaque graphique montre les fréquences des distributions des occurrences potentielles en fonction de la valeur prédite, et selon l'observation ou non de l'espèce sur le terrain, pour chaque modèle comparant les variables spatiales relatives aux données d'observations INPN.











Variable testée par chaque modèle (tous pas de temps confondus), par espèce potentiellement présente :
Modèle 1 : Nombre centré-réduit d'observations INPN dans la maille focale;
Modèle 2 : Nombre centré-réduit d'observations INPN dans les mailles adjacentes à la maille focale;
Modèle 3 : Nombre centré-réduit d'observations INPN dans les mailles dans un rayon de 50 km autour de la maille focale;
Modèle 4 : Nombre centré-réduit de mailles avec au moins une observation INPN dans un rayon de 50 km autour de la maille focale;
Modèle 5 : Proportion de mailles avec au moins une observation INPN dans un rayon de 50 km autour de la maille focale;

Annexe 3 Résumés statistiques des différents modèles comparés pour choisir l'échelle temporelle la plus pertinente pour la proportion de mailles dans un rayon de 50 km autour du centroïde de la maille focale ayant des données d'observations INPN. Les valeurs de paramètres sont calculées sur le sous-échantillon d'entrainement du jeu de données (80 %). Noms des zones bioclimatiques : ATC = Atlantic Central, CON = Continental, LUS = Lusitanean, MDM = Mediterranean Mountains, MDN = Mediterranean North, MDS = Mediterranean South.

Modèle 8 : Données d'observations INPN datant de moins de 10 ans (période 2009-2019)

Paramètre	Coefficient	Erreur standard	Valeur-z	Valeur-p
(Intercept)	8,12	1,04	7,79	<0,001
Variable spatiale	2,80	0,12	23,29	<0,001
ATC	-0,49	0,09	-5,28	<0,001
CON	-1,78	0,19	-9,54	<0,001
LUS	-0,37	0,09	-3,92	<0,001
MDM	0,02	0,06	0,40	0.704
MDN	-1,11	0,07	-15,22	<0,001
MDS	-0,01	0,18	-0,05	0.863
Latitude	-3,14	0,17	-18,81	<0,001
Longitude	-1,00	0,12	-8,11	<0,001

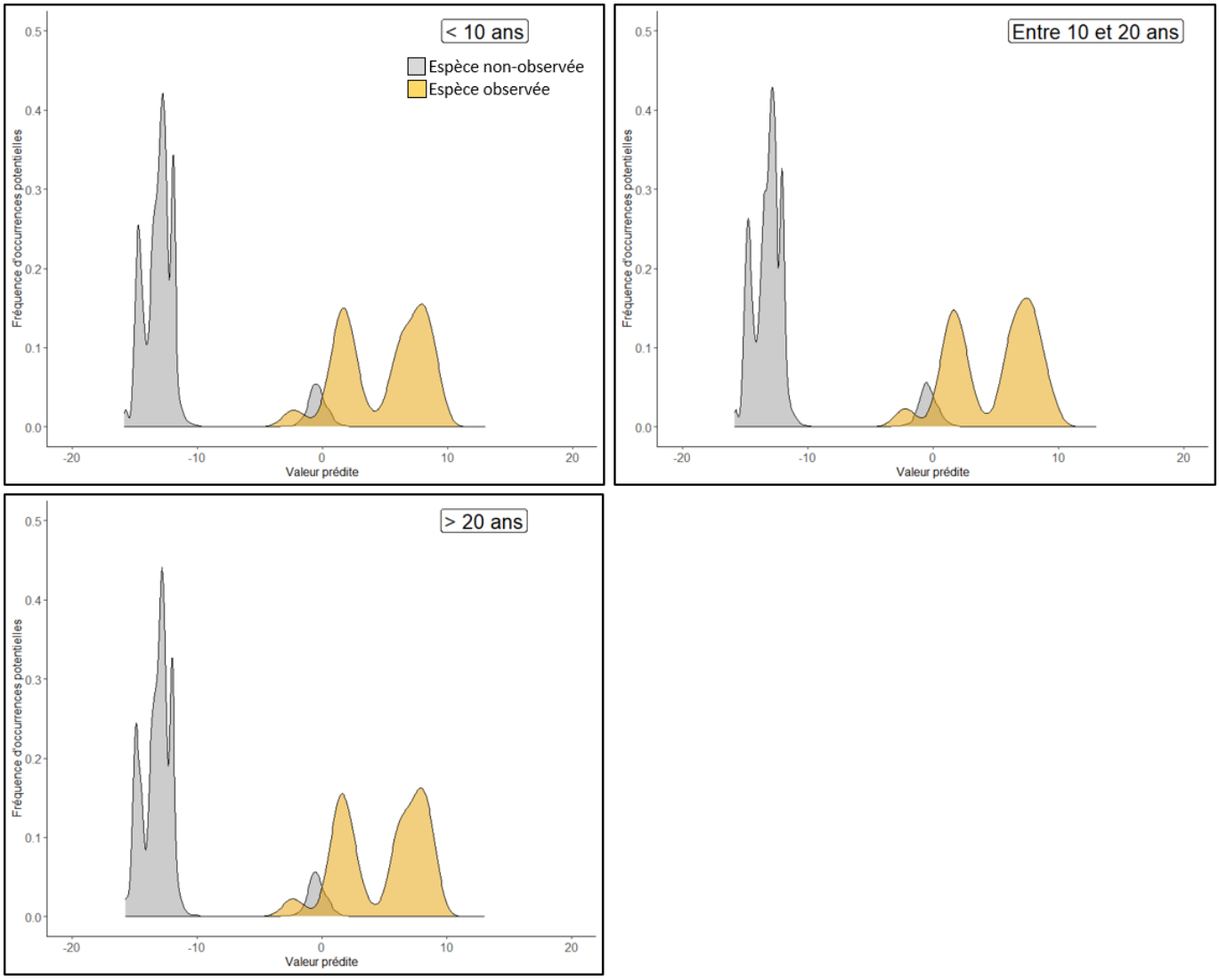
Modèle 9 : Données d'observations INPN datant d'entre 10 et 20 ans (période 1999-2009)

Paramètre	Coefficient	Erreur standard	Valeur-z	Valeur-p
(Intercept)	7,79	1,07	7,28	<0,001
Variable spatiale	2,73	0,13	21,78	<0,001
ATC	-0,48	0,09	-5,29	<0,001
CON	-1,75	0,19	-9,38	<0,001
LUS	-0,28	0,09	-2,99	0.006
MDM	0,04	0,05	0,69	0.518
MDN	-1,14	0,07	-15,69	<0,001
MDS	0,00	0,18	0,01	0.859
Latitude	-3,11	0,17	-18,54	<0,001
Longitude	-0,95	0,12	-7,66	<0,001

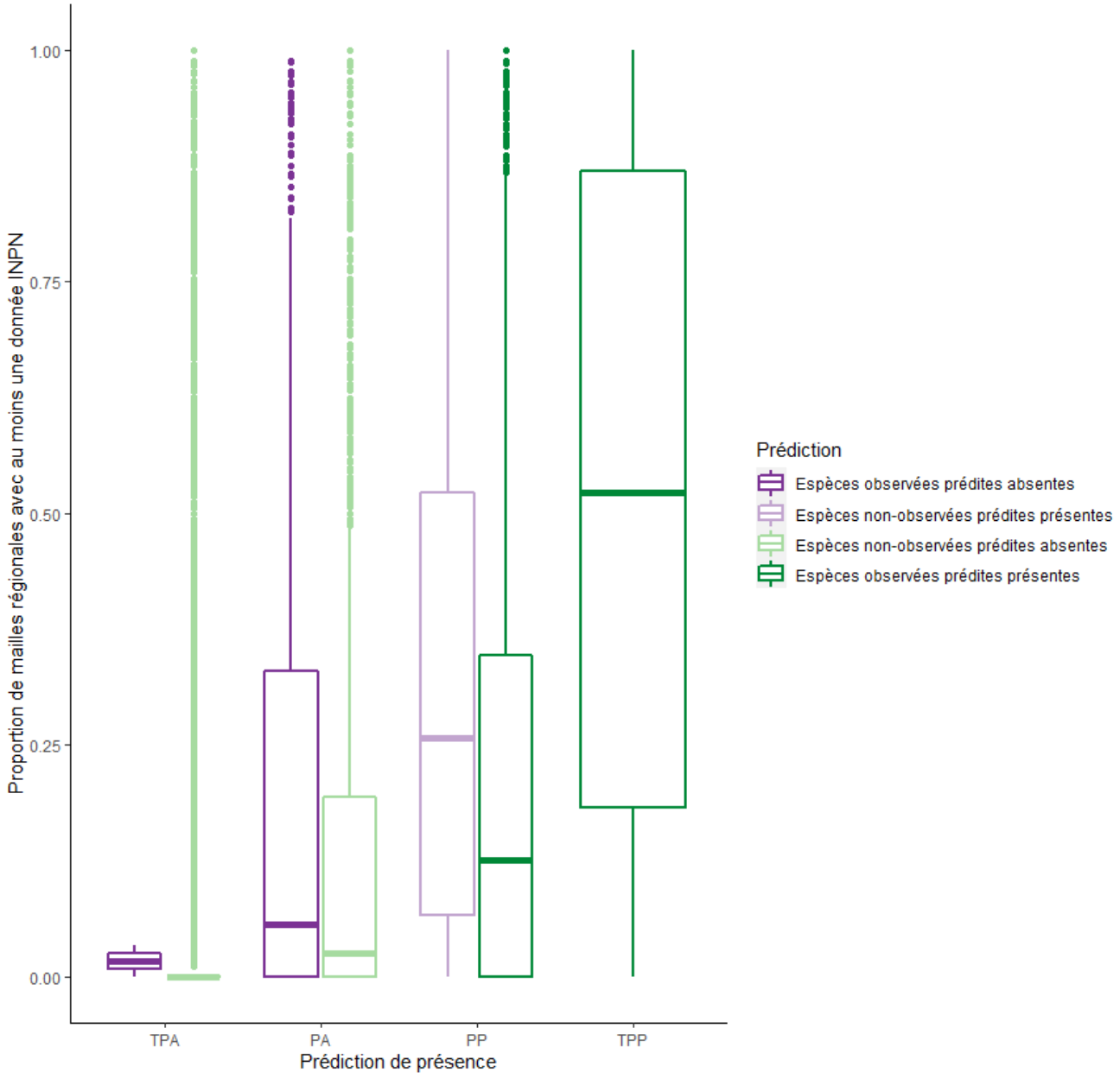
Modèle 10 : Données d'observations INPN antérieures à 20 ans (avant 1999)

Paramètre	Coefficient	Erreur standard	Valeur-z	Valeur-p
(Intercept)	8,94	1,04	8,58	<0,001
Variable spatiale	2,42	0,12	20,56	<0,001
ATC	-0,46	0,09	-5,01	<0,001
CON	-1,64	0,18	-8,95	<0,001
LUS	-0,29	0,09	-3,05	0.004
MDM	0,03	0,05	0,59	0.572
MDN	-1,11	0,07	-15,38	<0,001
MDS	0,01	0,18	0,02	0.872
Latitude	-3,31	0,17	-19,81	<0,001
Longitude	-0,78	0,12	-6,36	<0,001

Annexe 4 Comparaison des modèles prédictifs sur différents pas de temps. Les graphiques sont réalisés à partir des prédictions calculées sur un des sous-échantillons de test (20 % du jeu de données). Chaque graphique montre les fréquences des distributions des occurrences potentielles en fonction de la valeur prédite, et selon l'observation ou non de l'espèce sur le terrain, pour chaque modèle comparant l'échelle temporelle des données d'observations INPN.



Annexe 5 Proportions de mailles régionales avec au moins une donnée d'observation INPN en fonction de la catégorie de probabilité de présence à laquelle sont attribuées les occurrences potentielles. Une distinction est faite selon la bonne (en vert) ou la mauvaise (en violet) attribution des occurrences.



RÉSUMÉ

Ce rapport présente les analyses réalisées pour déterminer la potentialité de présence des espèces à l'échelle de la France métropolitaine, dans le cadre d'une convention entre l'UMS PatriNat et la Direction des infrastructures de transport du Ministère de la transition écologique.

A partir d'inventaires d'habitats réalisés sur divers sites métropolitains et de la base de connaissance des relations établies entre les espèces et les habitats qu'elles occupent, une liste des espèces potentiellement présentes est construite pour chaque site. Un jeu de données rassemblant 2 287 espèces (dont 84 % sont patrimoniales), alimenté d'observations de terrain dans 36 départements a ainsi été construit. Les espèces incluses appartiennent à sept groupes taxonomiques (amphibiens, angiospermes, odonates, oiseaux nicheurs, orthoptères, reptiles et rhopalocères). Une régression logistique permet de tester la qualité de la prédiction de différents critères généraux (données d'observations dans l'Inventaire national du patrimoine naturel, zone bioclimatique, latitude et longitude du site, effet de l'espèce).

Le modèle sélectionné est volontairement simple et permet de prédire correctement la présence ou l'absence des espèces pour 97 % du jeu de données. Certains groupes taxonomiques ont des espèces moins bien prédites, comme les orthoptères ou les angiospermes. Le critère de patrimonialité de l'espèce n'influe pas sur la qualité de la prédiction de présence.

Le modèle prédictif a ensuite été testé sur les données d'inventaires réalisés dans le cadre du projet du contournement ouest de Nîmes (30). Sur les 394 espèces observées lors de ces inventaires, 212 sont présentes dans la base de connaissance des relations espèces – habitats. La présence ou l'absence des espèces potentielles est correctement prédite par le modèle à hauteur de 76 %. D'autres tests avec des données réelles devront être réalisés afin de de perfectionner les prédictions de présence.

Les analyses statistiques permettent de donner une première version de l'outil prédictif de la présence d'espèce simple et relativement efficace. Cet outil sera utile en amont des inventaires d'études d'impact, mais également pour étudier les effets possibles de modifications paysagères sur les cortèges présents.