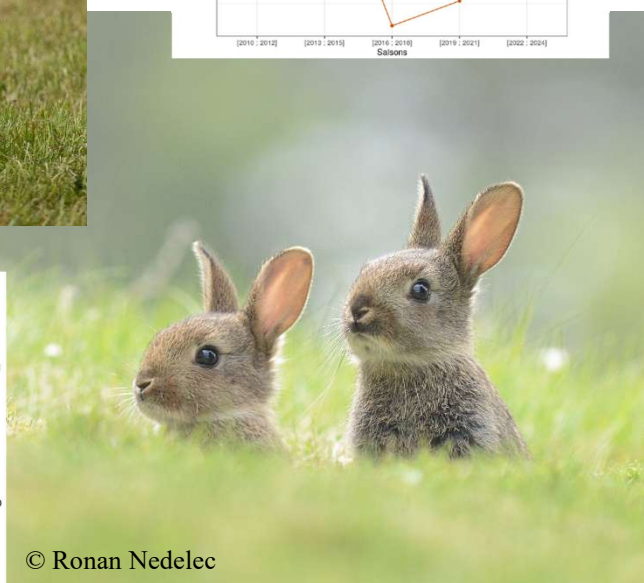
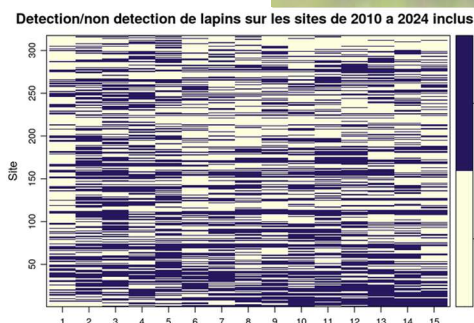


Analyse de données opportunistes pour rechercher une potentielle tendance d'évolution temporelle de populations de mammifères en Bretagne



Septembre 2025

COFINANCÉ PAR
UNION EUROPÉENNE



L'Europe s'engage
en Bretagne





Groupe Mammalogique Breton -www.gmb.bzh

Maison de la Rivière - 29450 Sizun

tél. : 02 98 24 14 00

Courriel : contact@gmb.bzh

Analyse de données opportunistes pour rechercher une potentielle tendance d'évolution temporelle de populations de mammifères en Bretagne

Valentine Cléac'h (ENSAI), Franck Simonnet (GMB)

Septembre 2025

Le GMB remercie toutes les personnes ayant contribué à la collecte des données.

Résumé

Cette étude explore la possibilité de détecter des évolutions temporelles dans les populations de mammifères en Bretagne à partir de données opportunistes de détection. Les données opportunistes sont des observations d'espèces collectées sans protocole scientifique prédéfini et de façon volontaire par des observateurs. L'analyse couvre une période de quinze ans et vise à évaluer dans quelle mesure ces données peuvent être exploitées pour le suivi de la faune.

Dans un premier temps, différentes régressions linéaires des proportions d'observations ont été testées. Ces modèles se sont révélés peu adaptés aux données opportunistes, en raison de la présence de nombreux biais. Nous avons donc eu recours à des modèles dits « *d'occupancy* », qui permettent de distinguer explicitement le processus écologique (présence réelle) du processus d'observation (probabilité de détection).

L'approche a permis d'obtenir un résultat robuste pour une seule espèce, la fouine. L'utilisation d'un modèle *d'occupancy* dynamique à trois périodes met en évidence une tendance à la hausse de sa probabilité de présence sur le territoire breton.

Les résultats de ce travail exploratoire illustrent à la fois le potentiel et les limites de l'utilisation de données opportunistes pour le suivi des populations de mammifères, et ouvrent des perspectives méthodologiques pour mieux exploiter ces sources d'information.

Table des matières

1	Introduction	1
2	Environnement du stage	2
3	Des données opportunistes	3
3.1	Une difficulté due à la nature des données	3
3.2	Nettoyage des données naturalistes	4
3.3	Les mailles définies par des variables géographiques.....	5
3.4	Les bases de données naturalistes présentent des différences.....	6
4	De meilleurs résultats avec les modèles <i>d'occupancy</i>	8
4.1	Les régressions linéaires sont inefficaces	8
4.1.1	Modélisation de proportions entre espèces par un modèle simple.....	8
4.1.2	Des modèles non valides	10
4.2	Le modèle <i>d'occupancy</i> : Un modèle séparant présence et détection.....	10
4.2.1	Différentes périodes pour différentes espèces.....	12
4.2.2	Une augmentation de la population des fouines	13
	Conclusion et perspectives	15
	Bibliographie	16
	Annexes	17

1 Introduction

*Cette étude a été menée dans le cadre d'un **stage** de 2e année du cursus « Ingénieur en Sciences de la donnée (spécialité biostatistiques) » de l'Ecole Nationale de la Statistique et de l'Analyse de l'Information. Le présent document est une version amendée du rapport de stage : quelques éléments issus de discussions avec les trois scientifiques référents (voir page suivante) à l'issue d'une restitution orale des résultats en fin de stage ont été ajoutés par le GMB au chapitre « Conclusion et Perspectives ». L'ensemble des analyses menées est consultable à l'adresse : https://rpubs.com/valentine_cleach/.*

Les données opportunistes correspondent à des données de détection renseignées de manière volontaire par des observateurs. Avec l'accroissement d'Internet au cours des dernières décennies, ce type de données est devenu de plus en plus répandu, notamment via des bases de données participatives, comme celles étudiées dans ce mémoire, mais également à travers les observations diffusées sur les réseaux sociaux. Ces données présentent un intérêt majeur lorsqu'elles sont correctement exploitées, car elles permettent de constituer de larges jeux de données à un coût relativement faible. Contrairement aux protocoles standardisés nécessitant un suivi sur plusieurs années et mobilisant des moyens importants pour de nombreuses espèces, les données opportunistes offrent une alternative peu coûteuse et facile à collecter.

Au sein du Groupe Mammalogique Breton, certaines espèces telles que la loutre d'Europe ou les chauves-souris font l'objet de suivis spécifiques à l'aide de protocoles standardisés. Cependant, la majorité des mammifères présents en Bretagne ne disposent pas de protocoles dédiés. Dans ce contexte, les données opportunistes constituent une ressource précieuse pour l'organisme naturaliste, car elles représentent souvent la seule source d'information disponible pour l'étude et le suivi de ces espèces.

Dans ce stage, nous chercherons à répondre à la question suivante : *dans quelle mesure est-il possible d'extraire des tendances temporelles sur différentes populations de mammifères à partir de données opportunistes ?*

2 Environnement du stage

Le Groupe Mammalogique Breton, ou GMB, est une association de protection de la nature qui couvre la Bretagne historique, les 4 départements de la Bretagne et la Loire-Atlantique. L'association possède 3 antennes, une à Sizun où s'est effectué le stage, une à Redon, et la dernière à Ploufragan. Elle s'intéresse aux mammifères sauvages présents sur le territoire, notamment la loutre, les chauves-souris et les micromammifères.

Le GMB se structure autour des 4 axes suivants :

- La sensibilisation, à travers des documents pédagogiques comme des plaquettes, livrets, vidéos ou encore des animations,
- La connaissance. Le GMB recense différentes espèces à l'aide d'inventaires et de suivis d'espèces.
- La protection. Le GMB incite les collectivités et les particuliers à se questionner sur la cohabitation avec l'homme et les aide dans leurs démarches. Il aide également en donnant des conseils d'aménagement du territoire afin de conserver ou rétablir des espaces où certains mammifères peuvent vivre et circuler.
- La formation. Le GMB forme à la protection et à l'étude des mammifères.

Mon maître de stage, Franck Simonnet, est Chargé de mission « Etudes et conservation » au sein du GMB. Il est spécialisé dans l'étude des loutres, et suit aussi d'autres espèces telles que les loups et les micromammifères. Nous avons pu bénéficier du concours de trois personnes ressources : Eric Petit chercheur à l'INRAE, Pascal Irz de l'OFB et Simon Lacombe, doctorant au Centre d'Ecologie Fonctionnelle et Evolutive du CNRS. Qu'ils soient vivement remerciés pour leur temps et leurs conseils.

Mon stage s'insère dans le programme de *l'Observatoire des Mammifères de Bretagne*, qui organise et pérennise le suivi des mammifères en Bretagne. Il a pour objectif d'étudier les données opportunistes recueillies par le GMB ainsi que celles de faune sauvage afin de voir si certaines populations de mammifères suivent des tendances temporelles. A terme, cette étude pourra contribuer à la Liste Rouge régionale des espèces menacées.

3 Des données opportunistes

Une difficulté due à la nature des données

Ma mission lors de mon stage au Groupe Mammalogique Breton est d'étudier une possible tendance temporelle d'évolution des populations de certains mammifères en Bretagne. Mon travail a pour rôle de confirmer ou non ce que le GMB observe sur le terrain en utilisant des méthodes statistiques.

Pour cela, on dispose de 3 bases de données différentes :

- GeoNature, une base de données recueillie par le GMB,
- VisioNature, une base de données participative en ligne sur le [portail Faune Bretagne](#)
- Collision faune sur route nationale, une base de données de la Direction Interdépartementale des Routes de l'Ouest, DIR Ouest.

—

Ces données sont particulières car elles sont opportunistes. Elles sont recueillies par des naturalistes sans suivre de protocole. Lorsqu'ils observent un animal ou des indices de présence ou entendent des bruits permettant d'affirmer la présence d'une espèce dans une partie de la Bretagne, ils peuvent insérer cette observation dans la base de données. Ces derniers précisent certaines informations comme la date, la localisation, le nombre d'individus et leur état de santé. Ces observations sont ensuite vérifiées par les propriétaires de la base de données. Les données sont donc assez fiables.

Néanmoins, la nature même de ces données peut entraîner plusieurs biais. Ces biais peuvent être regroupés en quatre grandes catégories [1] :

- **Biais d'observation** : variations de l'effort de prospection au cours du temps,
- **Biais de déclaration** : enregistrements incomplets et/ou sélectifs de la part des observateurs,
- **Biais géographiques** : distribution inégale des sites étudiés,
- **Biais de détection** : biais communs à l'ensemble des données de détection, qu'elles soient opportunistes ou issues de protocoles standardisés.

La base de données GeoNature correspond en grande partie à la période 2010 à 2014. A partir de 2015, le portail Faune Bretagne est mise en ligne. On retrouve cette différence dans les statistiques descriptives 2.

La base de données DIR Ouest a été créée différemment. Des informations concernant des collisions d'animaux sont récupérées, puis regroupées et saisies mensuellement dans un tableau. Il n'y a pas de vérification des données saisies et les employés de la DIR Ouest risquent potentiellement de confondre différentes espèces, notamment les mustélidés, car ils sont difficilement différenciables les uns des autres. Cette base de données a été mise en place en 2014.

Après un nettoyage de nos bases de données, on recense 55349 observations pour la base de données GeoNature, 68214 pour VisioNature, et enfin 24744 observations pour les données de la DIR Ouest.

L'une des difficultés de la mission est de trouver un modèle qui fonctionne sur toutes les espèces étudiées. En effet, le groupe des mammifères regroupe des espèces très différentes selon le mode de vie, l'espérance de vie et d'autres paramètres. A défaut de trouver ce modèle hypothétique, nous avons créé des modèles en sélectionnant les variables pertinentes selon le critère BIC (critère d'information bayésien). Au cours du stage, les espèces qui ont été analysées sont les suivantes :

- Le sanglier,
- Le renard,
- Le lapin de garenne,
- Le hérisson d'Europe,
- Le putois d'Europe,
- L'hermine,
- La belette,
- La fouine.

Le sanglier et le renard sont étudiés ici comme des espèces de référence car le niveau de connaissance de ces deux espèces est bon. La progression de leur population respective est connue en amont de l'analyse, ce qui permet de comparer les résultats du modèle choisi à ce qui aurait pu être observé selon un protocole. Concernant les autres espèces, une évolution à la hausse ou à la baisse est supposée ou pressentie.

Nettoyage des données naturalistes

Tout d'abord, afin de garantir la fiabilité des résultats, les données classées comme « Douteuses » ou « Invalides » ont été retirées. Nous avons ensuite supprimé certaines espèces qui nous intéressaient moins, comme les taupes, ainsi que les espèces ne provenant pas de Bretagne. De plus, les données manquant de précision sur l'espèce observée ont été écartées. Par exemple, les observations qui n'identifiaient pas une espèce précise mais un regroupement, tel que « petits mustélidés » ou « rongeurs », n'ont pas été conservées. Les techniques d'observation associées aux hérissons ont toutes été uniformisées en « Vu » quelle que soit la technique réellement utilisée (« Entendu », « Fèces/Guano/Epreintes », « Empreintes », etc).

Différentes techniques d'observation d'indices de présence ont ensuite été regroupées. Les enregistrements avec une technique d'observation « Inconnue » étaient nombreux (73382 observations parmi les 2 bases de données). L'examen des commentaires de ces observations qui contenaient les mots « taupinières », « empreintes », « traces » ou leurs variantes a permis

de reclasser ces données parmi les indices de présence. Il ne restait alors qu'un faible nombre de données présentant à la fois un commentaire non informatif et une technique d'observation non renseignée (0,6 % des observations avec une technique d'observation non renseignée). Pour le reste des techniques « Inconnues », il aurait été envisageable d'imputer les valeurs manquantes par des méthodes statistiques (par exemple, les k plus proches voisins ou une décomposition en valeurs singulières). Toutefois, après analyse d'histogrammes comparant les techniques d'observation selon les espèces, il a été décidé de les remplacer par « Vu ».

Un filtre a ensuite été appliqué sur les commentaires contenant la chaîne de caractères « rout », permettant d'identifier ceux mentionnant « routes » ou « routière ». Ces cas ont été interprétés comme des individus trouvés morts suite à un impact routier. Cette méthode reste néanmoins approximative, certains commentaires faisant référence à des animaux vivants observés sur ou depuis une route. Cependant, ce filtre n'ayant modifié qu'un faible nombre d'observations (92), ces exceptions n'ont pas été jugées significatives.

Concernant les observateurs, plusieurs étaient anonymes (attribués à un code individuel) et d'autres non renseignés (2063 observations, majoritairement issues de la base VisioNature). Afin de ne pas regrouper artificiellement des individus différents, un code distinct a été attribué à chaque observation dont l'observateur n'était pas connu.

Enfin, certaines observations étaient présentes à la fois dans les deux bases de données naturalistes. Celles-ci, signalées par un numéro d'identifiant spécifique, ont été supprimées afin d'éviter les doublons.

Par ailleurs, étant donné que la base VisioNature n'intègre pas la Loire-Atlantique, les données concernant ce département ont également été retirées de la base GeoNature. Les îles ont été supprimées car l'étude portait principalement sur le continent, dont les dynamiques diffèrent de celles des îles. Toutes les observations localisées à moins de 200 mètres du large ont été conservées.

Enfin, les bases de données GeoNature et VisioNature ont été fusionnées, après harmonisation de certaines différences de présentation.

Vous pouvez retrouver la liste des variables sélectionnées en Annexe 1

Les mailles définies par des variables géographiques

Nous avons importé une grille de 10 × 10 km, que nous avons superposée à la Bretagne continentale, afin de découper la Bretagne et observer plus finement l'évolution de la population de chaque espèce et potentiellement observer des différences entre les zones de la Bretagne.

Ensuite, plusieurs variables géographiques susceptibles d'expliquer certains phénomènes ont été ajoutées. Ces variables ont été traitées différemment selon leur nature.

Nous avons utilisé des images matricielles contenant les variables suivantes :

- Un indice de diversité des milieux dans un rayon de 500 m ;
- La densité de cultures dans un rayon de 500 m ;
- La distance à l'écotone arboré (haies + lisières) le plus proche ;
- La distance à un cours d'eau ou un plan d'eau ;
- La distance au littoral le plus proche.

Au sein de chaque maille de 10×10 km, la valeur moyenne de chaque variable a été conservée, indépendamment des observations de mammifères.

Concernant les différentes familles de paysages, le processus a été légèrement différent. Pour les mailles situées à l'intersection de plusieurs familles, la famille dominante (en termes de surface couverte) a été attribuée. Certaines mailles le long du littoral ne possédaient pas de famille de paysage attribuée. Dans ce cas, la famille la plus fréquente parmi les 8 mailles voisines a été retenue.

L'aire totale de chaque maille a ensuite été calculée. Bien que la plupart aient une aire de 100 km^2 , les mailles situées le long du littoral ou en limite régionale étaient parfois de taille inférieure. Environ 7% d'entre elles avaient une surface inférieure à 10 km^2 (22 sites sur un total de 337). Réparties aléatoirement en bordure de la Bretagne, ces mailles ont été supprimées car jugées trop petites.

Enfin, les coordonnées du centre de chaque maille ont été calculées afin d'obtenir la longitude et la latitude.

À chaque observation de notre base de données originale, nous avons attribué une maille ainsi que les différentes variables correspondantes.

Les bases de données naturalistes présentent des différences

Les bases de données naturalistes présentent des différences notables. Tout d'abord, la répartition géographique des données n'est pas la même. La base de données VisioNature représente davantage le littoral sud, tandis que GeoNature privilégie les données situées dans le nord-ouest du territoire 3. Étant donné que nous travaillons sur des données opportunistes, nous pouvons supposer que cette disparité n'est pas due à une densité de population plus élevée des mammifères, mais plutôt à une différence de pression d'observation entre les deux bases.

De plus, comme nous avons pu voir auparavant, les données de ces deux bases de données ne sont pas réparties de la même façon dans le temps 2. Bien que nous ayons des données datant d'il y a fort longtemps, il y a une nette augmentation du nombre des données à partir de 2010 du fait de la mobilisation de nombreux naturalistes bénévoles pour la réalisation d'un

Atlas des Mammifères régional. Pour cette raison, nous avons choisi d'écarter les données antérieures à 2010. En effet, si nous les avions conservées, les résultats auraient manqué de robustesse : une seule détection aurait pu modifier sensiblement les résultats, du fait du faible nombre d'observations. De plus, l'année 2025 n'a pas été utilisée car elle était incomplète et donc non représentative.

Les espèces sont également réparties différemment (4). On remarque que la base GeoNature contient plus d'observations de carnivores que de cétartiodactyles, tandis que la tendance est inverse dans VisioNature. Le GMB s'intéresse particulièrement à la loutre. Cependant, même en supprimant les observations relatives à cette espèce, nous retrouvons des distributions très similaires.

Enfin, un travail spécifique a été réalisé sur les observateurs. Nous avons remarqué que certains individus enregistrent un nombre particulièrement important d'observations. En particulier, un observateur a consigné plus de 5500 observations réparties entre les deux bases de données. Nous avons également constaté que cet observateur, ainsi que certains autres, concentrent leurs observations sur des zones précises. Cela n'est pas problématique pour notre étude, car, dans le cadre de modèles d'occupancy (présence-absence) appliqués à chaque maille de 10 km², une maille avec un grand nombre d'observations n'impacte pas les résultats globaux. En revanche, les données de deux personnes dans ce cas de figure ont été retirées pour les régressions linéaires. Nous notons par ailleurs que, dans GeoNature, la majorité des observateurs saisissent entre 2 et 10 observations, tandis que VisioNature est principalement alimentée par des observateurs n'ayant renseigné qu'une seule observation 5. Cette particularité peut s'expliquer par le grand nombre d'observateurs non identifiés que nous avons modifié en individus distincts. Enfin, une classification non supervisée a permis de regrouper nos observateurs en trois catégories, selon la diversité des espèces renseignées, le nombre total d'observations et les ordres taxonomiques privilégiés même si nous ne l'avons pas utilisée par la suite.

4. De meilleurs résultats avec les modèles *d'occupancy*

Pour rappel, nous souhaitons savoir s'il existe des tendances d'évolution des populations de mammifères en Bretagne.

Les régressions linéaires sont inefficaces

Modélisation de proportions entre espèces par un modèle simple

Tout d'abord, nous utilisons des régressions linéaires afin de déterminer si une tendance linéaire est apparente. Pour ce faire, nous calculons la proportion entre deux espèces pour chaque année entre 2010 et 2024 pour évaluer l'évolution des populations :

- une espèce témoin dont la population varie peu, qui est facilement détectable, présente sur chaque maille, pour chaque année et bien répartie sur l'ensemble du territoire ;
- l'espèce d'intérêt.

La proportion de l'espèce d'intérêt dans une maille constitue la variable expliquée de nos modèles de régressions. Les variables explicatives utilisées sont les variables géographiques présentées auparavant, mais également la proportion aux deux dates précédentes `prop_tmoins1`, et `prop_tmoins2`.

Différentes espèces témoins ont été retenues afin de comparer les modèles. Les espèces choisies sont celles dont les populations sont les plus communes et largement réparties sur le territoire, et soit les plus stables dans le temps soit celles pour lesquelles la tendance d'évolution est relativement bien identifiée :

- le sanglier (en augmentation);
- le chevreuil (stable) ;
- le renard (en augmentation après une baisse) ;
- le blaireau (stable).

Le critère de sélection des variables explicatives est le BIC. Nous avons utilisé le package `leaps`, qui permet de sélectionner efficacement les variables par une procédure de remplacement séquentiel. La variable *annee* pouvait être considérée soit comme quantitative, soit comme qualitative. Le choix s'est porté automatiquement sur la formulation correspondant au BIC minimal. Dans le cas où *annee* était quantitative, nous avons créé une variable *annee2*, terme quadratique permettant de prendre en compte une éventuelle non-linéarité (par exemple, des tendances de type concave ou convexe, telles qu'une croissance suivie d'un déclin).

Nous avons ainsi étudié non pas directement la variation de présence dans le temps, mais la variation relative de détection par rapport à une autre espèce. En effet, les données disponibles concernaient la détection et non la présence avérée. Il était donc nécessaire de

prendre en compte cette distinction. C'est pourquoi nous avons choisi de modéliser les proportions, plutôt que le nombre brut d'observations. Enfin, une seule espèce témoin a été utilisée à la fois, et non l'ensemble des espèces disponibles dans la base, afin de limiter la complexité et de conserver une comparaison claire. [10]

Définition d'un modèle de régression linéaire

La régression linéaire est une méthode statistique qui consiste à ajuster une droite (ou un hyperplan) à un nuage de points, de façon à minimiser l'erreur entre les valeurs observées et les valeurs prédites par le modèle. Cette méthode repose sur le principe des moindres carrés.

Mathématiquement, à partir des variables explicatives $X^{(1)}, \dots, X^{(n)}$ et d'une variable expliquée Y , le modèle peut s'écrire ainsi :

$$Y_i = \beta_0 + \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \dots + \beta_n X_i^{(n)} + \varepsilon_i,$$

où l'on cherche à estimer les coefficients $\beta_0, \dots, \beta_n$ pour chaque individu i . Les termes ε_i représentent les erreurs résiduelles, c'est-à-dire la différence entre la valeur observée et la valeur prédite par le modèle. L'objectif est de minimiser la somme des carrés de ces erreurs.

Ce modèle repose sur plusieurs hypothèses fondamentales :

- **Non-colinéarité** : les variables explicatives ne doivent pas être fortement corrélées entre elles ;
- **Indépendance des erreurs** : les résidus ε_i doivent être indépendants les uns des autres ;
- **Exogénéité** : l'espérance des erreurs conditionnellement aux variables explicatives doit être nulle ;
- **Homoscédasticité** : la variance des erreurs ε_i doit être constante pour toutes les observations i ;
- **Normalité des erreurs** : les résidus suivent approximativement une loi normale.

Nous avons réalisé ces régressions sur différentes bases de données. Afin de réduire les biais potentiellement liés à l'échantillonnage, nous avons décidé d'inclure la base de données de la DIR Ouest, présentée précédemment. Nous avons ensuite voulu étudier la proportion de chaque espèce en utilisant uniquement les collisions routières, car celles-ci constituent une source d'information moins dépendante de l'effort d'observation.

Ainsi, nous avons mené des régressions sur les combinaisons de données suivantes :

- Les collisions routières issues des naturalistes ;
- Les collisions routières issues des naturalistes et de la DIR Ouest ;
- Les données « Vu » des naturalistes et celles de la DIR Ouest.

Ces différentes combinaisons de données ne donnaient pas toujours les mêmes résultats.

Des modèles non valides

Pour valider nos modèles, nous avons examiné les résidus. En effet, afin de conclure qu'un modèle est robuste, les résidus doivent respecter les hypothèses énoncées précédemment. Étant donné la taille importante de notre base de données, la normalité stricte des résidus n'est pas essentielle : la loi des grands nombres garantit une approximation satisfaisante.

Cependant, nos différentes régressions se sont révélées peu efficaces. Les graphiques de diagnostic des résidus montrent clairement que les hypothèses de base ne sont pas respectées. Nous avons également effectué les tests de Breusch–Godfrey et de Breusch–Pagan afin de vérifier respectivement les hypothèses d'autocorrélation et d'homoscédasticité. Ces tests ont confirmé qu'une grande partie de nos modèles n'étaient pas valides au seuil de 95%. Les modèles dont les tests vérifiaient les hypothèses, avaient des coefficients non significatifs pour la variable « année ». Ces modèles ne nous intéressent donc pas.

Enfin, notre hypothèse selon laquelle l'espèce témoin ne fluctue pas au cours du temps peut introduire des incohérences supplémentaires dans l'interprétation des résultats.

Les modèles de régression ne permettent donc pas de modéliser correctement nos données. Nous avons donc choisi d'étudier des modèles plus complets que nous allons voir dans la partie suivante.

Le modèle *d'occupancy* : Un modèle séparant présence et détection

Après différentes régressions non fructueuses, nous avons choisi de recourir aux modèles *d'occupancy*. Ces modèles hiérarchiques utilisent des données de détection / non-détection en tenant compte de la possibilité qu'une espèce soit présente sur une maille sans être observée.

Définition d'un modèle *d'occupancy* [2]

Un modèle *d'occupancy* dynamique ou modèle *d'occupancy* sur plusieurs périodes décrit l'évolution de la probabilité de présence d'une espèce sur un ensemble de mailles au fil du temps. Il utilise des probabilités de présence, intégrées dans une structure hiérarchique, plutôt que de calculer des composantes déterministes comme le fait les modèles de régression.

Soient

- $t = 1, \dots, T$ les périodes (regroupement d'années),
- $i = 1, \dots, I$ les mailles,
- $j = 1, \dots, J$ les années

Dans ce modèle, on utilise une chaîne de Markov à 2 états :

- 0 : "La maille est inoccupée"
- 1 : "La maille est occupée"

Pour rappel, une chaîne de Markov à temps discret est un processus tel que la probabilité à un temps t ne dépend uniquement de la probabilité au temps précédent $t - 1$. Mathématiquement, ceci se représente de la façon suivante : Soit $Z_{t \in \mathbb{R}^{++}}$ la variable aléatoire représentant la présence a un temps t . Alors

$$P(Z_{t+1} = j | Z_t = i) = P(Z_{t+1} = j | Z_t = i, Z_{t-1} = i_{t-1}, \dots, Z_0 = i_0)$$

On définit 2 probabilités de transition afin de changer d'un état (occupé ou inoccupé) à un autre:

- γ : La probabilité de colonisation. La probabilité qu'une maille inoccupée devienne occupée, autrement dit $P(Z_t = 1 | Z_{t-1} = 0) = \gamma$
- ε : La probabilité d'extinction. La probabilité qu'une maille occupée devienne inoccupée, i.e. $P(Z_t = 0 | Z_{t-1} = 1) = \varepsilon$

Ce qui nous donne la matrice de transition suivante :

$$\begin{pmatrix} 1-\gamma & \gamma \\ \varepsilon & 1-\varepsilon \end{pmatrix} \quad \begin{pmatrix} \gamma \\ 1-\varepsilon \end{pmatrix} \quad (1)$$

L'atout de ce modèle est qu'il est hiérarchique et prend en compte une détection imparfaite des espèces. En effet, bien qu'un individu puisse être présent au temps t , il n'est pas forcément détecté tout au long de cette période. Plusieurs sondages j sont alors réalisés au sein de chaque période. Soit $H_{t \in \mathbb{R}^{++}} \in \{0; 1\}$ la probabilité de détecter l'espèce sur une maille. Alors $P(H_t = 1 | Z_t = 1) = p_{i,j} < 1$ et $P(H_t = 1 | Z_t = 0) = 0$ car on supposera dans notre cas qu'il n'y a pas de fausses détections.

On se retrouve donc avec la matrice de probabilité d'observation suivante pour une maille i à l'année j

$$P_{i,j} = \begin{matrix} & Z_t = 0 & Z_t = 1 \\ \begin{matrix} H_t = 0 \\ H_t = 1 \end{matrix} & \begin{pmatrix} 1 & 1-p_{i,j} \\ 0 & p_{i,j} \end{pmatrix} \end{matrix}$$

Posons $X_i^{(1)}, \dots, X_i^{(n)}$ les variables géographiques décrivant la maille i (distance au littoral, indice de diversité ...).

Notre modèle peut ainsi être représenté de la façon suivante :

Probabilité de détection :	$(p_{i,j}) = \alpha_0 + \alpha_1 \text{pression_observation}_{i,j}$
Probabilité de présence :	$(\psi_i) = \beta_0 + \beta_1 X_i^{(1)} + \dots + \beta_n X_i^{(n)}$
Loi de présence (1ère période) :	$z_{i,1} \sim \text{Bernoulli}(\psi_1)$
Loi de présence (t selon $t - 1$) :	$(z_{t,i} z_{t-1,i} = 0) \sim \text{Bernoulli}(\gamma), \quad t = 2, \dots, T$
	$(z_{t,i} z_{t-1,i} = 1) \sim \text{Bernoulli}(1 - \varepsilon), \quad t = 2, \dots, T$
Loi de détection :	$h_{t,i,j} z_{t,i}, p_{i,j} \sim \text{Bernoulli}(z_{t,i} \times p_{i,j})$

Le modèle complet est ensuite estimé par maximum de vraisemblance, en construisant la fonction de vraisemblance à partir de l'ensemble des historiques de détection.

Les hypothèses principales [5] de ce modèle sont :

- L'état d'une maille ne change pas au cours d'une période ;
- La probabilité d'occupation est identique sur toutes les mailles (pas d'hétérogénéité inexpliquée entre les mailles) ;
- La probabilité de détection, sachant que l'espèce est présente, est identique à conditions égales (pas d'hétérogénéité inexpliquée entre mailles) ;
- Les détections sont indépendantes entre elles ;
- L'historique de détection est indépendant d'une maille à l'autre ;
- Il n'existe pas de fausses détections.

Dans ce rapport, le package `unmarked` [7] est utilisé. Ce package utilise une approche dite fréquentiste au problème, en calculant des estimateurs de maximums de vraisemblance.

Différentes périodes pour différentes espèces

Dans chacune des mailles, nous considérons qu'il y a détection dès lors qu'au moins une observation de l'espèce est réalisée au cours de l'année. Si aucune observation n'est enregistrée, nous supposons que l'espèce n'a pas été détectée dans cette maille.

La base de données de la DIR Ouest ne contenant des données qu'à partir de 2014, elle n'a pas été utilisée dans ce modèle.

Il est souvent avantageux de définir des périodes d'une durée variable selon l'espèce étudiée. [2] Nous avons choisi d'utiliser des périodes plus longues (5 ans) pour les espèces difficiles à détecter, et des périodes plus courtes (3 ans) pour les autres. Nous avons considéré que les petits mustélidés (par exemple le putois et la fouine) étaient moins facilement détectables, à cause de leur taille et de leur mode de vie.

Les périodes retenues sont donc les suivantes :

- Espèces moins facilement détectables : [2010–2014], [2015–2019], [2020–2024] ;
- Espèces plus facilement détectables : [2010–2012], [2013–2015], [2016–2018], [2019–2021], [2022–2024].

Analyse menée pour chaque espèce :

Nous retirons les variables présentant une forte corrélation au sein d'un même modèle 6. Pour cela, nous utilisons le coefficient de corrélation de Pearson et excluons les combinaisons de variables où le coefficient dépasse 0,5. ([3]).

Par ailleurs, une variable de pression d'observation a été construite à partir de l'ensemble des données relevées. Contrairement aux régressions linéaires où les proportions prenaient implicitement en compte la pression d'observation, il a été nécessaire de l'inclure

explicitement ici. Cette variable est définie indépendamment du nombre d'observations de l'espèce étudiée.

Nous centrons et réduisons ensuite les variables susceptibles d'influencer la détection ou la présence des espèces. Bien que cela rende l'interprétation des paramètres moins intuitive, cette étape est essentielle à la construction du modèle.

Comme mentionné précédemment, nous utilisons le package `unmarked` afin d'appliquer différents modèles dynamiques. Nous sélectionnons ensuite les variables explicatives à tester, en retenant au maximum deux variables lorsque le $\Delta AIC < 2$ [5]. Le critère d'information d'Akaike, AIC, est une mesure de la qualité d'un modèle statistique qui trouve un compromis entre la complexité et la qualité d'un modèle. On choisit en général un modèle avec un AIC faible. On choisit donc le modèle dont l'écart avec l'AIC le plus faible est peu élevé.

Peu de travaux se sont intéressés à la validation des modèles d'occupation [6]. Parmi ceux-ci, la majorité recourent à des *goodness-of-fit tests*, en particulier celui proposé par MacKenzie et Bailey [8]. D'autres études se contentent de sélectionner le modèle présentant la plus faible valeur d'AIC ou d'AIC corrigé, sans véritablement aborder la question de la validation.

Un *goodness-of-fit test*, ou test d'adéquation, est une procédure statistique visant à évaluer dans quelle mesure un modèle parvient à représenter correctement les données observées. Le principe repose sur la comparaison entre les probabilités ou fréquences prédites par le modèle et celles effectivement observées dans l'échantillon. Lorsque les écarts sont faibles, le modèle est considéré comme adéquat. Dans le cas contraire, il est jugé insuffisant pour décrire correctement les données.

Dans notre étude, nous avons soumis les modèles à deux tests d'adéquation. Le premier repose sur un *bootstrapping* paramétrique [7], tandis que le second correspond au test de MacKenzie et Bailey, fondé sur une statistique du χ^2 . Le test par *bootstrapping* paramétrique accepte la grande majorité des modèles, alors que celui de MacKenzie et Bailey, plus restrictif, ne valide qu'un seul modèle : celui concernant la fouine.

Une augmentation de la population des fouines

Étant donné que le modèle concernant la fouine est le seul pour lequel les deux tests statistiques sont validés, nous choisissons de consacrer la suite de notre analyse sur cette espèce.

Le modèle retenu s'écrit ainsi :

$$\text{logit}(p_{i,j}) = -1,515 + 0,535 \times \text{PressionObservation}_{i,j}$$

$$\text{logit}(\psi_{1,i}) = 0,602 - 0,524 \times \text{Dist_EcotoneArbore}_i - 0,358 \times \text{Dist_Littoral}_i$$

$$z_{1,i} \sim \text{Bernoulli}(\psi_{1,i})$$

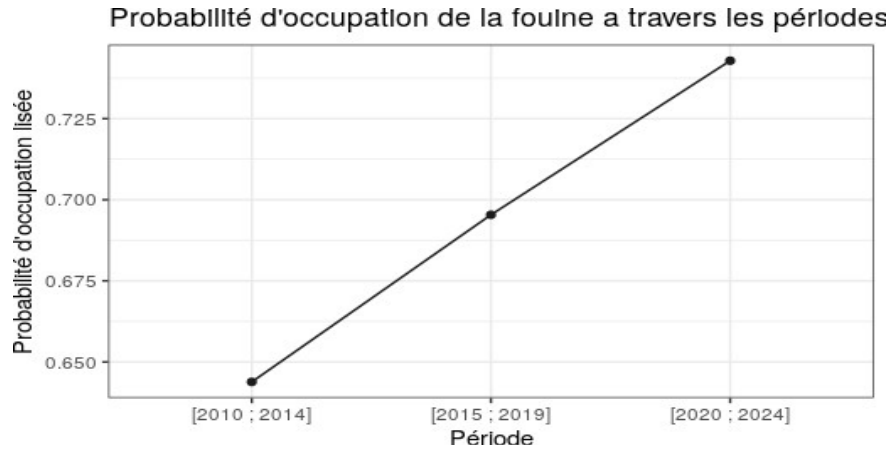
$$(z_{t,i} = 1 \mid z_{t-1,i} = 0) \sim \text{Bernoulli}(0,345), \quad t = 2, 3$$

$$(z_{t,i} = 1 \mid z_{t-1,i} = 1) \sim \text{Bernoulli}(0,902), \quad t = 2, 3$$

$$h_{t,i,j} \mid z_{t,i}, p_{i,j} \sim \text{Bernoulli}(z_{t,i} \times p_{i,j})$$

Ces équations indiquent que la probabilité de présence de la fouine est plus élevée à proximité des écotones arborés ainsi que des zones littorales. En moyenne, l'espèce occupe légèrement plus de la moitié du territoire étudié, mais elle reste difficile à détecter. En effet, sans tenir compte de la pression d'observation, la probabilité de détection d'une fouine est seulement de 0,180 ($\frac{1}{1+e^{1,515}} \approx 0,180$).

La probabilité d'occuper une maille précédemment inoccupée est de 0,345, tandis que la probabilité de disparaître d'une maille précédemment occupée est de $1 - 0,902 = 0,098$. Comme la probabilité d'occupation initiale est de l'ordre de 0,6, cela suggère que la population de fouines est en expansion. Cette tendance est visible dans le graphique suivant :



Évolution de la probabilité de présence de la fouine au cours des quinze dernières années

Lecture : Entre 2015 et 2019, la probabilité d'occupation moyenne était d'environ 0,7. Échantillon : 231 mailles présentant au moins une détection sur les 15 années d'étude.

Enfin, il est possible d'estimer l'état stationnaire, c'est-à-dire la probabilité d'occupation à long terme d'une maille i , à l'aide de la formule suivante :

$$\lim_{t \rightarrow \infty} (\psi_t) = \frac{\gamma}{\gamma + \varepsilon}$$

[9]. En appliquant cette relation, et en supposant que la dynamique observée au cours des 15 dernières années se maintienne, la probabilité d'occupation de la fouine en Bretagne tendrait vers une valeur d'équilibre de 0,779.

Conclusion et perspectives

Il est probable que plusieurs de nos modèles *d'occupancy* n'aient pas fonctionné en raison du non-respect de l'hypothèse d'indépendance entre les sites. En effet, si une espèce est détectée dans une maille donnée, la probabilité de la retrouver dans les mailles voisines est généralement plus élevée que dans une maille choisie aléatoirement. Cette autocorrélation spatiale remet en cause l'indépendance des observations, hypothèse centrale des modèles *d'occupancy*.

Deux approches peuvent être envisagées pour remédier à ce problème. La première consiste à recourir à une stratification [5], c'est-à-dire à attribuer des probabilités de sélection aux mailles, puis à les choisir en fonction de ces probabilités. L'intégration d'informations sur les mouvements des espèces permettrait d'améliorer cette sélection et de limiter les dépendances entre mailles.

Une seconde solution consiste à utiliser des modèles *d'occupancy* intégrant explicitement la dépendance spatiale entre sites [11]. Certains développements méthodologiques récents permettent en effet de modéliser des probabilités de colonisation plus élevées pour les sites situés à proximité immédiate d'une détection. Ces approches pourraient être appliquées afin de mieux rendre compte des dynamiques spatiales des populations.

Par ailleurs, il est possible que la pression d'observation dans la période considérée soit trop variable dans le temps et l'espace. Couplé à la faible probabilité de détection de la plupart des espèces, cette trop grande variabilité empêche peut-être de dégager des résultats robustes. Des variables liées aux conditions météorologiques (par exemple le pourcentage de jours de pluie, les épisodes de chaleur extrême ou d'orages intenses) pourraient également être intégrées afin de mieux caractériser l'hétérogénéité de la probabilité de détection. Il serait également intéressant de tester l'utilisation de la période en tant que variable aléatoire. En outre, l'ajout d'autres sources de données (collisions routières, prélèvements chasse, réseaux sociaux, études spécifiques, par exemple sur la fécondité...) pourrait être utile.

D'autres pistes d'amélioration résident dans des approches spécifiques :

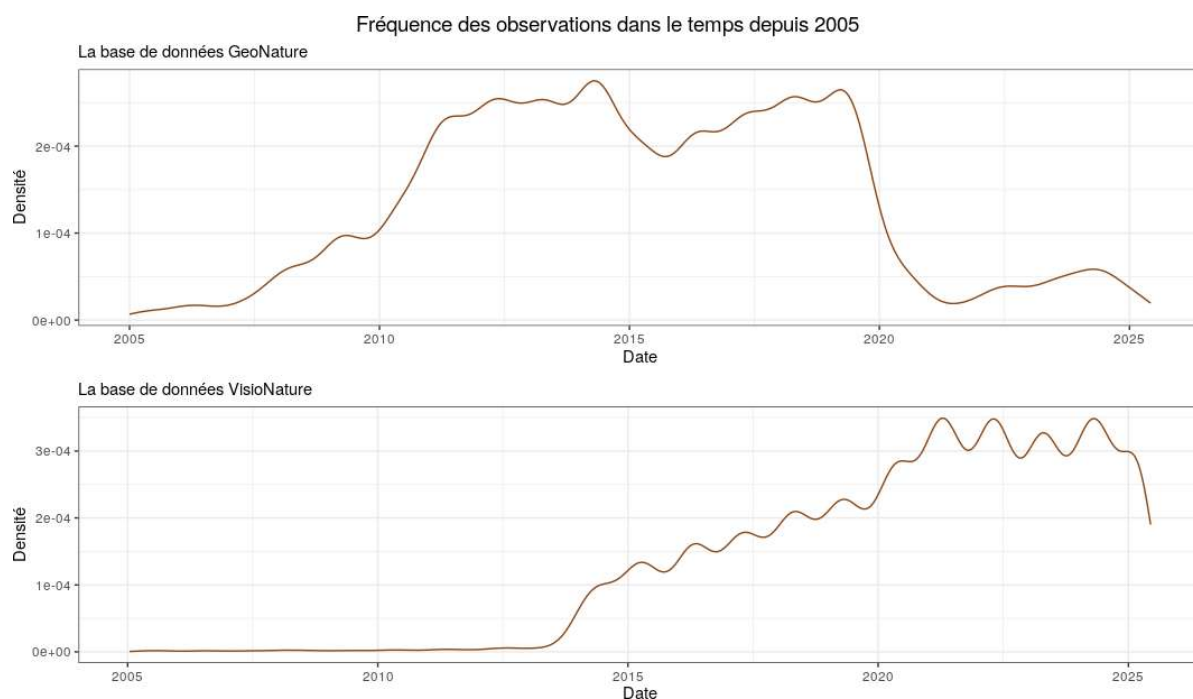
- Travailler sur les groupes d'observateurs afin de grouper les espèces ayant un processus d'observation voisin,
- Tester des modèles multiespèces pour trouver des dynamiques associées (similaires ou approchantes, ou opposées). Ce type de modèle permettrait, pour les espèces à faible présence ou détection, de tirer des informations des autres espèces semblant réagir dans le même sens.

Malgré ces limites, nos résultats montrent qu'il est possible de tirer parti des données opportunistes pour analyser les dynamiques de certaines populations, à condition de prendre en compte les biais potentiels liés à leur nature. Dans le cadre de ce travail, un seul modèle robuste a pu être validé, celui concernant la fouine. Ce résultat illustre à la fois le potentiel et les contraintes de l'utilisation de données opportunistes en écologie, et souligne l'importance de développer des méthodes adaptées pour valoriser au mieux ces jeux de données hétérogènes.

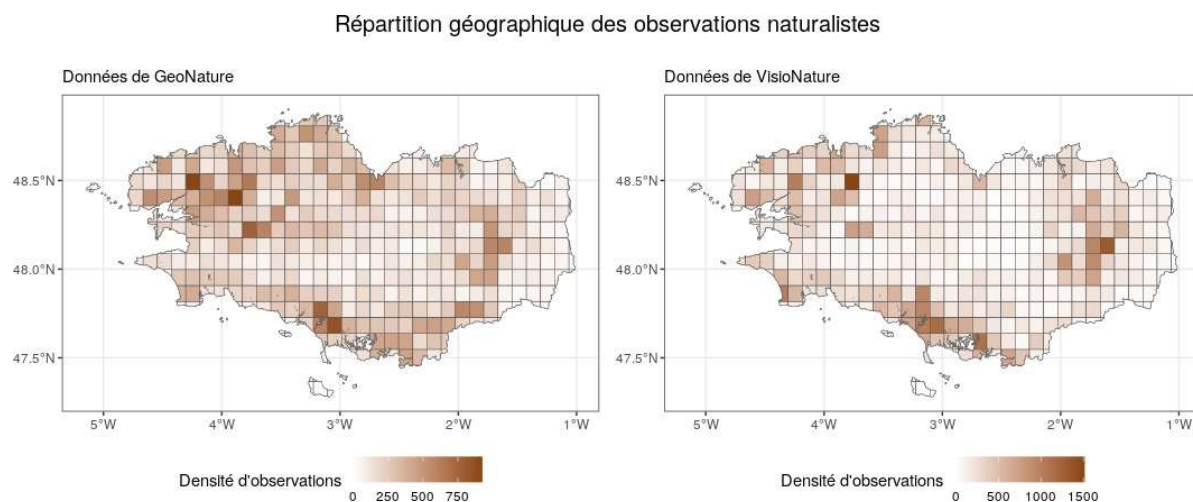
Bibliographie

- [1] Isaac, N.J.B., van Strien, A.J., August, T.A., de Zeeuw, M.P. and Roy, D.B. (2014), *Statistics for citizen science : extracting signals of change from noisy ecological data*. *Methods Ecol Evol*, 5 : 1052-1060. <https://doi.org/10.1111/2041-210X.12254>
- [2] Darryl I. MacKenzie, James D. Nichols, J. Andrew Royle, Kenneth H. Pollock, Larissa L. Bailey, James E. Hines, (2018) *Occupancy Estimation and Modeling (Second Edition)*, <https://doi.org/10.1016/B978-0-12-407197-1.00002-8>.
- [3] Matthew J. Gould, William R. Gould, James W. Cain, Gary W. Roemer, *Validating the performance of occupancy models for estimating habitat use and predicting the distribution of highly-mobile species : A case study using the American black bear*, *Biological Conservation*, Volume 234, (2019), <https://doi.org/10.1016/j.biocon.2019.03.010>.
- [4] S. Adamantopoulou, A A Karamanlidis, P. Dendrinou, O. Gimenez. *Citizen science indicates significant range recovery and defines new conservation priorities for Earth's most endangered Pinniped in Greece*. *Animal Conservation*, Wiley, (2022), 10.1111/acv.12806. hal-03718409
- [5] Bailey, Larissa Lynn and Michael J. Adams. *Occupancy models to study wildlife*. (2005).
- [6] Warton, D. I., Stoklosa, J., Guillerá-Arroita, G., MacKenzie, D. I., Welsh, A. H. (2017). *Graphical diagnostics for occupancy models with imperfect detection*. *Methods in Ecology and Evolution*, 8(4), 408-419. <https://doi.org/10.1111/2041-210x.12761>
- [7] Fiske, I., & Chandler, R. (2011). *unmarked : An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance*. *Journal of Statistical Software*, 43(10), 1–23. <https://doi.org/10.18637/jss.v043.i10>
- [8] MacKenzie, D. I., Bailey, L. L. (2004) *Assessing the fit of site-occupancy models*. *Journal of Agricultural, Biological, and Environmental Statistics* 9, 300–318.
- [9] Hanski, Ilkka. *A practical model of metapopulation dynamics*. *Journal of animal ecology* (1994) : 151-162. <https://doi.org/10.2307/5591>
- [10] Hill, M.O. (2012), *Local frequency as a key to interpreting species occurrence data when recording effort is not known*. *Methods in Ecology and Evolution*, 3 : 195-205. <https://doi.org/10.1111/j.2041-210X.2011.00146.x>
- [11] Solà, O., Kéry, M., Aquilué, N., & Brotons, L. (2025). *Sampling effort required for fitting spatially explicit models of species distribution dynamics*. *Ecological Modelling*, 501, 110998. <https://doi.org/10.1016/j.ecolmodel.2024.110998>

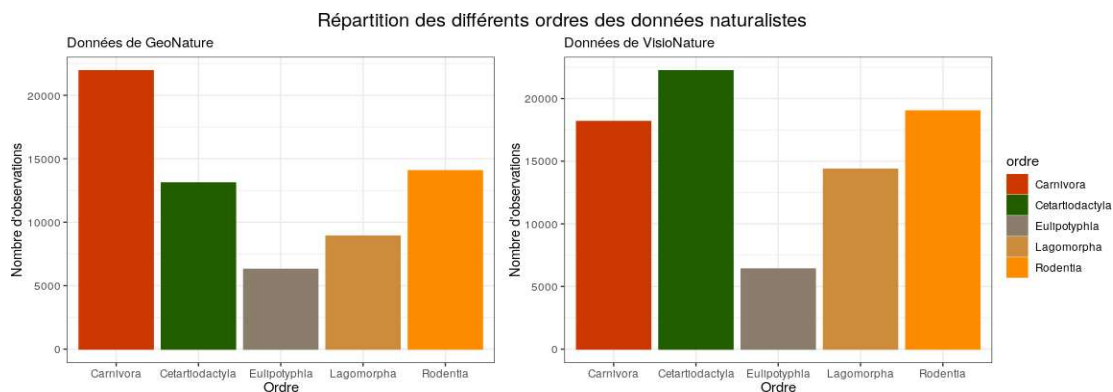
Annexes



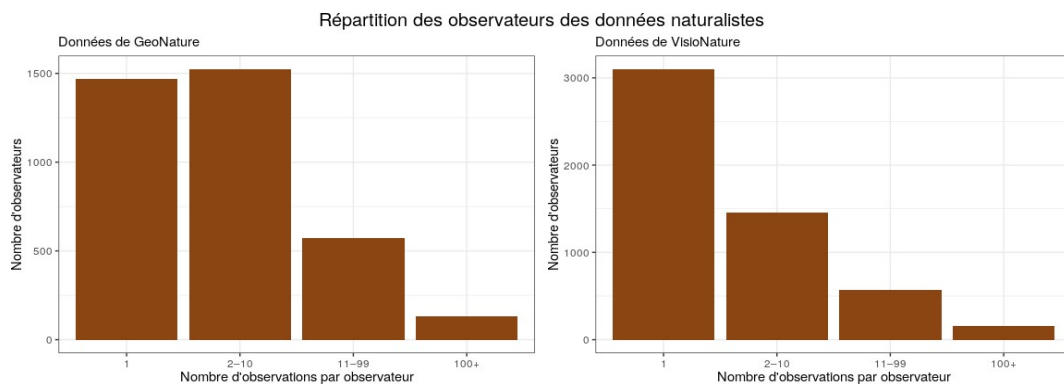
ANNEXE 1 - La fréquence des distributions varie dans le temps, et en fonction des bases de données - Lecture : Le nombre de données opportunistes de GeoNature a chuté aux alentours de 2020. Échantillon : 70245 GN, 86027 VN à partir de 2005 .



ANNEXE 2 - La fréquence des distributions varie géographiquement et en fonction des bases de données - Lecture : Près de 750 données de GéoNature sont situées dans la maille où se situe la maison de la rivière. - Échantillon : 72142 GN, 79585 VN, entre 2010 et 2024, après suppression des 2 principaux observateurs de VN dans les 2 bases de données.

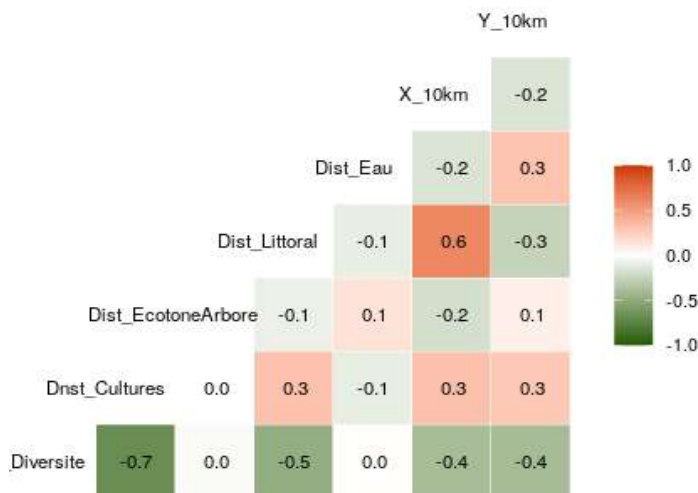


ANNEXE 3 - La répartition entre taxons est proche entre les deux bases, excepté pour les Carnivores et les Cetartiodactyles - Échantillon : 64291 GN, 80154 VN situées entre 2010 et 2024.



ANNEXE 4 - Les observateurs ne renseignent pas les deux bases de la même façon (ils diffèrent dans leurs pratiques) - Lecture : Près de 1500 observateurs ont inséré entre 2 et 10 observations dans la base de données de VisioNature - Échantillon : 73796 GN, 86958 VN .

Graphique de corrélation des différents variables explicatives quantitatives



ANNEXE 5 - Les variables géographiques sont parfois linéairement corrélées entre elles - Lecture : La longitude et la distance au littoral ont un coefficient de corrélation linéaire de 0,6. - Échantillon : Les 315 mailles de 10 × 10km².

ANNEXE 6 - Données présentes dans nos jeux de données après nettoyage

Nom	Description
id_synt	Un numéro de l'observation
date	La date à laquelle l'espèce a été observée
cd_nom	Un nombre qui identifie l'espèce observée
nom_valide	Le nom valide de l'espèce observée
nom_vernaculaire	Le nom vernaculaire de l'espèce observée
ordre	L'ordre de l'espèce observée
nombre	Le nombre d'animaux observés
observateurs	Le nom et prénom, ou l'identifiant de l'espèce
cmmt_c	Un commentaire que l'observateur peut faire lors de sa saisie
technique_observation	La technique d'observation, après modification : - Vu, - Entendu, - Indices de présence
etat_biologique	L'état biologique de l'animal lors de la saisie : - Vivant, - Trouvé mort, - Trouvé mort : collision routière, - Inconnu
bdd_originale	La base de données de laquelle la donnée provient : - GeoNature, - VisioNature, - DIRO
geometry	Les coordonnées géographiques où l'espèce a été observée

ANNEXE 7 - Les espèces ne sont pas présentes sur tous les sites (=mailles 10x10 km) à toutes les saisons (=périodes) - Lecture : 225 sites retrouvent le Renard sur toutes les saisons - Échantillon : Les 315 sites sur observations de la base de données complète

	Nombre de présences sur x saisons					
	x=0	x=1	x=2	x=3	x=4	x=5
Sanglier	42	57	67	69	43	39
Renard	3	8	15	19	47	225
Lapin	10	17	27	51	60	152
Hérisson	2	3	8	28	52	224
Putois	98	106	63	50		
Hermine	245	56	15	1		
Fouine	86	91	83	57		

ANNEXE 8 - Le nombre de sites par période ne reflète pas forcément les résultats obtenus -
*Lecture : Sur la période [2019 ; 2021], 251 sites ont eu une détection de lapin. - Échantillon :
 Les 315 sites sur observations de la base de données complète*

	Nombre de sites avec une présence				
	[2010 ; 2012]	[2013 ; 2015]	[2016 ; 2018]	[2019 ; 2021]	[2022 ; 2024]
Sanglier	129	143	161	160	172
Renard	262	276	286	286	298
Lapin	222	248	257	251	246
Hérisson	273	289	285	287	297
	[2010 ; 2014]	[2015 ; 2019]	[2020 ; 2024]		
Putois	137	119	126		
Hermine	30	25	34		
Fouine	144	140	144		