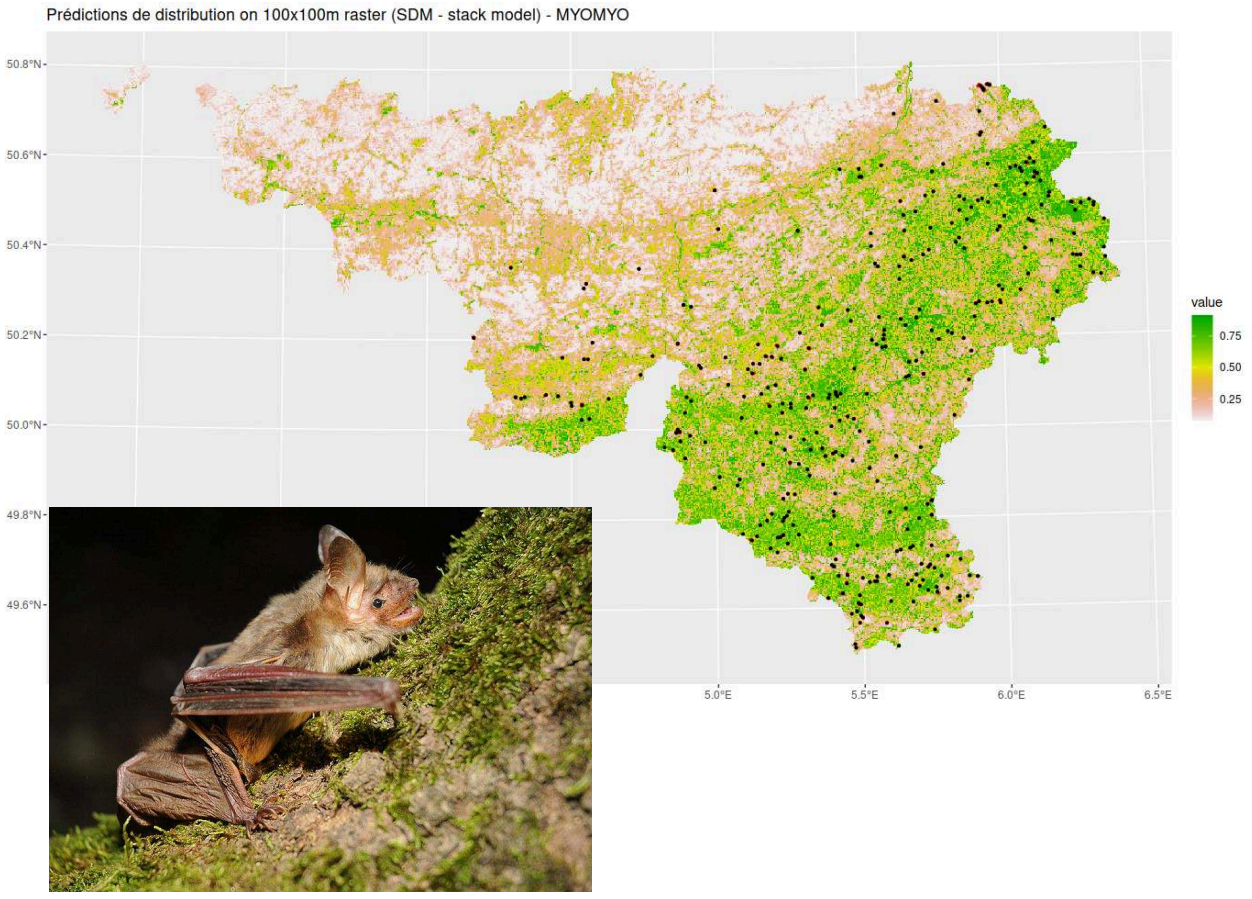
Marché public de services visant à étudier l'activité et la mortalité des chauves-souris sur des parcs éoliens forestiers en Wallonie

LOT 3 - Modélisation des habitats naturels des chauves-souris forestières sur base des données acoustiques existantes

Rapport final



Colophon

Étude et rapport réalisés par

Ecofirst SC
Société coopérative
BE 0692.806.959
www.ecofirst.eu
Adresse postale : Grand-Rue 12 - 6870 Awenne - Belgique
Administrateur responsable : Gérard Jadoul gerard.jadoul@ecofirst.eu
Personnes de contact pour ce dossier :
Pierrette Nyssen 0473 265 264 pierrette.nyssen@ecofirst.eu
et Jean-François Godeau 0472 94 48 47 jf.godeau@ecofirst.eu

pour le compte de

Service Public de Wallonie (SPW)
Agriculture, Ressources Naturelles et Environnement (ARNE)
Département de l'Etude du Milieu Naturel et Agricole (DEMNA)
Direction de la Nature et de l'Eau (DNE)
Adresse postale : Avenue Maréchal Juin, 23 à 5030 Gembloux - Belgique
Pouvoir adjudicateur : Bénédicte Heindrichs, Directrice générale du SPWARNE
Fonctionnaire dirigeant : Annick Terneus annick.terneus@spw.wallonie.be
Personne de contact pour ce dossier :
Thierry Kervyn 0477 26 03 99 thierry.kervyn@spw.wallonie.be

dans le cadre du

marché public de services visant à étudier l'activité et la mortalité des chauves-souris sur des parcs éoliens forestiers en Wallonie - Cahier spécial des charges n°O3.02.03-22-3285
et en particulier son LOT 3 intitulé "Modélisation des habitats naturels des chauves-souris forestières sur base des données acoustiques existantes"

Composition du comité d'accompagnement :

- SPWARNE – DEMNA : Thierry Kervyn, Quentin Smits, Annick Terneus, Jérémy Simar
- SPWARNE – DNF : Jean-Philippe Bizoux, Corentin Laroy
- ECOFIRST : Jean-François Godeau, Pierrette Nyssen

Rédaction et mise en page : Jean-François Godeau, Pierrette Nyssen

Photos :

- couverture : grand murin - cc Ján Svetlík
- les images utilisées dans ce rapport, sauf mention explicite, sont d'Ecofirst : Jean-François Godeau, Pierrette Nyssen

Date de finalisation du rapport : 11/12/2023

Référence : Godeau, J.-F. & Nyssen, P., Modélisation des habitats naturels des chauves-souris forestières sur base des données acoustiques existantes, Rapport final, Décembre 2023, Ecofirst.

Table des matières

1. Contexte de l'étude et termes du marché

Cette étude constitue le lot 3 du marché public de services visant à étudier l'activité et la mortalité des chauves-souris sur des parcs éoliens forestiers en Wallonie (Cahier spécial des charges n°O3.02.03-22-3285) émis par le Service Public de Wallonie. Ce marché porte sur une mission visant à organiser la récolte et l'analyse de données de terrain relatives aux chauves-souris, notamment en milieu forestier en Wallonie. Cet appui contribuera au monitoring régional, à l'appui aux gestionnaires et aux rapportages internationaux. Le marché a été conclu pour une durée de 12 mois, débutant en date du 13/12/2022, et le lot 3 a été remporté par **Ecofirst SC**.

Le lot 3 porte sur la modélisation des habitats naturels des chauves-souris forestières sur base des données acoustiques existantes. Le cahier spécial des charges définit le contexte et les objectifs de ce lot en ces termes :

La Wallonie contribue tous les 6 ans au rapportage européen imposé par la Directive « Habitats ». Une analyse statistique modélisant la répartition (RANGE) des espèces de chauves-souris forestières contribuera à préciser cet élément lors du prochain rapportage européen.

Objectifs :

- *Identifier de façon détaillée par modélisation spatiale des données acoustiques les parties du territoire wallon susceptibles d'être exploitées par espèces forestières de chauves-souris : le Grand murin, la Noctule commune, la Noctule de Leisler, la Pipistrelle commune et la Pipistrelle de Nathusius*
- *Appuyer des mesures de protection d'espèces sensibles par la diffusion des informations pertinentes aux gestionnaires concernés et par la collecte des compléments d'information nécessaires*
- *Analyser l'importance de la variable relative à la distance aux lisières forestières*

Dans le cadre de sa mission, l'adjudicataire abordera les aspects suivants :

1. *Collecte, rassemblement, validation et traitement par modélisation spatiale des données acoustiques enregistrées par le SPW-ARNE depuis 2016*
2. *Production d'une carte de probabilité de présence de chaque espèce sur base des variables environnementales explicatives pertinentes*
3. *Synthèse des résultats au format (RANGE) du rapportage pour la Directive européenne « Habitats »*

2. Méthodologie de la modélisation

Ce rapport présente les manipulations et analyses de données que nous avons mises en œuvre pour créer des modèles de distributions de 6 espèces de chauves-souris à l'échelle de la Région Wallonne. Nous y détaillons chaque étape avec force détails pour que l'ensemble soit reproductible. Les scripts R d'analyse sont délivrés en annexe et le texte cite clairement les packages et fonctions clé ainsi que certains termes de jargon qui ont été utilisés.

2.1 Utilisation de la librairie R 'tidysdm' : principe général du flux d'analyse

Depuis la première réunion du comité d'accompagnement où nous avions présenté l'approche méthodologique pressentie, les outils statistiques de modélisation de distribution d'espèces (*SDM - Species Distribution Models*) ont évolué. Un nouveau package R nommé **tidysdm** a vu le jour en 2023, accompagné de documents de référence : un site web et une publication encore au stade de prépublication à ce jour. Cet outil de modélisation est particulièrement intéressant car il permet d'utiliser les deux modèles prévus de manière uniformisée pour créer un jeu de données de prédiction.

Tidysdm est une adaptation de tidymodels orientée pour les Species Distribution Models. Ce package fait appel à plusieurs packages pré-existants :

L'ensemble de la procédure comprend les étapes suivantes :

- A. **Préparation des données** :
  - (1) sélectionner le jeu de données de l'espèce à modéliser (*outcome* = présence) et les facteurs environnementaux (*descripteurs*)
  - (2) affiner la sélection pour minimiser l'autocorrélation spatiale
  - (3) créer/sélectionner des données de pseudoabsence/d'absence
  - (4) constituer aléatoirement un lot de données d'entraînement (*training*) et de test, ces dernières ne sont pas utilisées pour l'ajustement des paramètres des modèles mais serviront à valider la qualité des prédictions.

- B. **Définition d'une procédure (*workflow*)** spécifiant une '*recette*' qui identifie les données à analyser, si elles sont de nature binaire ou multiclasse, et spécifie le ou les modèles qui seront utilisés. La '*recette*' et les modèles sont groupés dans un objet nommé *workflowset*. Cette étape consiste à ajuster les hyperparamètres[1] des modèles pour qu'ils expliquent au mieux les résultats attendus (*outcome*). Pour entraîner les modèles, les hyperparamètres sont ajustés sur le lot d'entraînement par validation croisée de groupes géographiques (*block cross-validation*), où N groupes contenant 10% vs. 90% des données *training* sont confrontés N fois. La force de *tidymodels* (et donc de *tidysdm*) est d'uniformiser la formulation du *workflowset* pour que des modèles traditionnellement définis sous des formes diverses (au travers de différents packages) puissent facilement traiter le même jeu de données sans multiplier les scripts d'analyse.

[1] On utilise le terme 'hyperparamètres' car il s'agit de paramètres propres à des algorithmes eux-même inclus dans les fonctions du package tidymodels.

- C. **Test de l'adéquation de chaque modèle** : L'adéquation des modèles à prédire correctement les données test est mesurée via plusieurs métriques, ici aussi selon une formulation uniformisée.

- D. **Combinaison des modèles** : Chaque modèle et ses meilleurs hyperparamètres sont ensuite combinés en un ensemble afin de calculer des prédictions sur base de la moyenne des valeurs prédites indépendamment par chaque modèle (voir détails au point 3.7).

- E. **Empilement des modèles** : On gagne ensuite encore en précision en créant un ensemble 'empilé' (*stack*) qui combine cette fois un ou plusieurs modèles, voire plusieurs versions du même modèle doté des différentes combinaisons d'hyperparamètres. Les métriques décrites précédemment sont enfin calculées pour cet ensemble empilé.

- F. **Production de la prédiction finale :** Les prédictions du modèle empilé sont calculées en tous points de l'aire étudiée (emprise des rasters des facteurs environnementaux) pour produire le résultat final, à savoir une probabilité de détection (une valeur allant de 0 à 1). Ce résultat est exporté sous forme d'un geotif exploitable en SIG.

2.2 Travail préparatoire

2.2.1 Collecte des jeux de données

Les données utilisées proviennent de la base de données acoustiques fournie par le DEMNA, auxquelles on a ajouté les résultats d'EIE de projets éoliens fournis par les bureaux d'étude CSD Ingénieurs et Sertius et une partie des données de l'étude menée par Ecofirst sur Nassonia.

2.2.2 Nettoyage

Une base de données qui agglomère des sources variées (bénévoles, EIE de bureaux d'études, programmes d'inventaires planifiés…) doit toujours faire l'objet d'une validation de son intégrité et de quelques corrections. Les corrections suivantes ont été opérées sur base des données mises à disposition par le DEMNA :

- uniformisation de l'encodage des caractères spéciaux (accents), de majuscules/minuscules ou d'espaces surnuméraires
- ajout d'un identifiant univoque par donnée
- filtrage (exclusion) de certains jeux de données ne provenant pas d'un protocole d'enregistrement passif en un point fixe, ou de nuits complètes d'inventaire
- corrections de quelques coordonnées géographiques (erreur de format)
- redondances en un même point (X-Y) ou sur plusieurs points de nombreuses données
- exclusion de données situées hors de la Région wallonne

La base de données nettoyée est fournie en délivrable de cette mission, ainsi que le script R, lequel pourra être adapté et réutilisé pour répéter ce travail à l'avenir si nécessaire.

Étapes des opérations de nettoyage de la base de données

La description détaillée ci-dessous a pour but d'expliciter le contenu du script R mais aussi de consigner, étape par étape, ce nettoyage.

Chargement des bibliothèques
Le script commence par charger les bibliothèques nécessaires à l'exécution, notamment *sf*, *tidyverse*, *mapview* et *stringi*.

Fonction de formatage
Ensuite, une fonction nommée *clean_by_site()* est définie. Cette fonction prend en paramètre un nom de site (`*SiteName*`) et effectue le nettoyage voulu sur les données du site correspondant.

Chargement des données DEMNA
Les données DEMNA sont chargées à partir d'un fichier CSV et converties en un objet R (data.frame). Certaines opérations de base sont effectuées, telles que la conversion des dates et l'ajout d'un identifiant unique (`*Serial*`) à chaque ligne.

Correction des problèmes d'encodage et de majuscules/minuscules
Des corrections sont apportées à certaines colonnes (notamment `*Habitat1*`, `*Habitat2*`, `*Id*`, et `*Site*`) pour uniformiser les noms et résoudre des problèmes d'encodage ou de typographie.

Exclusion de certains sites
Certains sites (notamment des transects sur la Semois ou sur le site des Marais d'Harchies) sont filtrés dès lors que leurs noms génériques '*Site*' ont pu être identifiés. Les données provenant d'enregistrements en sortie de cavités souterraines ont également été supprimées du jeu de données visant la modélisation des terrains de chasse.

Correction ou élimination des coordonnées géographiques erronées
Les coordonnées géographiques sont corrigées au cas par cas, et des problèmes de coordonnées manquantes ou incorrectes (tronquées lors d'un import-export dans un tableur) sont résolus.

Boucle pour épurer les sites à redondances
Une boucle est utilisée pour nettoyer les sites contenant des données redondantes, en utilisant la fonction définie précédemment (*clean_by_site()*).

Élimination finale des duplicats
Les duplicats sont éliminés en fonction de la concaténation de plusieurs champs univoques : une espèce présente observée en un site à une date précise avec un nombre de contacts et un nombre de minutes positives doit nécessairement être unique.

Filtration des données en RW uniquement
Une couche cartographique de la Région wallonne est chargée, et seules les données incluses dans les limites régionales sont conservées.

Chargement des données fournies par CSD
Les données de CSD sont chargées à partir d'un fichier CSV, et certaines opérations de traitement sont effectuées pour obtenir une représentation propre des données.

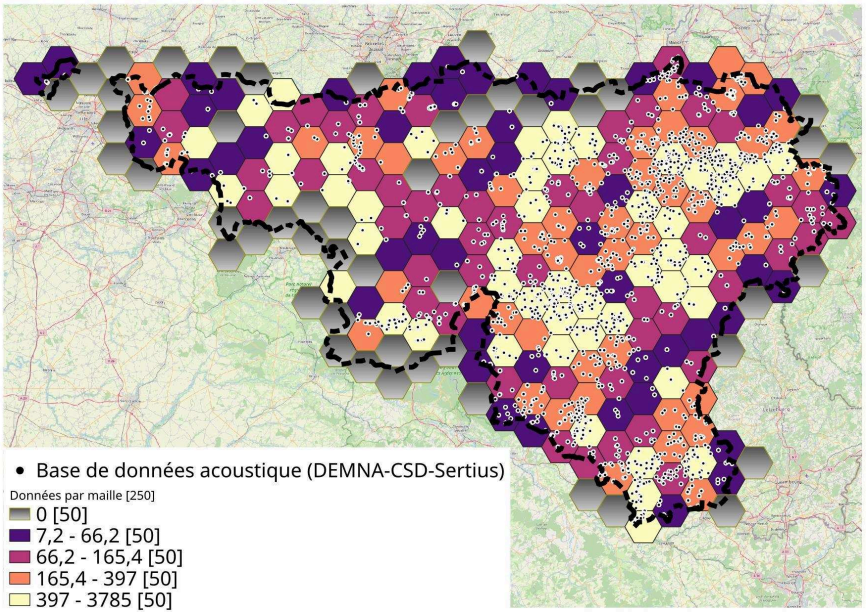
Chargement des données fournies par Sertius
Les données Sertius sont chargées à partir d'un fichier CSV, et des corrections sont apportées à certaines colonnes pour uniformiser les noms et résoudre des problèmes d'encodage.

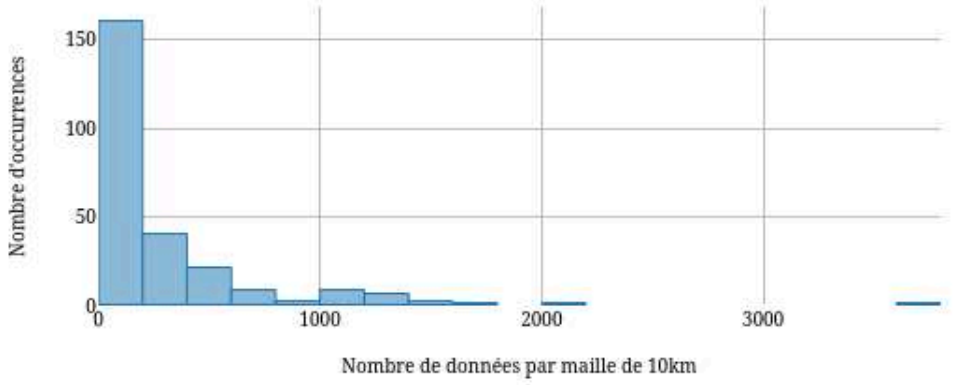
Fusion des bases de données
Les trois bases de données (DEMNA, CSD, et Sertius) sont fusionnées et exportées sous forme d'une couche cartographique geopackage `All_Bats.gpkg`.

Base de données finale en quelques chiffres
En résumé, la base fournie par le DEMNA comptait **54.476** entrées (lignes ou données). Après élimination des jeux de données à exclure pour des raisons de protocole, il restait 49.275 données.

Après suppression des jeux de données par nom de site où des répétitions étaient observées soit en différents lieux X-Y, soit en un même point, il restait 42.020 données.
On a supprimé 5.492 données hors Région wallonne.
On a ajouté 18.376 données résumées provenant de CSD.
On a ajouté 9.642 données résumées provenant de Sertius.
La base de données finalement utilisée pour les analyses de modélisation compte **65.068** données.

La répartition du nombre de données de la base finale sur l'emprise de la Région wallonne sur une maille de 10 km est la suivante :





3. Description du flux d'analyse et des paramétrages

Les étapes détaillées de l'analyse sont présentées ci-dessous pour la modélisation de *Pipistrellus nathusii*[2]. La description, étape par étape, de l'analyse permet de développer tous les choix méthodologiques opérés. Ceux-ci s'appuient sur l'article en preprint décrivant le fonctionnement du package *tidysdm* :

> *M. Leonardi, M. Colucci, A. Manica, 2023. tidysdm: leveraging the flexibility of tidymodels for Species Distribution Modelling in R. bioRxiv - doi: https://doi.org/10.1101/2023.07.24.550358*

La même procédure est valide pour l'analyse de chaque espèce, nous n'en détaillons qu'une seule ici par souci de concision. Nous avons décidé d'effectuer **chaque flux d'analyse selon deux méthodes** : l'une utilisant des **pseudoabsences** et l'autre des **vraies absences**.

En effet, par nature, un SDM se base sur la collection de données qui ne résultent pas d'un protocole d'inventaire standardisé. Autrement dit, on a toujours la certitude d'une présence de l'espèce (une ou plusieurs observations en un point) alors que l'absence reste toujours une hypothèse : soit l'espèce est réellement absente jusqu'à preuve du contraire, soit aucune prospection n'a été menée. Cette contrainte conduit les modèles à se baser sur des points de **pseudoabsence** choisis aléatoirement sur le territoire par le modèle en postulant que l'espèce ne s'y trouve probablement pas.

Les données d'inventaires acoustiques résultent pourtant de une ou plusieurs nuits complètes d'enregistrement passif, il est donc assez probable qu'une espèce non détectée sur une nuit d'enregistrement était effectivement absente au lieu et à l'endroit où le détecteur d'ultrasons a fonctionné. Cette hypothèse est d'autant plus vraie qu'il y a de nuits d'observation au même endroit, ce qui est éminemment variable et impossible à établir en l'absence de métadonnées provenant des opérateurs. On peut donc postuler que tous les points d'inventaire de la base de données où une espèce n'a pas été observée constitue une -très probable- **vraie absence** !

Les étapes décrites ci-dessous suivent le protocole d'analyse de présences vs. pseudoabsences mais il est identique lorsqu'on analyse des présences vs. vraies absences. Les deux scripts d'analyse sont fournis en annexe à ce rapport.

3.1 Chargement des facteurs environnementaux

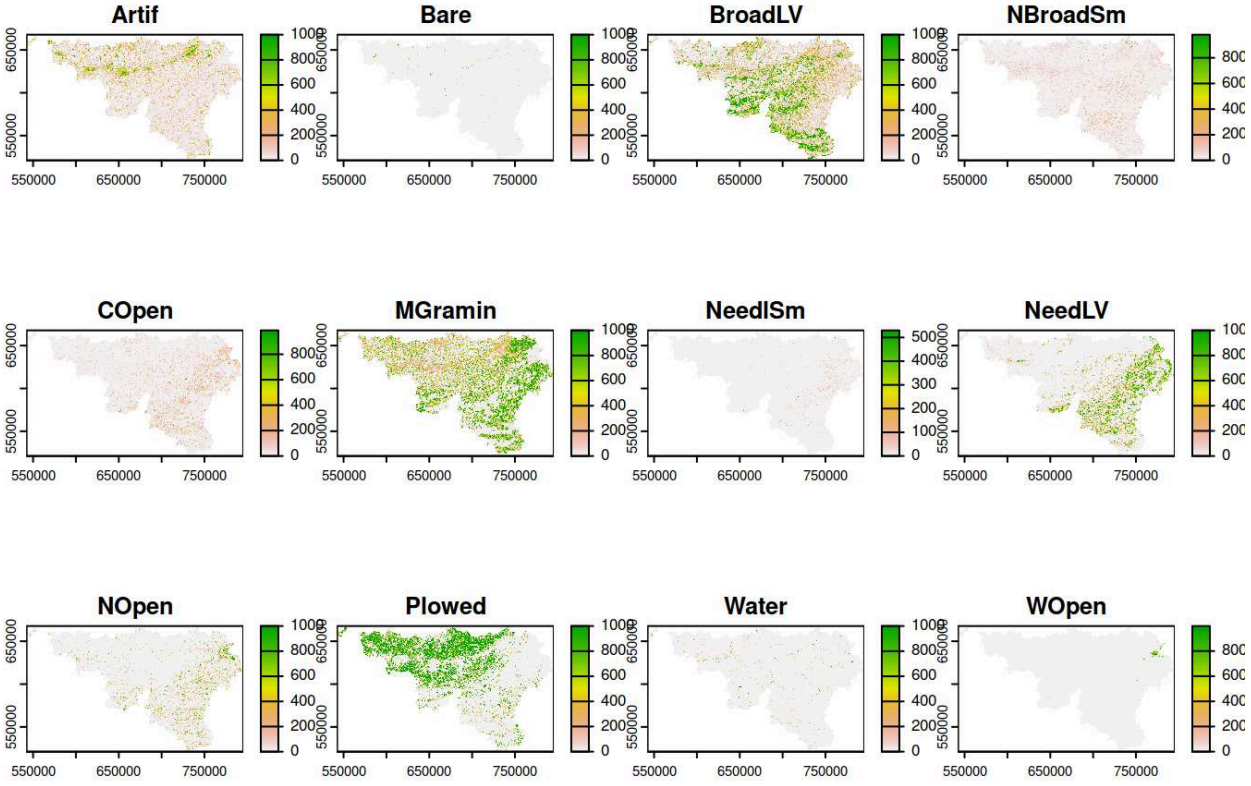
On a "rastérisé" une sélection arbitraire de 25 facteurs contenus dans la couche SIG Ecotope (lifewatch) en coordonnées projetées ETRS89/Lambert Conique Conforme 2008 (epsg:3812) selon une maille de 100x100m.
Chaque paramètre est extrait dans une couche raster via l'outil dédié de QGIS (*gdal:rasterize*).
Les facteurs environnementaux décrivent le pourcentage de présence de 12 descripteurs pour chaque polygone (exprimé par une valeur allant de 0 à 1000, autrement dit des 'pourmilles').

Ces facteurs sont les suivants (ils sont ensuite représentés sur des cartes en petit format). La définition reprise ici est une traduction du document de référence lifewatch qui décrit les différentes variables contenues dans le jeu de données géographique.

[2] Toutes les données reprises dans la base de données sous l'identification PIPNAT, PIP35 et PIPKUH ont été regroupées sous la dénomination PIPNAT/KUH et présentées ici par simplicité sous *P. nathusii* avec le code PIPNAT. La Pipistrelle de Kuhl n'étant a priori connue que de la région bruxelloise et identifiable avec certitude qu'en présence de cris sociaux, il semble plausible d'associer cette espèce ici au taxon P. *nathusii*.

 # Land cover

- **BroadLV** (alias *broadleaved trees*) : feuillus situés dans des forêts ou d'autres utilisations du sol (parcs, vergers), d'une hauteur supérieure à 3 mètres.
- **NeedlLV** (alias *needleleaved trees*) : conifères situés dans des forêts ou d'autres utilisations du sol (haies, jardins…), d'une hauteur supérieure à 3 mètres.
- **BroadSm** (alias *broadleaved shrubs*) : arbres ou arbustes feuillus, situés dans les forêts ou dans d'autres utilisations du sol (vignobles, jardins...), d'une hauteur inférieure à 3 mètres.
- **NeedlSm** (alias *needleleaved shrubs*) : arbres ou arbustes conifères situés dans des forêts ou d'autres utilisations du sol (jardins, arbres de Noël, etc.), d'une hauteur inférieure à 3 mètres.
- **Plowed** (alias *ploughed herbaceous cover*) : terres arables (cultures annuelles et couverture herbacée temporaire).
- **MGramin** (alias *permanent monospecific graminoid cover*) : couverture permanente de graminoïdes monospécifiques résultant d'utilisations intensives des terres, telles que les prairies, parcs et jardins.
- **NOpen** (alias *open area with relatively dry soils*) : couverture herbacée permanente mélangée à d'autres végétaux non ligneux. Cette classe couvre un grand nombre de sites à haut potentiel pour la biodiversité, tels que les landes ou les prairies extensives.
- **WOpen** (alias *open area with humid soil*) : couverture herbacée inondée mélangée à d'autres végétaux non ligneux. Cette classe couvre un grand nombre de sites à haut potentiel pour la biodiversité, tels que les zones humides et les tourbières.
- **COpen** (alias *disturbed open area*) : couverture permanente d'herbacées et d'arbustes provenant de plantes adventices forestières et de jeunes arbres. Cette classe comprend les coupes à blanc récentes, les petites trouées forestières où le couvert arboré est absent, ainsi que la végétation rudérale des zones récemment perturbées.
- **Water** (alias *permanent water bodies*) : masses d'eau permanentes.
- **Bare** (alias *bare soils*) : sols nus ou peu couverts de végétation (< 15%), principalement des carrières en Wallonie.
- **Artif** (alias *artificialised surface and building*) : surface du sol recouverte de surfaces imperméables artificielles (par ex. béton ou bitume) et de bâtiments. Cette classe comprend les routes, les parkings, les ponts, les maisons et autres bâtiments.



 # Hauteur de végétation

Proportion d'occupation de chacune des classes suivantes pour chaque polygone.

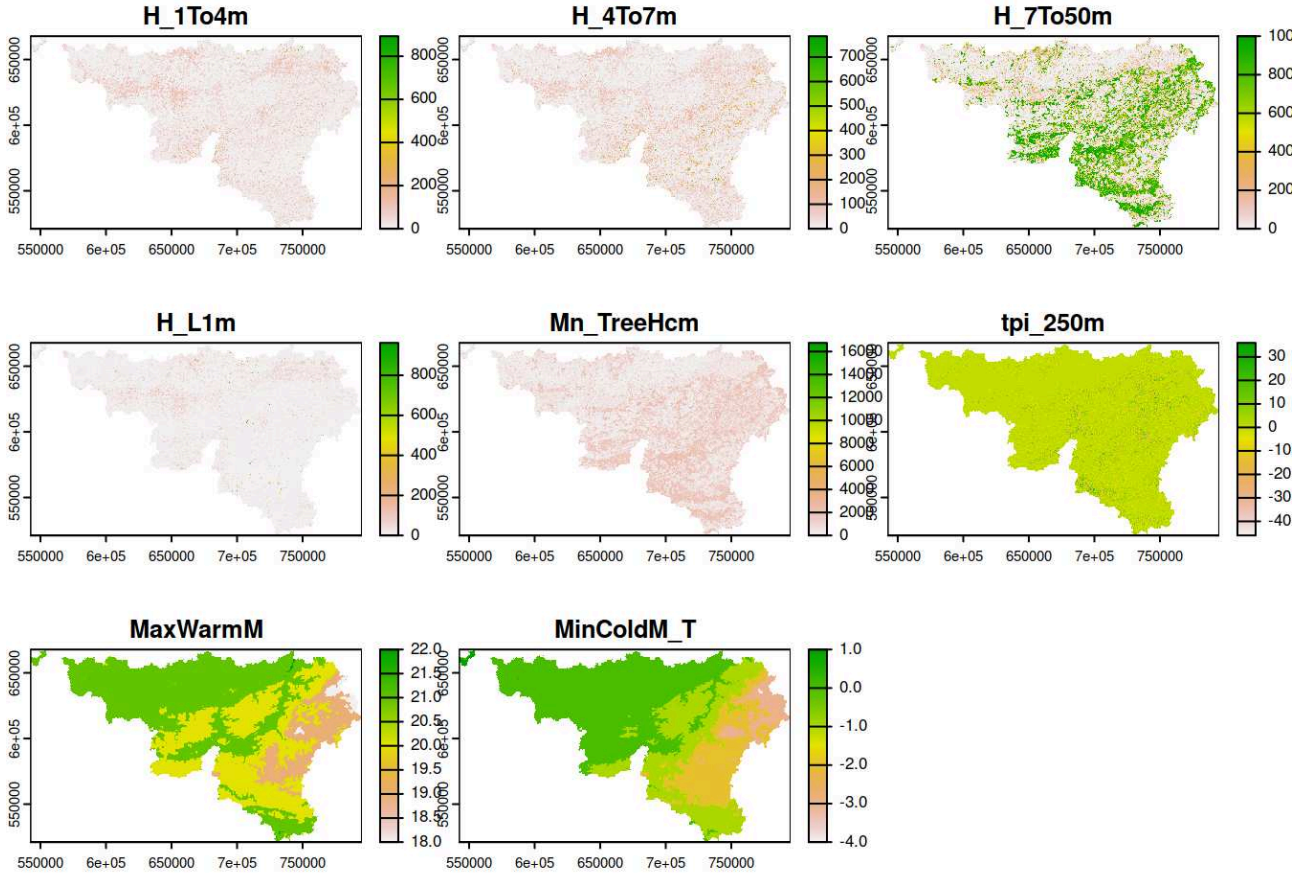
- **H_L1m** : végétation de plus de 25 cm et de moins d'1 m.
- **H_1To4m** : végétation de 1 à 4 m (buissons)
- **H_4To7m** : végétation de 4 à 7 m (arbustes)
- **H_7To50m** : végétation de 7 à 50 m (arbres)
- **Mn_TreeHcm** : hauteur moyenne de la canopée en cm

 # Indice de position topographique

- **tpi_250m** : moyenne de la position (élévation) relative par rapport à un environnement de 250 m depuis ce point (valeur négative dans un fond de vallée, valeur positive sur une colline).

 # Variables climatiques

- **MaxWarmM** : température maximale du mois le plus chaud
- **MinColdM_T** : température minimale du mois le plus froid

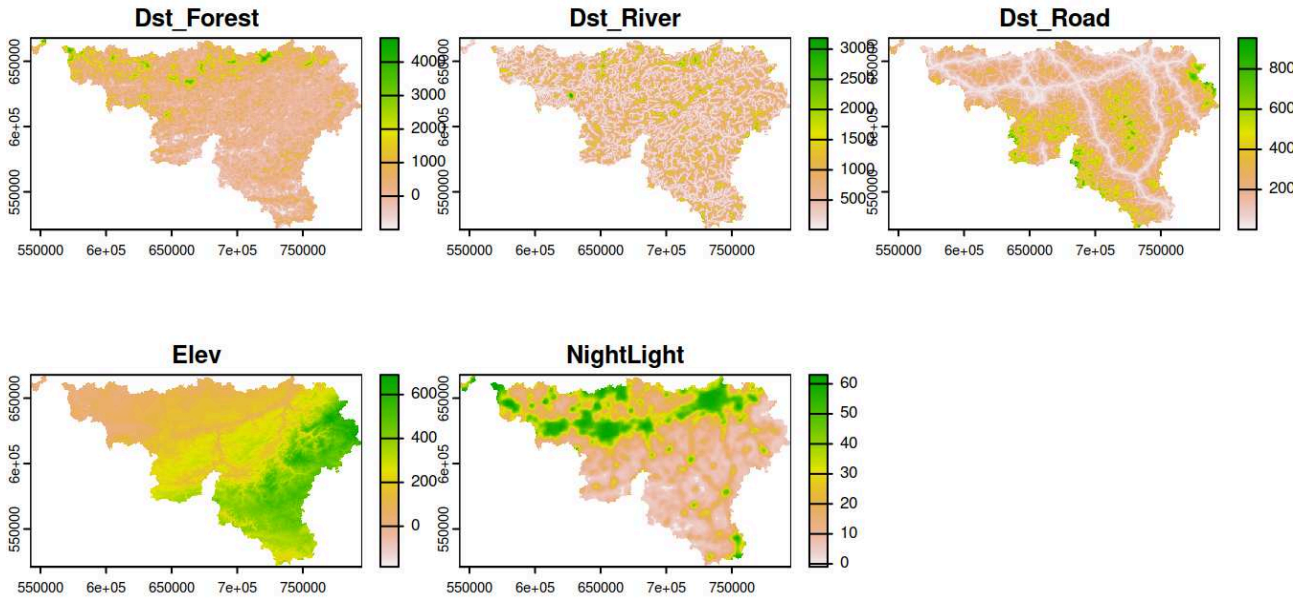


 # Distance aux éléments structurants du paysage

- **Dst_Forest** : distance moyenne aux massifs forestiers (valeur négative lorsque le point est localisé dans un massif forestier !).
- **Dst_River** : distance moyenne par rapport aux cours d'eau, pondérée selon la catégorie.
- **Dst_Road** : distance moyenne par rapport au réseau routier, pondérée selon la catégorie.

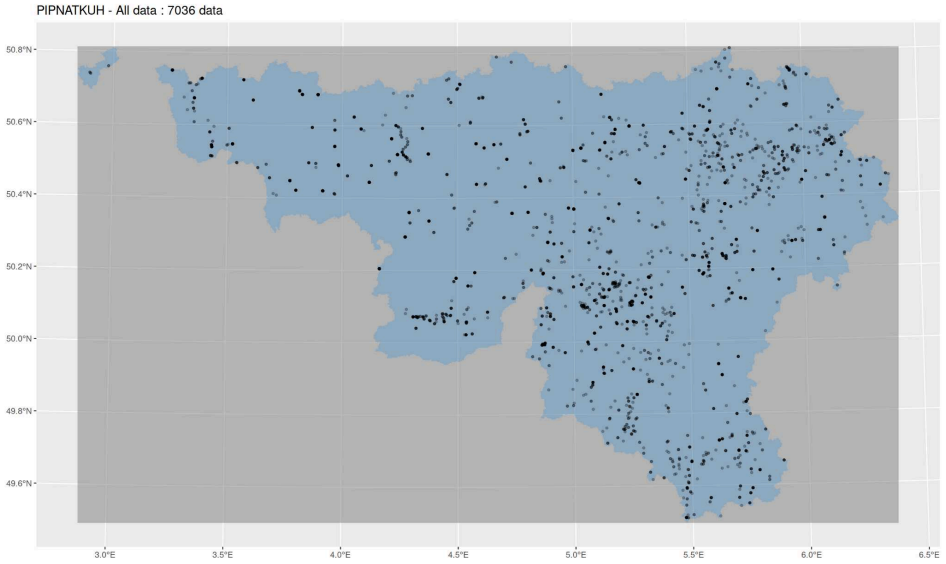
 # Autres facteurs

- **Elev** : élévation moyenne par rapport au niveau de la mer
- **NightLight** : intensité de luminosité nocturne mesurée par le programme de satellites météorologiques de la défense (DMSF)



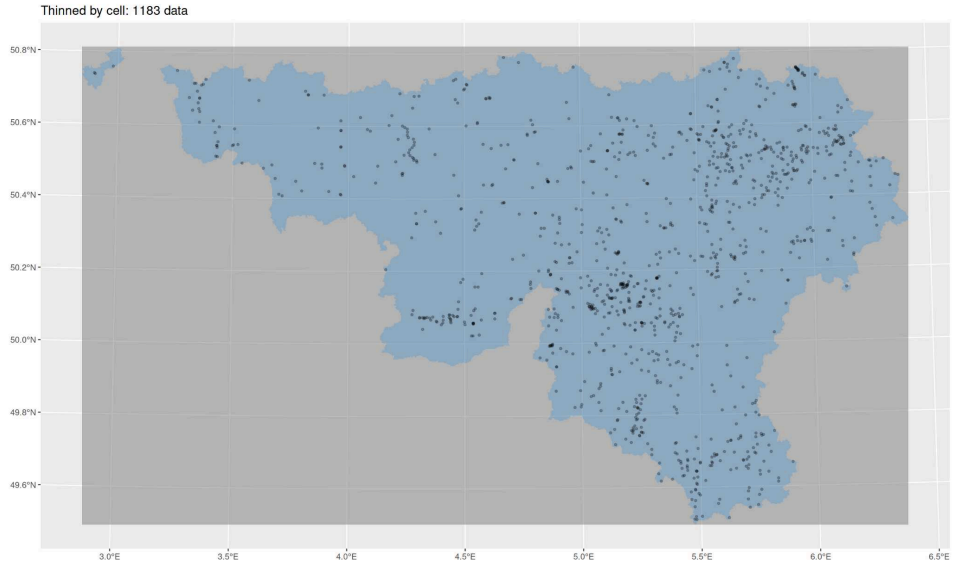
3.2 Chargement des données d'observation

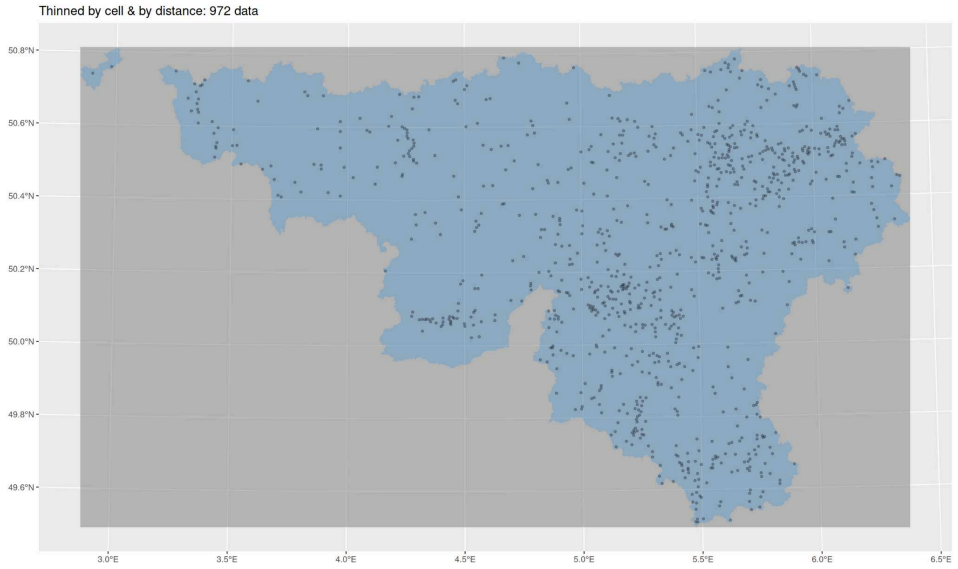
On filtre le taxon pipnat/kuh, nommé PIPNAT ou *P. nathusii* par facilité dans ce rapport, dans la base de données acoustique DEMNA-CSD-Sertius pour obtenir 7036 données.



3.3 Affinage des données

Deux opérations successives visent à réduire le biais d'autocorrélation spatiale en prélevant de manière aléatoire des données réparties au mieux sur le territoire wallon. La fonction [*thin_by_cell*] réduit les points à une seule observation par cellule du raster, puis [*thin_by_dist*] supprime les points à moins d'une distance seuil définie, ici 500 m. Pour cette espèce, les 7036 données sont réduites à 1183 après sélection d'une donnée par cellule, puis encore réduites à 973 après exclusions de points trop proches les uns des autres.

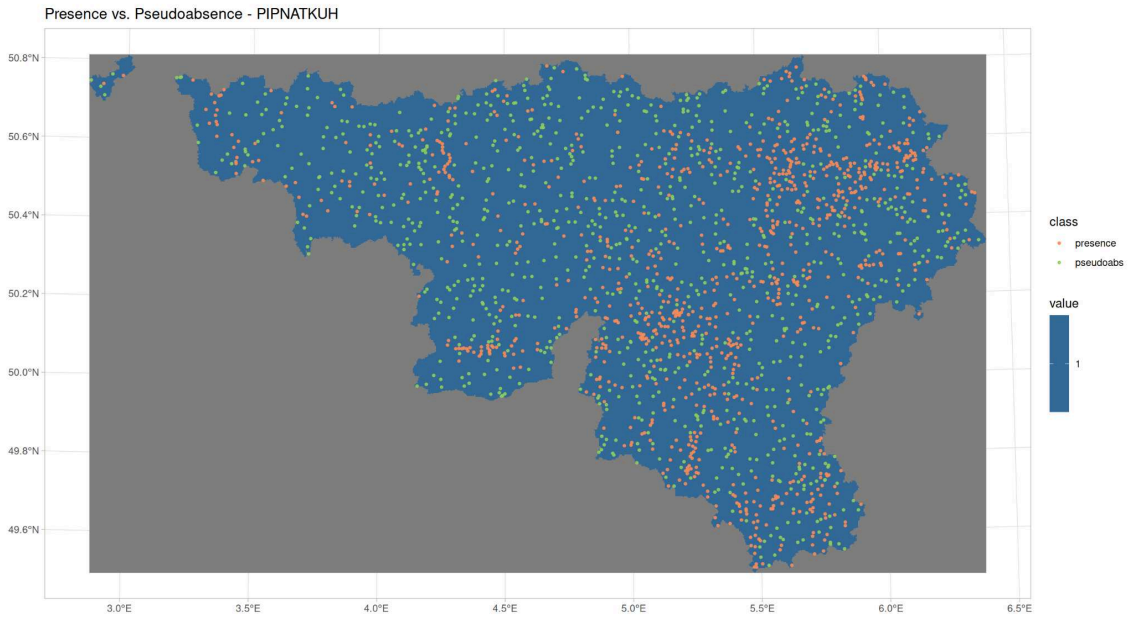




3.4 Génération des pseudoabsences

On génère des pseudoabsences de manière aléatoire, à savoir des localisations X-Y où l'espèce n'a pas été observée à ce jour, mais à au moins 1 km des points de présence. Ces "supposées absences" seront utilisées pour déduire quels facteurs environnementaux (et avec quelle intensité) permettraient d'expliquer la distribution réellement observée pour la généraliser au territoire wallon.
On génère autant de données de pseudoabsences que de données de présence réelles après filtration : 972 (présences) + 972 (pseudoabsences).

Dans le cas des vraies absences, on a appliqué la procédure d'affinage par cellule et par distance qui vient d'être décrite aux points d'inventaire où *P. nathusii* n'a pas été observée pour obtenir 972 présences et 528 absences.



3.5 Extraction des valeurs des facteurs environnementaux

Pour chaque point du jeu de données constitué, on extrait les valeurs des facteurs environnementaux avec *terra::extract*.

Une brève analyse descriptive permet de sélectionner *a priori* les facteurs environnementaux à retenir pour la modélisation. D'abord, un indice du pourcentage de non-recouvrement de la distribution de chaque facteur pour les présences avec les pseudoabsences est calculé :

```
BATS_Pseudo %>% dist_pres_vs_bg(class) %>% round(., 2)
## Water       NeedLV        NeedlSm       Plowed        NOpen         COpen         BroadLV
## 0.65        0.49          0.39          0.36          0.34          0.29          0.27
## Dst_Forest  NBroadSm      H_1To4m       tpi_250m      H_4To7m       H_7To50m      H_L1m
## 0.26        0.23          0.22          0.21          0.21          0.21          0.19
## Mn_TreeHcm  MinColdM_T     WOpen         Artif         Dst_River     Elev          MGramin
## 0.18        0.17          0.17          0.15          0.12          0.12          0.11
## NightLight  Bare          MaxWarmM      Dst_Road
## 0.08        0.05          0.03          0.03
```

Dans le but de retenir les facteurs séparant au mieux les présences et les pseudoabsences, on ne garde que ceux dont l'indice est supérieur à 30% (0.3). Cette valeur est celle utilisée par les auteurs du package *tidysdm* (Leonardi et al. 2023). On a testé différentes valeurs et 30% semble être un bon compromis pour sélectionner des facteurs explicatifs pertinents. Toutefois, pour les analyses de présence vs. vraies absences, on a abaissé ce seuil à 15% voire 10% pour retenir suffisamment de facteurs[3].

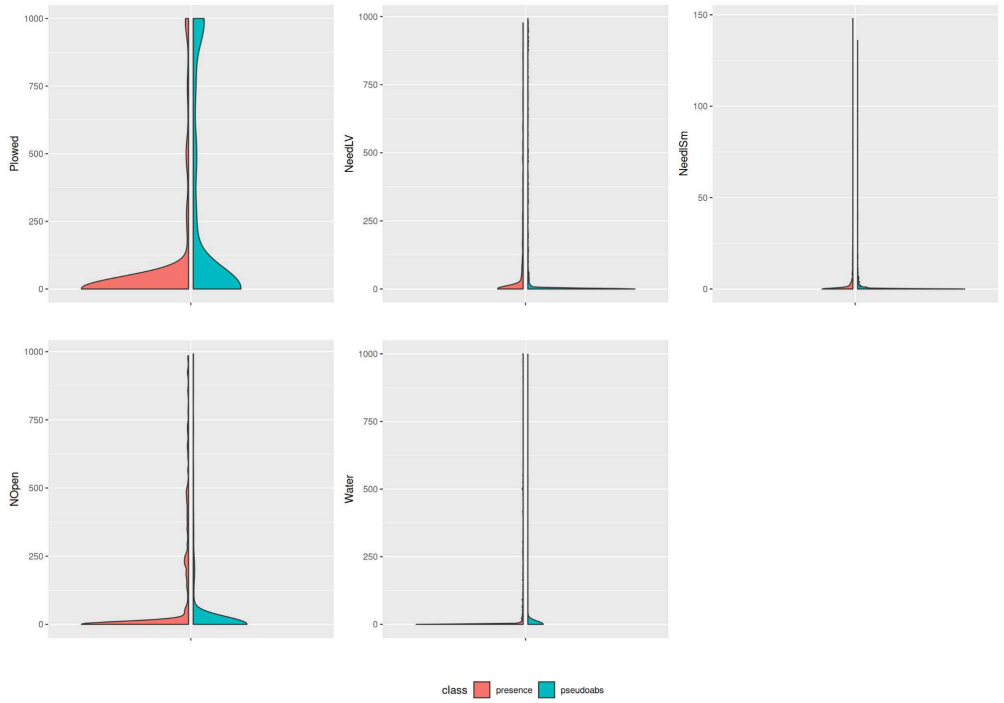
```
vars_to_keep <- BATS_Pseudo %>% dist_pres_vs_bg(class)
names(vars_to_keep[vars_to_keep > 0.3])
## [1] "Water"   "NeedLV"  "NeedlSm" "Plowed"  "NOpen"
```

On mesure ensuite le coefficient de corrélation entre toutes les combinaisons des facteurs sélectionnés (5 dans ce cas-ci) et on décide d'éliminer ceux qui seraient corrélés à plus de 70% (Leonardi et al. 2023) pour réduire le biais de colinéarité entre les facteurs, ce qui est une condition d'application de certains modèles (*glm* et *gam* notamment). Pour les données de *P. nathusii,* aucun facteur n'a dû être éliminé pour cette raison.

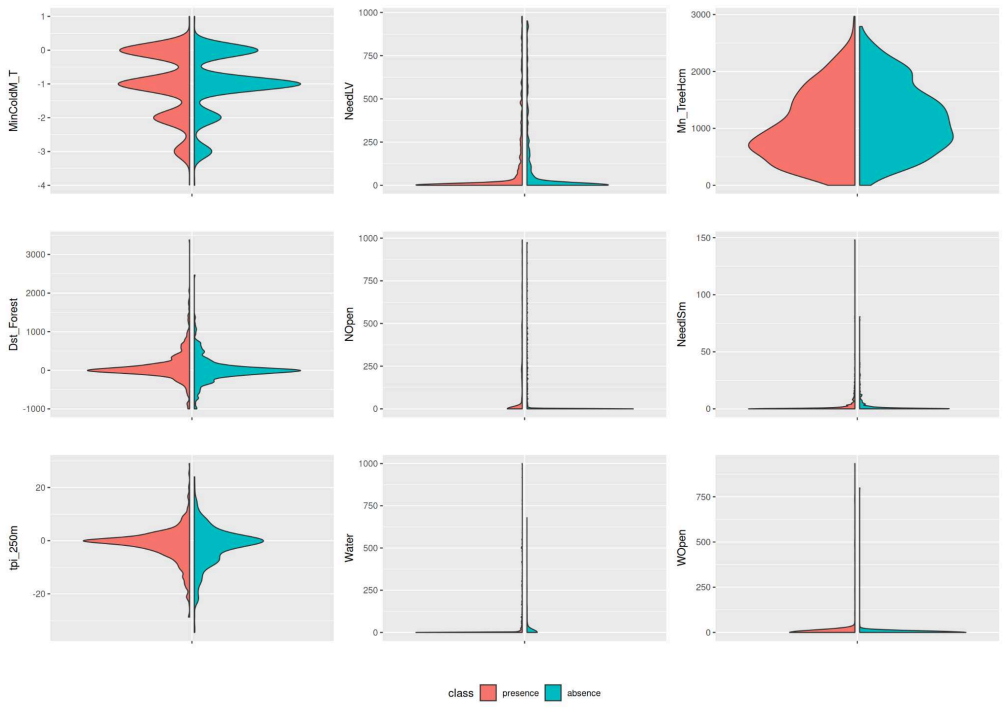
[3] Les effectifs des absences étant basés sur les points d'inventaires affinés, ils sont souvent inférieurs à ceux des pseudoabsences. La conséquence de cette sous-représentation est de réduire la variabilité des valeurs des facteurs environnementaux et donc la probabilité de contraste avec les présences.

Les facteurs finalement retenus pour la méthode des **pseudoabsences** dans le cas de *P. nathusii* sont **Plowed, NeedLV, NeedlSm, NOpen** et **Water** (le nombre et la nature des facteurs retenus est différent pour chaque espèce, ainsi que selon que l'on travaille en pseudoabsences ou en vraies absences)**.**

La distribution des facteurs retenus pour les deux classes (présence vs. pseudoabsence) sont illustrés par des graphiques "en violon" uniquement pour donner une idée de leur représentation dans le jeu de données.



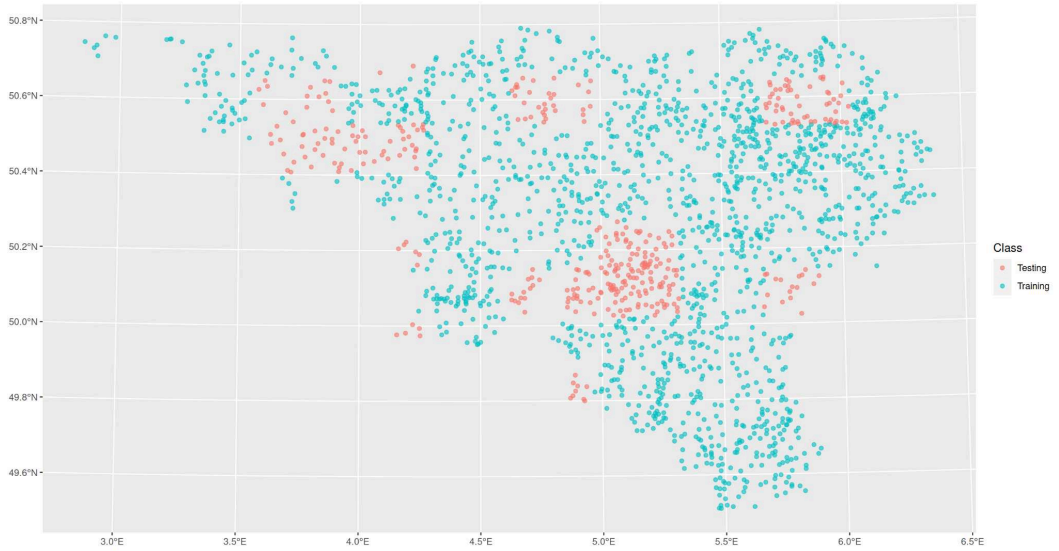
Les facteurs finalement retenus pour *P. nathusii* avec la méthode des **vraies absences** sont **MinColdM_T, NeedLV, Mn_TreeHcm, Dst_Forest, NOpen, NeedlSm, tpi_250m, Water et WOpen**.



3.6 Ajustement (*fitting*) des modèles et réglage (*tuning*) des hyperparamètres

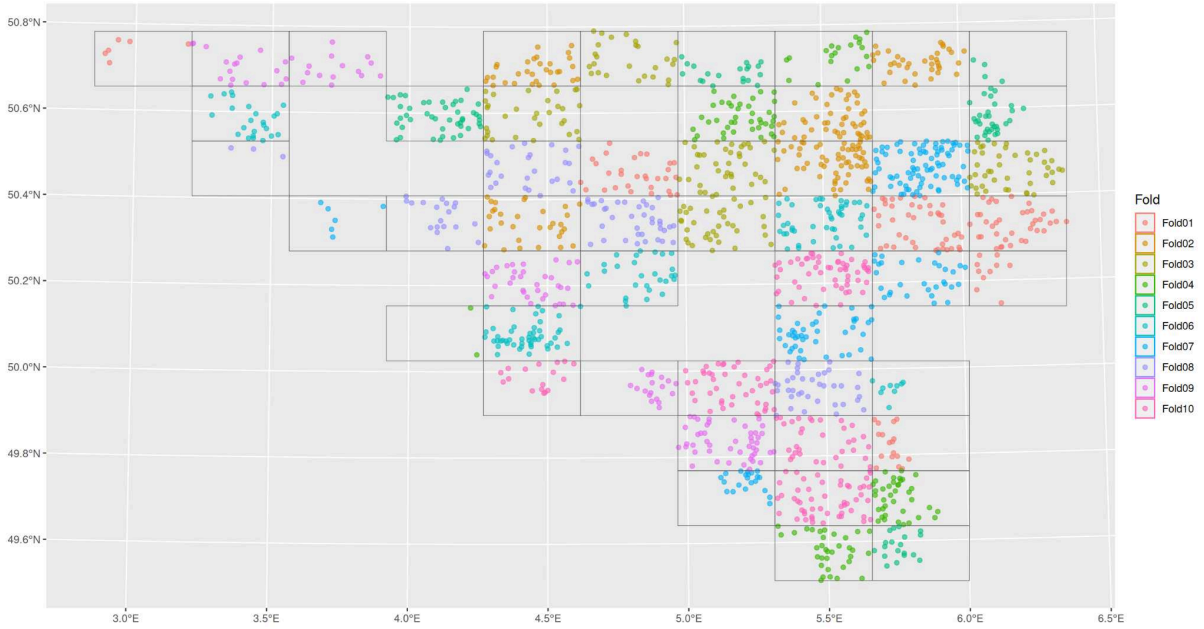
3.6.1 Split du jeu de données en *training:test* (80%:20%)

On isole 20% de l'ensemble du jeu de données présences + (pseudo)absences sélectionnés aléatoirement par blocs (maille géographique).



3.6.2 Création des schémas de validation croisée pour les données training

Des blocs de données sont déterminés au sein des données d'entraînement pour l'ajustement des modèles et leur évaluation en comparant à 10 reprises 10% vs. 90% des points, c'est-à-dire en créant 10 "*plis*" (*fold* en anglais) de validations croisées dans le jeu de données initial.



3.6.3 Création de la recette

Une "*recette*" (fonction recipes::recipe()) prépare les données pour les utiliser dans le workflow d'ajustement des modèles. Dans notre cas, cette formulation consiste simplement à identifier la variable de

réponse (colonne “class”), dont les deux niveaux possibles sont “presence” ou “pseudoabsence” et que l’analyse va consister à entraîner le modèle pour prédire les conditions d’obtenir le niveau de référence (“présence”).

3.6.4 Définition des modèles

C’est ici que l’on choisit quels modèles seront ajustés et évalués. Nous avons opté pour les 4 modèles suivants :

- **glm: Generalized Linear Model** sdm_spec_glm()
  La régression logistique en modèle linéaire généralisé est un modèle de machine learning très communément utilisé en statistique. À la différence des 3 modèles suivants, il n’y a pas d’hyperparamètre à déterminer pour ce modèle.

- **rf: Random Forest** sdm_spec_rf()
  Le modèle random forest avait été pressenti pour être utilisé lors de ces analyses, pour sa robustesse et parce que c’est probablement le modèle à arbres décisionnels le plus utilisé, à la fois pour des outcomes catégoriques et quantitatifs. Un seul hyperparamètre doit être déterminé pour les SDM: *mtry.*

- **gbm: Boosted Tree model** sdm_spec_boost_tree()
  Ce modèle est assez similaire au précédent par son fonctionnement par arbres décisionnels, son algorithme est le *xgboost* et 6 hyperparamètres sont à déterminer dans le cas des SDM: *mtry*, *trees*, *tree_depth*, *learn_rate*, *loss_reduction* et *stop_iter*.

- **maxent: MaxEnt - Maximum of Entropy** sdm_spec_maxent()
  Le modèle de Maximum of Entropy était pressenti en premier lieu pour ce projet de SDM. Il est spécifiquement conçu pour analyser des données de présence seules. Malgré sa réputation d’outil “simple à utiliser”, ses algorithmes sont basés sur la détermination d’hyperparamètres très complexes et plutôt obscurs. Son intégration uniformisée avec les autres modèles dans *tidysdm* permet d’en ajuster les hyperparamètres automatiquement.

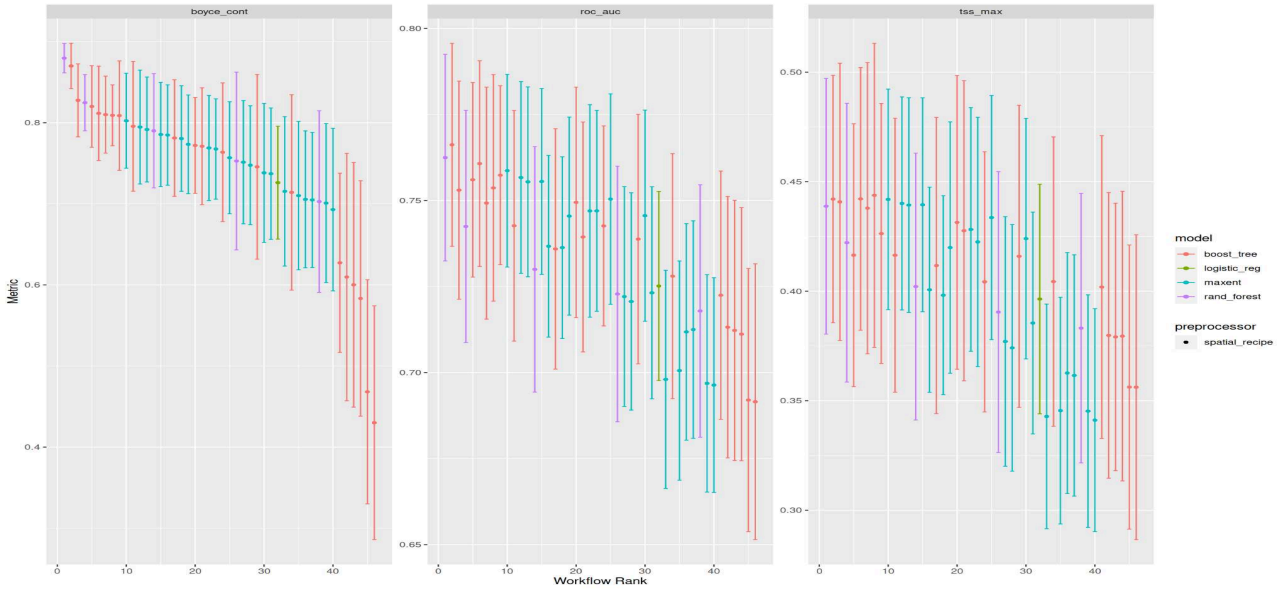
On a décidé d’utiliser ces 4 modèles parce que ce choix est proposé par Leonardi et al. pour ce type de modélisation. Cette approche ouvre la possibilité de profiter des forces de chacun des modèles séparément en retenant les (hyper)paramétrages expliquant au mieux la distribution des espèces sur base des facteurs environnementaux sélectionnés. Les deux modèles à arbres décisionnels présentent des caractéristiques qui les rendent complémentaires. Alors que Random Forest crée ses arbres en parallèle, le Boosted Tree Model le fait de manière séquentielle, augmentant ainsi les chances qu’un des deux s’avère le plus adéquat selon l’espèce modélisée. Les hyperparamètres déterminés sont listés ci-dessus sous forme des codes utilisés par les fonctions. La signification de ces codes est :

- ➔ *mtry* : le nombre de facteurs prédictifs échantillonnés aléatoirement à chaque ‘*split*’, c’est-à-dire chaque confrontation d’un bloc de données avec les autres blocs
- ➔ *trees* : nombre d’arbres créés dans l’ensemble
- ➔ *tree_depth* : nombre maximal de ‘*splits*’
- ➔ *learn_rate* : taux d’adaptation de l’algorithme à chaque itération
- ➔ *loss_reduction* : nombre de données exposées à la routine d’ajustement
- ➔ *stop_iter* : le nombre d’itérations au-delà duquel l’absence d’amélioration du modèle interrompt la boucle de calculs

3.6.5 Réglage des modèles (*model tuning*)

Le *workflowset* contenant les 4 modèles choisis va maintenant effectuer des essais itératifs pour déterminer les meilleurs hyperparamètres par validations croisées des blocs qui ont été déterminés précédemment sur les données d’entraînement. Dans le cas du *glm* le modèle est ajusté en une fois (pas d’itérations) puisqu’il

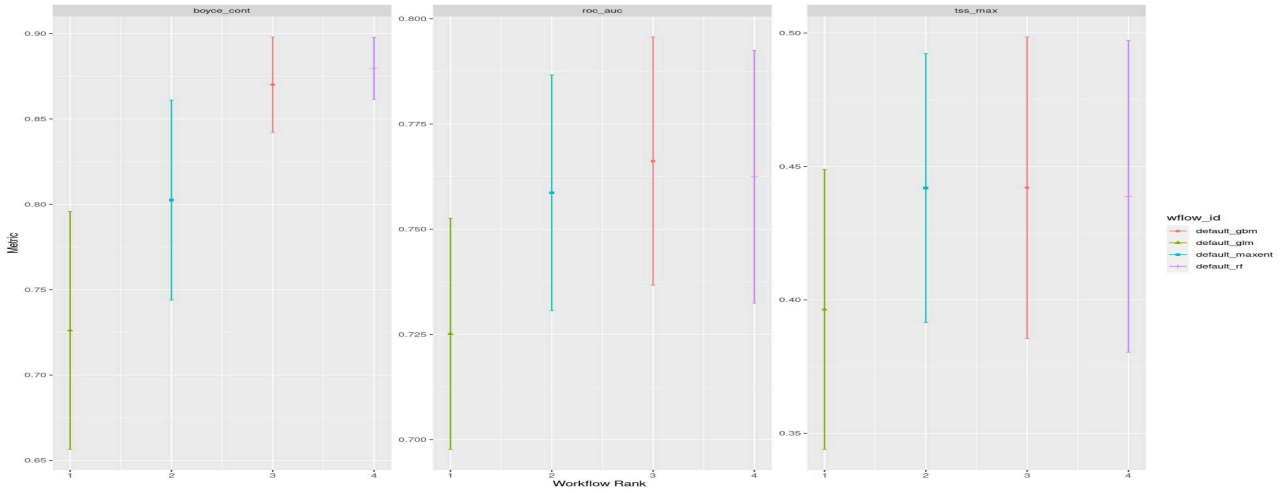
n'a pas d'hyperparamètres. Pour les 3 autres modèles les itérations s'arrêtent lorsque le modèle a atteint une valeur optimale. Les métriques (cf. explications au point 4.1) obtenues pour chaque version des modèles sont représentées sur le graphique suivant (à gauche pour le *Boyce Continuous index*), triées par ordre décroissant. Cette représentation nous donne une idée de la fourchette des valeurs et du nombre de versions testées des modèles.



Représentation des métriques obtenues pour les différents modèles : Boyce continuous index (← Boyce_cont), Area Under the Receiver Operator Curve (↑ ROC_AUC) et Maximum of the True Skills Statistics (→ TSS_MAX)

3.6.6 Création d'un ensemble

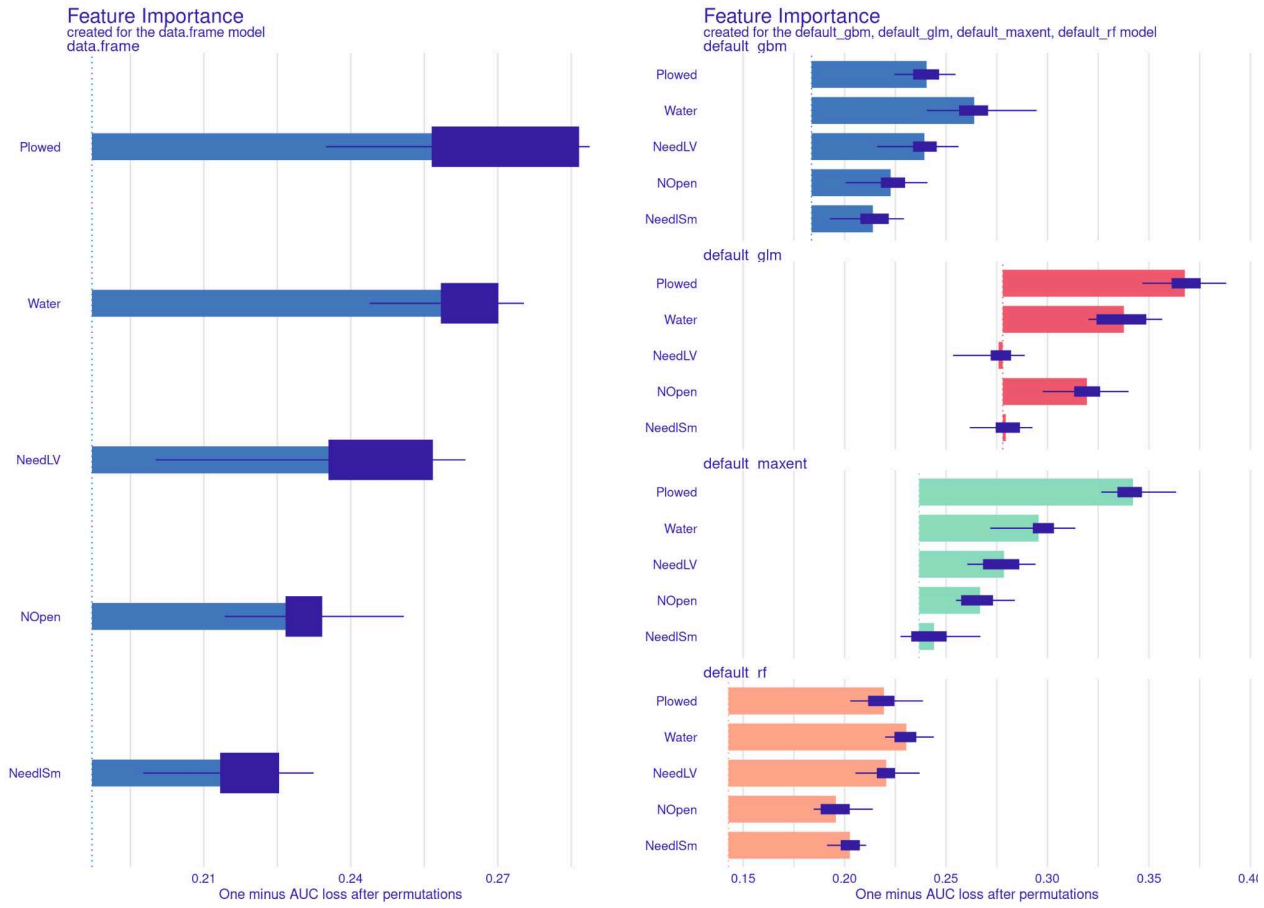
La meilleure version de chacun des 4 modèles est ensuite conservée dans un '*ensemble*'. Ces 4 meilleures métriques sont illustrées graphiquement ci-dessous. L'*ensemble* pourrait déjà être utilisé tel quel pour générer les prédictions de SDM, cependant on va encore l'améliorer en confrontant les prédictions calculées sur le lot de 20% de données test mises de côté au début de la procédure (voir "empilement des modèles").



3.6.7 Exploration de l'importance des facteurs environnementaux

Grâce à une fonction dédiée de *tidysdm* (explain_tidysdm()), on peut déterminer l'importance de chaque facteur environnemental, soit pour *l'ensemble* soit pour chacun des 4 modèles qui le constitue. Cette information est importante pour notre compréhension du lien entre les facteurs de l'environnement et la probabilité de détection de l'espèce qui sera calculée en résultat final.

Certains facteurs avaient été retenus en première étape parce qu'ils semblaient différencier les points de présence des points de pseudoabsence (ou de vraie absence) mais peuvent très bien s'avérer finalement très peu informatifs sur la structuration des données. On voit sur le graphique suivant que le caractère agricole (pourcentage de labour - *Plowed*) et la surface d'eau libre (*Water*) d'un site sont de loin plus influents que la proportion de la strate arbustive en résineux (*NeedlSm* et *NeedLV*) pour expliquer la présence de *P. nathusii*.



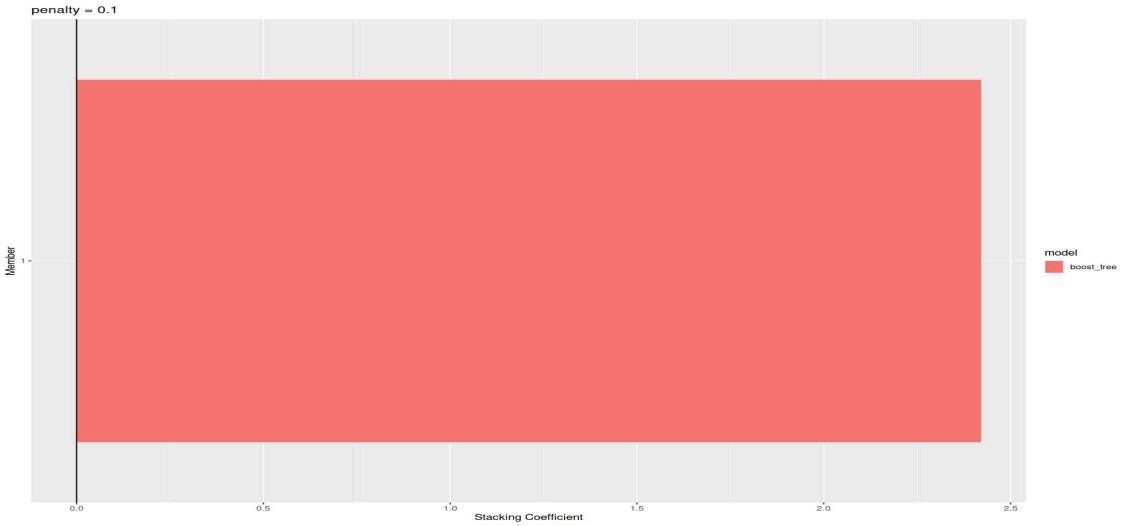
On peut aussi tracer graphiquement comment évolue la prédiction de présence de l'espèce en fonction de chaque paramètre, pris indépendamment. Les courbes de prédiction en fonction des 5 facteurs environnementaux permettent d'interpréter visuellement que l'espèce *P. nathusii* plus souvent rencontrée lorsqu'un site comprend plus de 25% (càd 250 ‰ sur les graphes) de couverture d'eau libre (*Water*), que la proportion de surface labourée (*Plowed*) tend vers 0‰ et qu'il y a moins de 25% (250 ‰) de grands résineux.

3.7 Création du modèle final (modèle empilé)

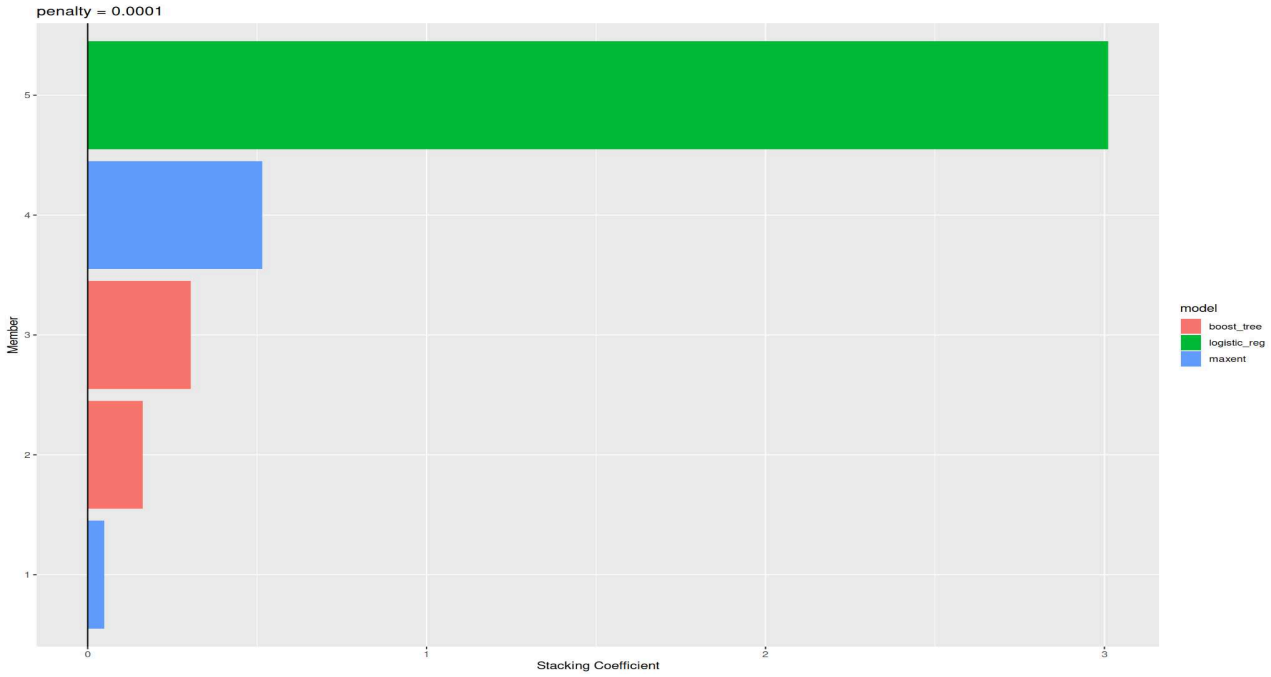
L'ensemble obtenu peut maintenant être amélioré en créant un "empilement" (*stack*) sur base d'une procédure de meta-learning permettant d'apprendre comment combiner et pondérer au mieux des modèles multiples, notamment en incluant différentes versions du même algorithme avec différents hyper-paramètres. Un ou plusieurs modèles sont "empilés" et pondérés dans le modèle final, qui peut être différent pour chaque espèce et même entre les deux méthodes que nous avons systématiquement comparées (pseudoabsences ou vraies absences).

Pour *P. nathusii*, c'est uniquement le *boosted tree model* qui est retenu avec la méthode des pseudoabsences alors que 5 (versions de) modèles sont retenus lorsqu'on utilise les vraies absences.

Présences vs. pseudoabsences



Présences vs. vraies absences



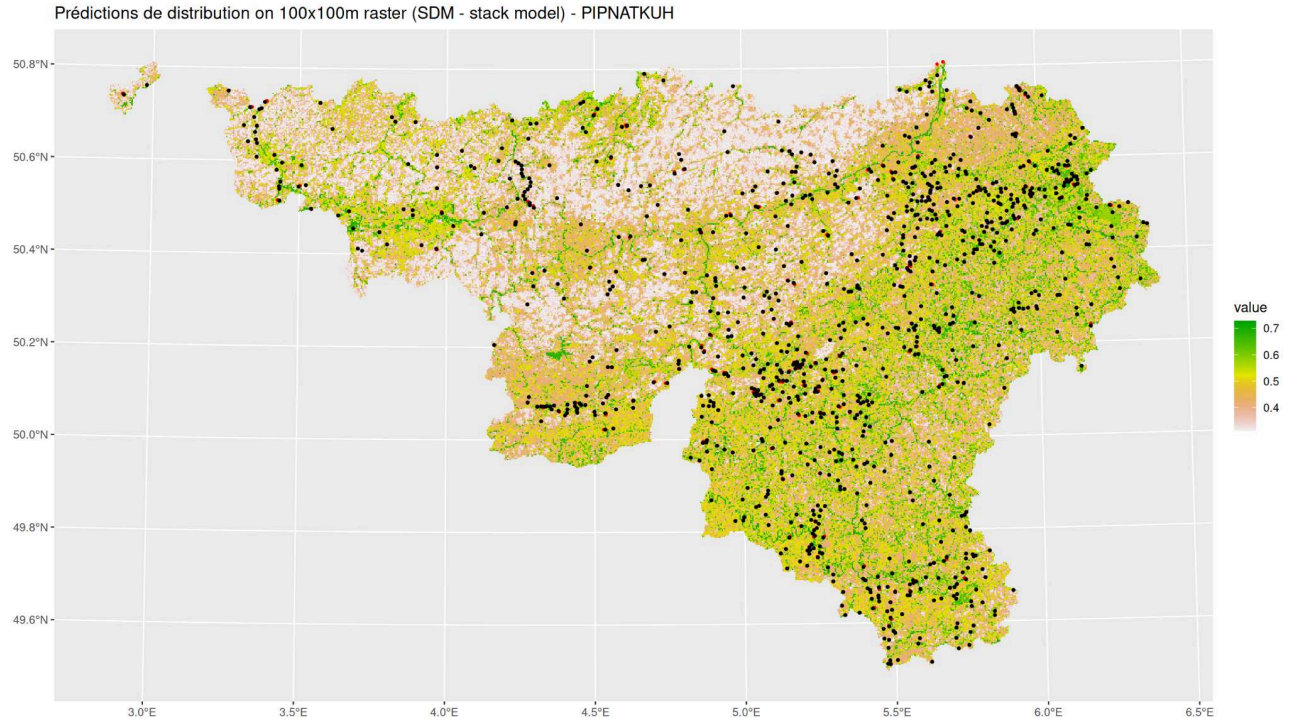
Confrontation des données de test aux prédictions du modèle empilé

Le modèle peut enfin être évalué sur les données test pour calculer les métriques de sa valeur prédictive avant de produire une prédiction finale. Ces métriques sont récapitulées, pour chaque espèce et pour les deux méthodes comparées (présences vs. pseudoabsences ou vraies absences).

.metric	.estimator	.estimate
<chr>	<chr>	<dbl>
boyce_cont	binary	0.9065239
roc_auc	binary	0.7178131
tss_max	binary	0.3659467

3.8 Prédictions du modèle

On utilise la fonction predict_raster() pour prédire une probabilité de détection de l'espèce uniformément sur l'ensemble de la Région wallonne. Le modèle est appliqué aux rasters des facteurs environnementaux sélectionnés au début pour attribuer une valeur à chaque pixel du raster de distribution de l'espèce. La carte ci-dessous présente ce résultat où la probabilité prédite est symbolisée par un gradient de couleur. Les points noirs sont les points de présence utilisés dans l'analyse et les points rouges[4] sont les présences qui ont été éliminées lors de l'affinage initial.



4. Synthèse des résultats, discussion, perspectives

4.1 Discussion des résultats - Commentaire analytique sur les résultats pour les espèces ciblées

4.1.1 Commentaires généraux

Choix des facteurs environnementaux

Parmi les nombreux facteurs potentiellement utilisables pour une analyse de modélisation, nous en avons extrait 25 de la couche Ecotope (cf. point 3.1). Pourtant, plusieurs autres ressources avaient été identifiées à l'entame de la mission. Les raisons de ce choix sont les suivantes :

- Priorité de l'usage de la couche Ecotope car elle agglomère déjà la majorité des ressources récentes de description des habitats, la topologie, la pédologie, la hauteur de la végétation, etc. Ces facteurs sont déjà pré-traités et sont exprimés sous une forme directement utilisable dans les analyses de modélisation.
- Les ressources telles que Forestimator et le parcellaire agricole ne représentent qu'une partie du territoire, respectivement forestière et agricole, or les raster descripteurs environnementaux doivent idéalement ne pas contenir de valeurs nulles pour que l'entraînement des modèles

[4] On en distingue peu sur la carte à cette échelle puisqu'il s'agit des points qui ont été éliminés lors de l'affinage à cause de leur proximité avec d'autres données. Ils sont donc masqués par les points noirs sur cette carte.

fonctionne. En plus de compléter les valeurs mannquantes, il aurait fallu aussi ré-échantillonner le raster pour obtenir la même maille que pour ceux provenant d'Ecoptope.

- Le MNT et une information sur la pollution lumineuse sont déjà intégrées dans les données Ecotope. Il serait probablement intéressant de diversifier les sources, peut-être plus à jour que ce qui se trouve dans Ecotope, mais ici à nouveau, on a considéré avoir à disposition une source de données satisfaisante.
- Le jeu de données Bioclim a aussi été investigué. Ecotope intègre déjà les variables Bioclim au travers des différents facteurs descriptifs stationnels

Pseudoabsences et vraies absences

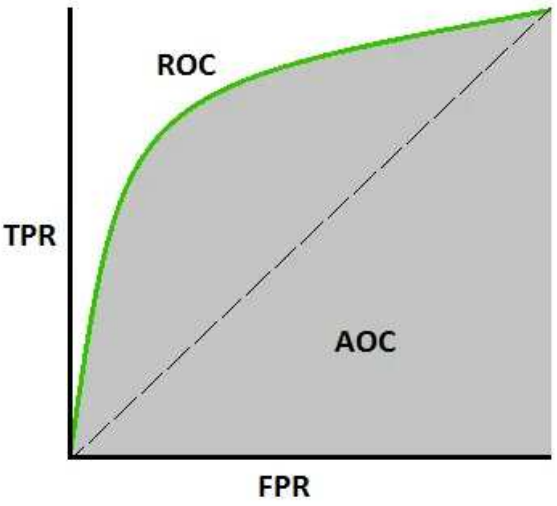
Pour construire un modèle de distribution, il faut confronter les facteurs environnementaux qui décrivent la présence et l'absence de l'espèce étudiée. Le plus souvent, on ne possède que des données de présence issues d'observations principalement fortuites ou parfois plus ou moins planifiées, mais jamais de protocoles standardisés de recherche sur un large territoire. Pour cette raison, les procédures de modélisation de distribution font appel à des pseudoabsences, des lieux tirés au sort aléatoirement où on postule que l'espèce est probablement absente. Le risque est que les points de pseudoabsence soient en réalité des habitats / régions favorables à l'espèce mais où elle n'a simplement pas été cherchée! D'autre part, l'hypothèse selon laquelle les points d'inventaires où l'espèce étudiée n'a pas été contactée constitue de vraies absences est aussi biaisée puisque les observateurs orientent souvent leurs recherches vers des habitats favorables (par exemple les milieux humides et cours d'eau, les zones à faible pollution lumineuse, …). Dans un cas comme dans l'autre l'absence est toujours entachée d'une incertitude due au postulat qu'on pose.

Les "pseudoabsences" et les "vraies absences" sont donc probablement biaisées et relativisent l'interprétation des résultats. Une solution à cette part de doute pourrait être de combiner les deux résultats en calculant une moyenne des deux rasters résultant de chaque modélisation. Nous avons produit ce résultat de compromis et ajouté aux résultats délivrés.

Définition et comparaison des métriques utilisées dans les analyses

Boyce continuous index (Boyce_cont) mesure l'adéquation d'un modèle. Il est calculé au moyen de fenêtres multiples se recouvrant et considéré particulièrement approprié pour les SDM basés sur des données de présence seule.

Area Under the [Receiver Operator] Curve (ROC_AUC) est une métrique très communément utilisée en Machine Learning, tant pour des données binaires que multiclasses. Elle représente la surface (grisé sur l'illustration suivante) contenue sous la courbe (verte). Cette 'Receiver Operator Curve' est la relation entre le taux de vrais positifs (TPR) au taux de faux positifs (FPR). Un modèle parfait serait théoriquement représenté par une courbe verte droite horizontale de valeur TPR = 1 (auquel cas la ROC_AUC vaut 1). Plus la ROC_AUC est grande (proche de 1), meilleur est le modèle.



Maximum of the True Skills Statistics (TSS_MAX) est une métrique appropriée à la modélisation de réponses binaires (présence-absence).

Le tableau suivant reprend une synthèse des métriques calculées. On y compare systématiquement les valeurs obtenues pour la méthode des pseudoabsences avec celles obtenues pour la méthode des vraies absences. À une exception près, c'est toujours la méthode des pseudoabsences qui obtient le meilleur résultat. Ceci est probablement dû au fait que dans ce cas on génère de nombreux points qui maximisent la variabilité des facteurs environnementaux alors que pour les vraies absences, il s'agit toujours des sites réellement inventoriés, qui répondent donc probablement, au moins en partie à certains a priori de l'observateur. Remarquons aussi que les métriques n'ont pas pu être calculées pour la pipistrelle commune, conséquence du faible nombre de points d'absence pour cette espèce extrêmement commune et largement répandue.

Espèce	Méthode	Boyce_cont	ROC_AUC	TSS_MAX
NYCNOC	**Prés./Pseudoabs.**	**0,826**	**0,812**	**0,571**
NYCNOC	Prés./Abs.	0,649	0,754	0,51
NYCLEI	**Prés./Pseudoabs.**	**0,96**	**0,81**	**0,49**
NYCLEI	Prés./Abs.	0,464	0,579	0,165
EPTSER	**Prés./Pseudoabs.**	**0,944**	**0,735**	**0,377**
EPTSER	Prés./Abs.	0,222	0,497	0,055
PIPPIP	**Prés./Pseudoabs.**	**0,976**	**0,721**	**0,326**
PIPPIP	Prés./Abs.[5]	-	-	-
PIPNAT	**Prés./Pseudoabs.**	**0,906**	**0,718**	**0,366**
PIPNAT	Prés./Abs.	0,887	0,673	0,28
MYOMYO	Prés./Pseudoabs.	0,796	**0,801**	**0,575**
MYOMYO	Prés./Abs.	**0,925**	0,702	0,318

Tableau des métriques obtenues pour les 6 espèces modélisées. Les deux méthodes sont comparées et la meilleure métrique est mise en évidence en gras. Voir texte pour les définitions des métriques.

4.1.2 Résultats des modélisations pour les 6 espèces

Les résultats des modélisations sont fournis sous forme de données brutes (un fichier geotif et un shapefile) en annexe numérique. Afin de faciliter l'interprétation de ces résultats, on détaille ci-dessous, espèce par espèce :

- le nombre de données utilisées pour les analyses
- la carte de localisation de ces données
- le graphique de l'importance des facteurs environnementaux pour les deux méthodes comparées
- les courbes de dépendance des facteurs environnementaux aux prédictions
- les cartes de probabilité de détection selon les deux méthodes utilisées
- la carte unifiant les deux méthodes

Conformément à la demande formulée dans le cahier des charges, on a créé une couche shapefile sur base du raster généré par la modélisation. Sachant que les rasters résultant des analyses nous donnent une valeur de probabilité de détection comprise entre 0 et 1, la déduction d'un polygone qui pourrait être interprété comme une enveloppe de distribution dépend d'un choix arbitraire d'un seuil définissant la

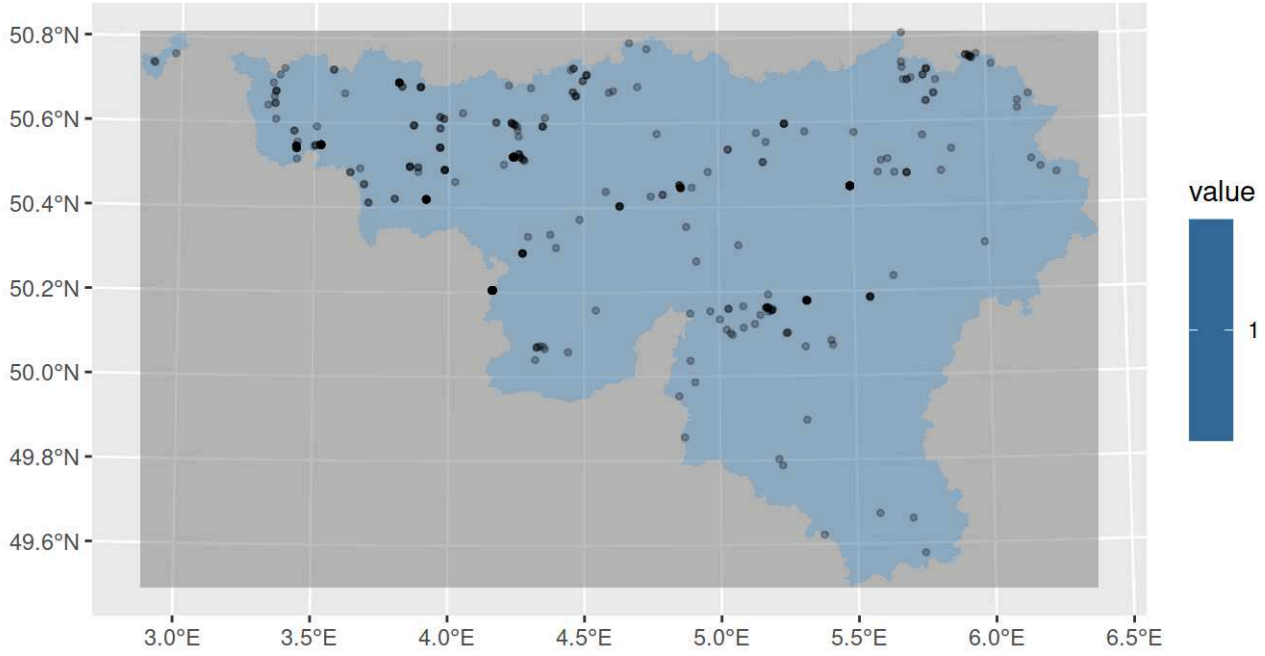
[5] En raison d'un très faible nombre de points d'absence, la modélisation n'a pas abouti et les métriques n'ont pas pu être calculées. Un raster de prédictions a toutefois pu être produit, mais sur base de l'ensemble non-empilé. Intuitivement, on comprend facilement que cette espèce, de loin la plus abondante dans la grande majorité des habitats, ne permette pas d'utiliser des données d'absence suffisamment nombreuses pour entraîner les modèles.

probabilité à laquelle on considère l'espèce présente ou absente. Ce seuil a été fixé à une valeur arbitraire de probabilité de 0.5. On regroupe donc dans un polygone “présent” tous les pixels du raster de valeur > 0.5 et un second polygone “absent” les pixels du raster de valeur < 0.5. Ce seuil totalement arbitraire pourrait être adapté, par exemple en se basant plutôt sur des valeurs quantiles (50 % ?) pour chaque espèce que sur une valeur arbitraire.

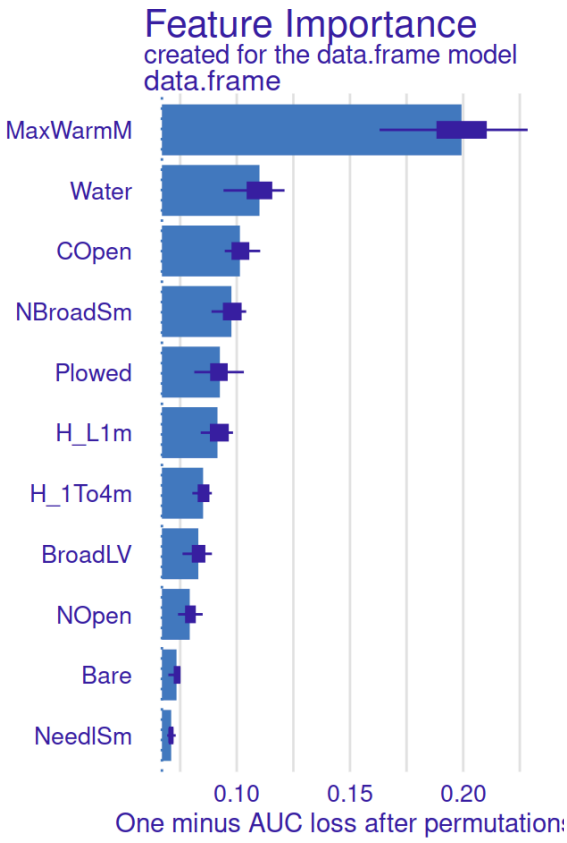
Noctule commune (*Nyctalus noctula*)

Nombre de données de présence **avant** affinage : 690

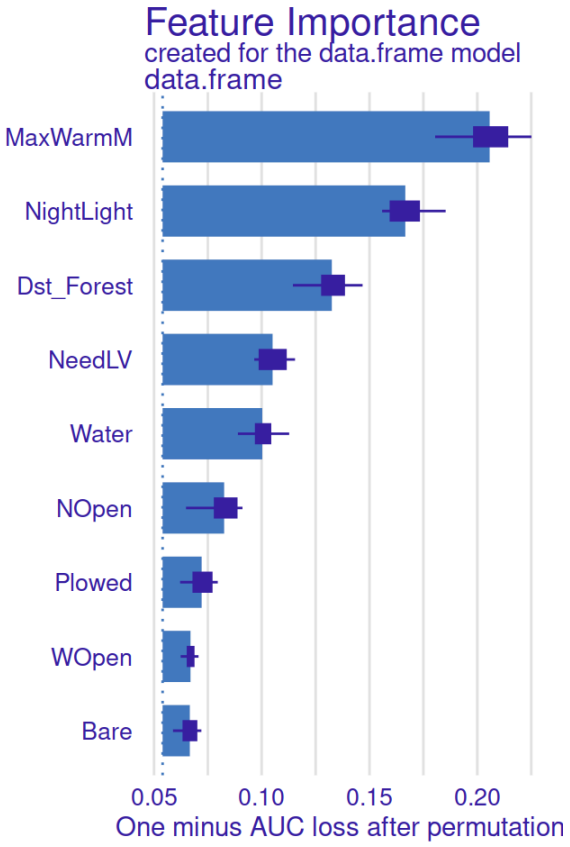
Nombre de données de présence **après** affinage : 162



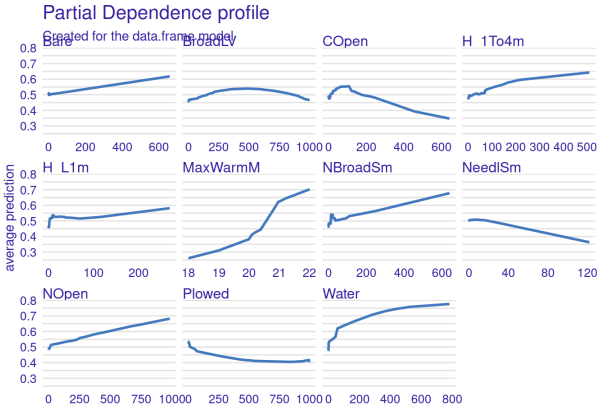
Méthode 1 : pseudoabsences



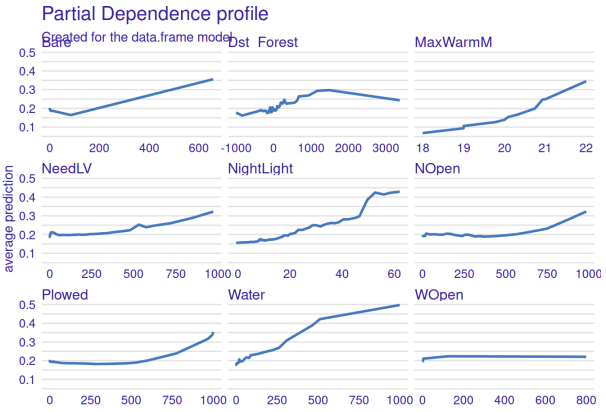
Méthode 2 : vraies absences



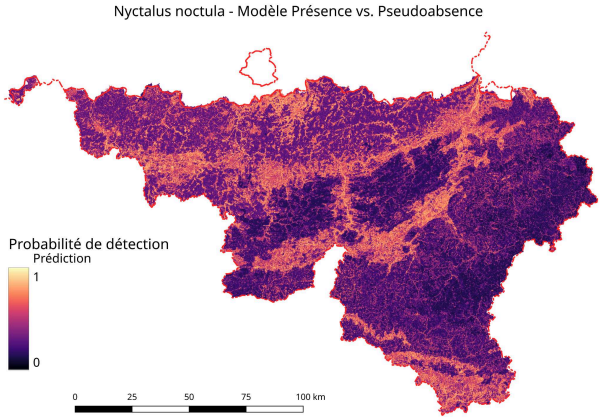
Méthode 1 : pseudoabsences



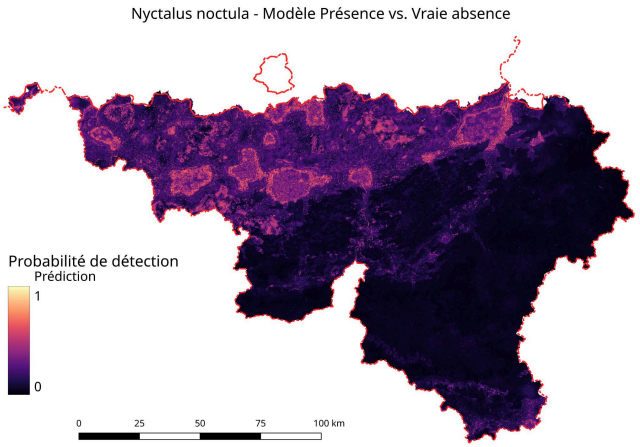
Méthode 2 : vraies absences



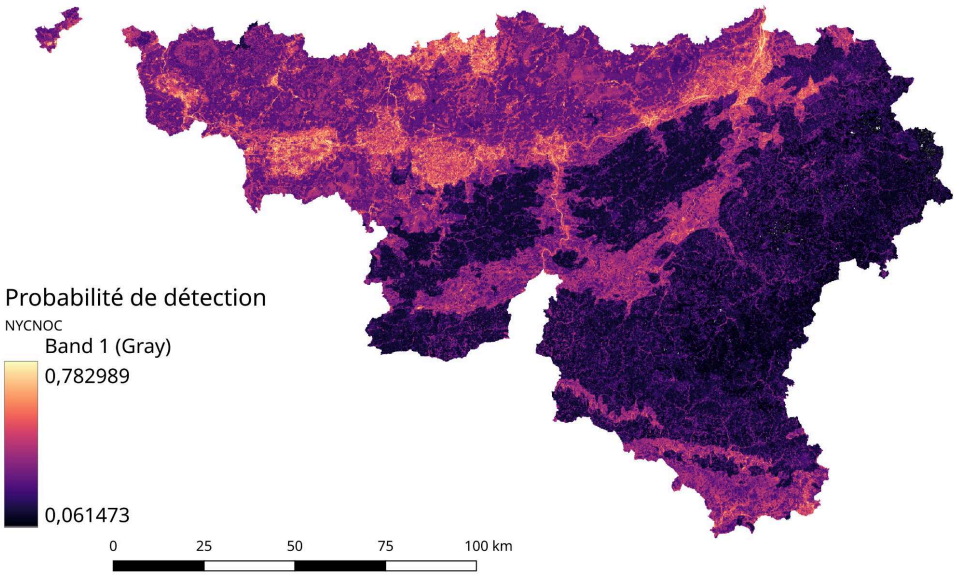
Méthode 1 : pseudoabsences



Méthode 2 : vraies absences



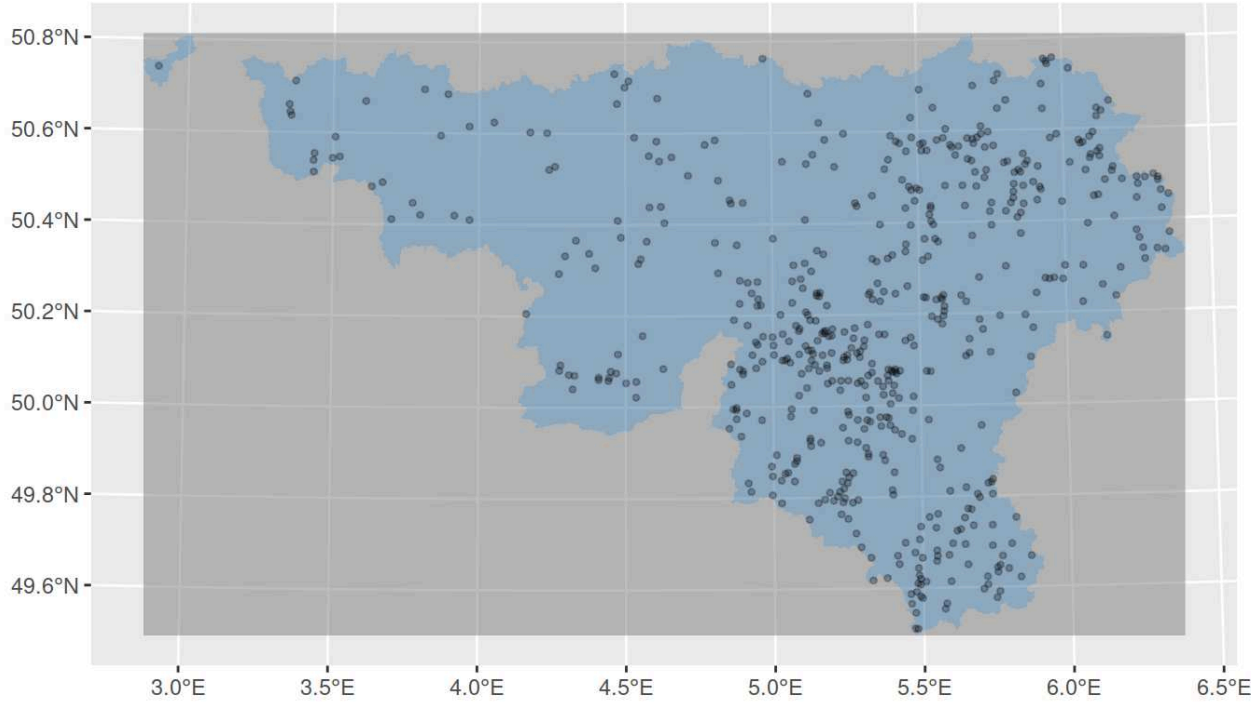
Nyctalus noctula - Moyenne des modèles
(Méthode 1 + Méthode 2) / 2



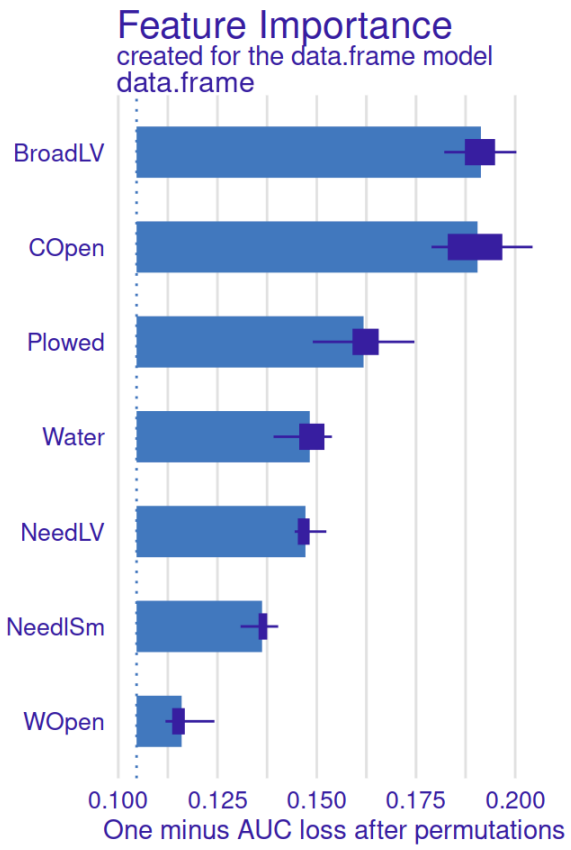
Noctule de Leisler (*Nyctalus leisleri*)

Nombre de données de présence **avant** affinage : 2051

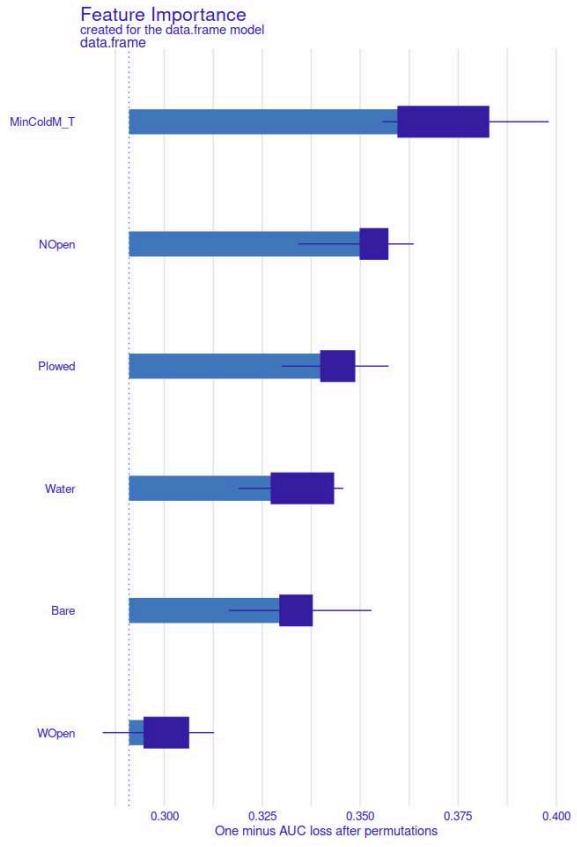
Nombre de données de présence **après** affinage : 562



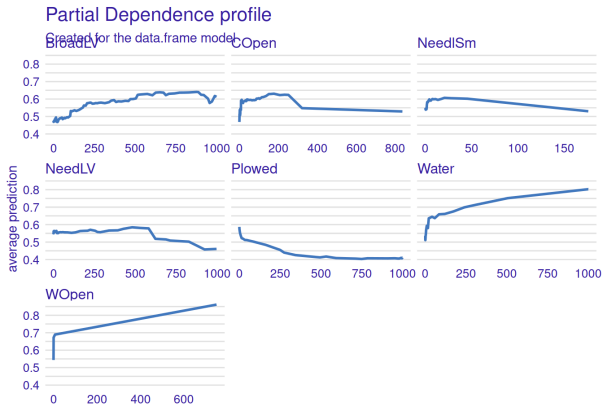
Méthode 1 : pseudoabsences



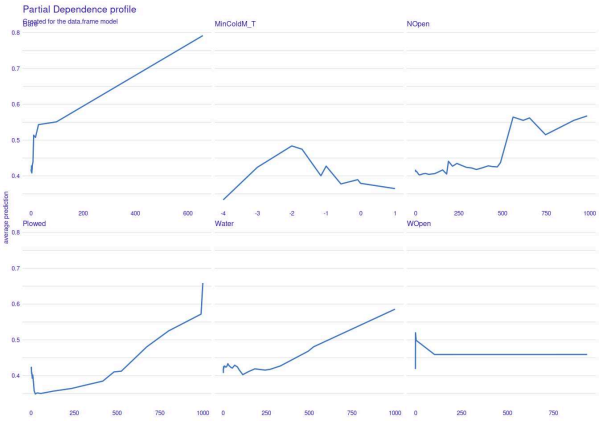
Méthode 2 : vraies absences



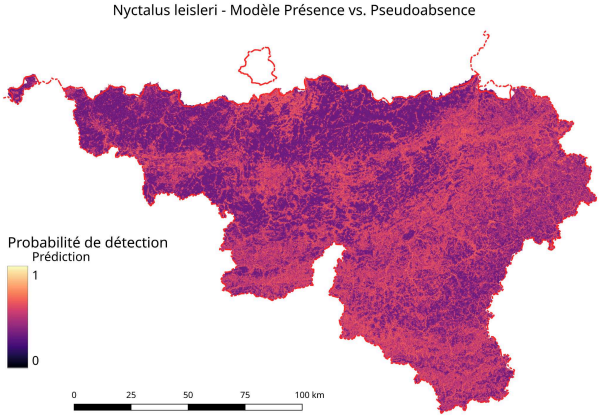
Méthode 1 : pseudoabsences



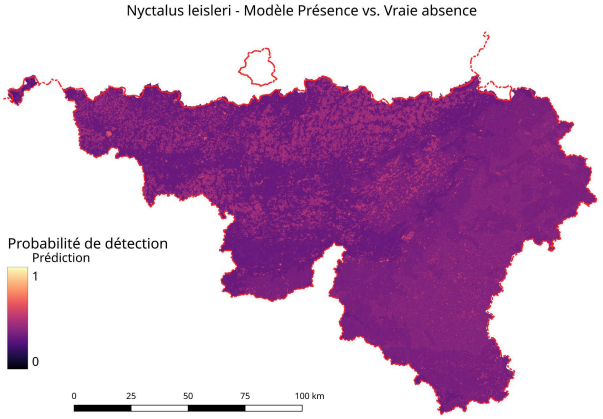
Méthode 2 : vraies absences



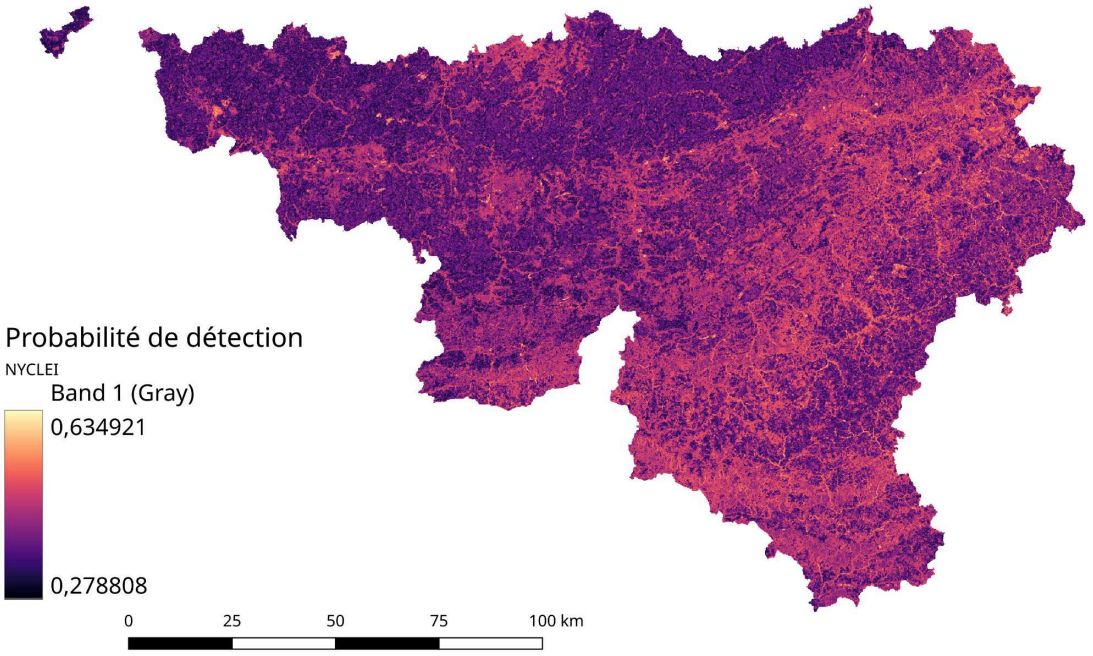
Méthode 1 : pseudoabsences



Méthode 2 : vraies absences



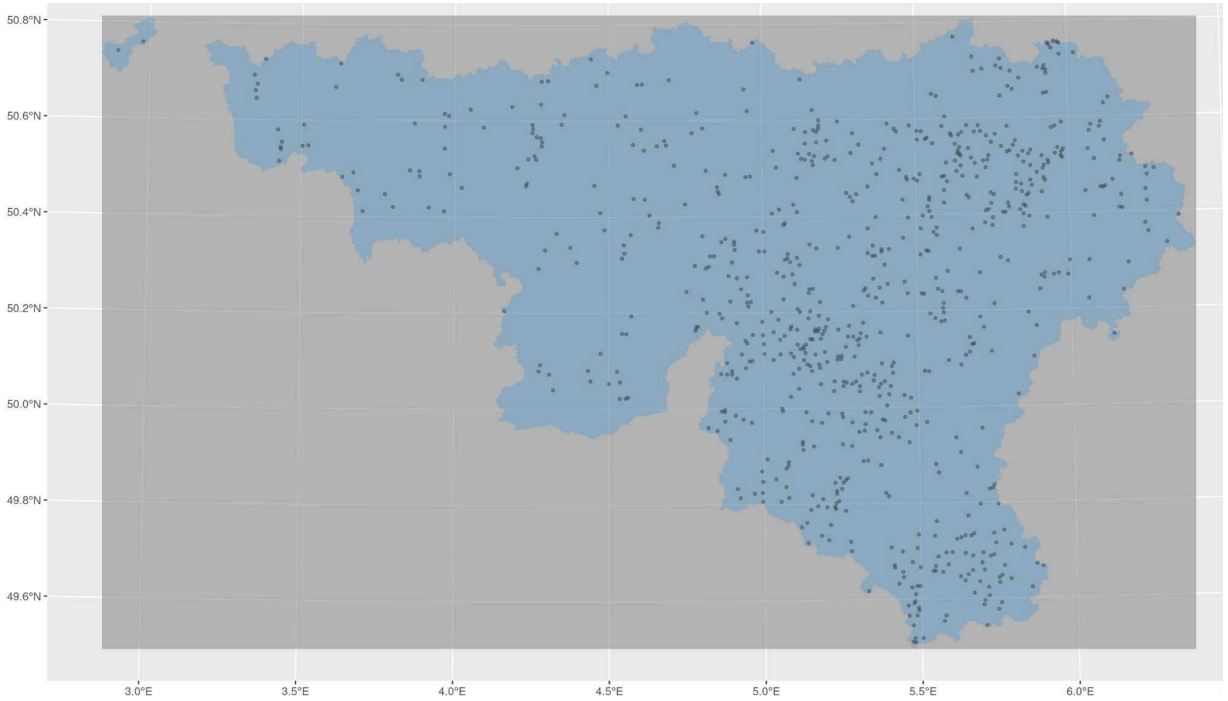
Nyctalus leisleri - Moyenne des modèles
(Méthode 1 + Méthode 2) / 2



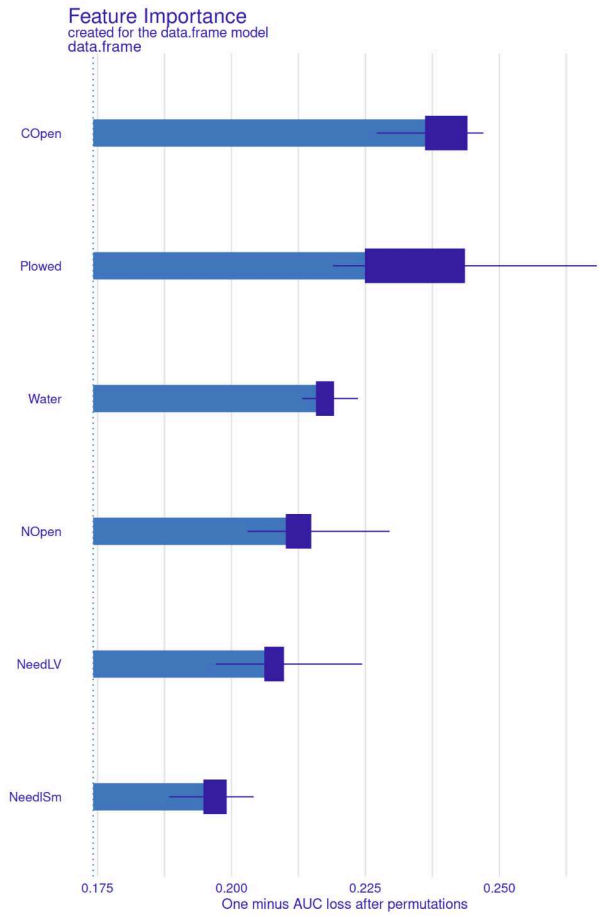
Sérotine commune (*Eptesicus serotinus*)

Nombre de données de présence **avant** affinage : 3006

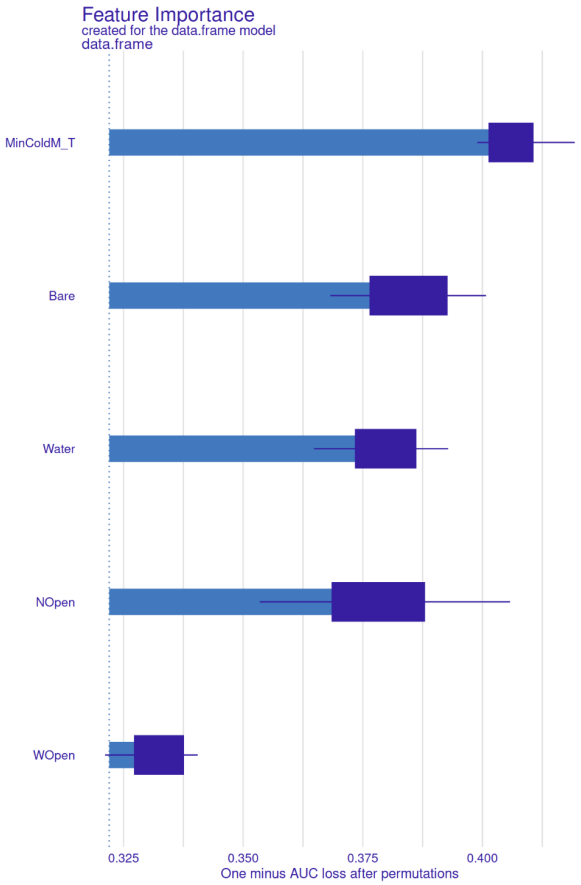
Nombre de données de présence **après** affinage : 767



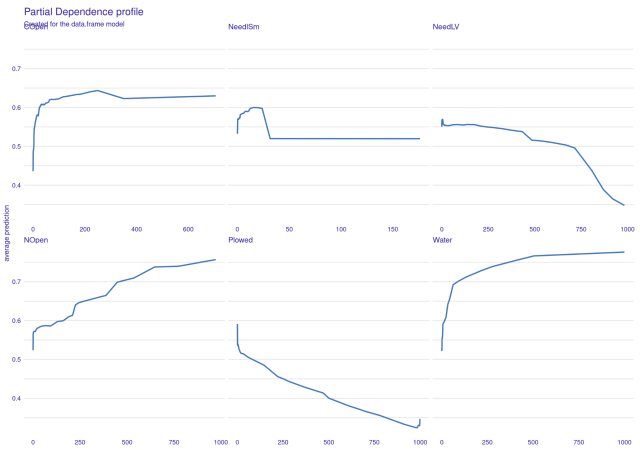
Méthode 1 : pseudoabsences



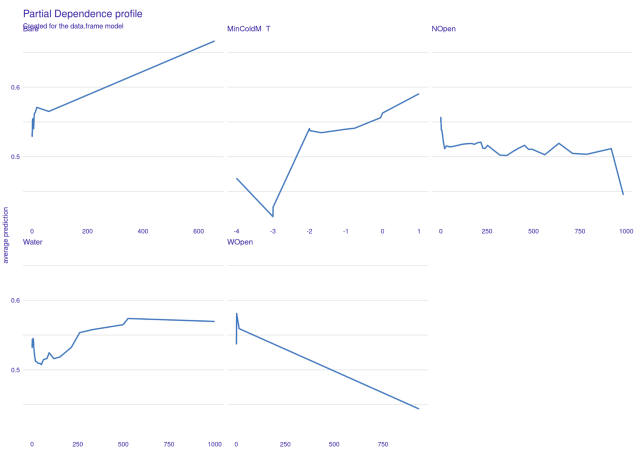
Méthode 2 : vraies absences



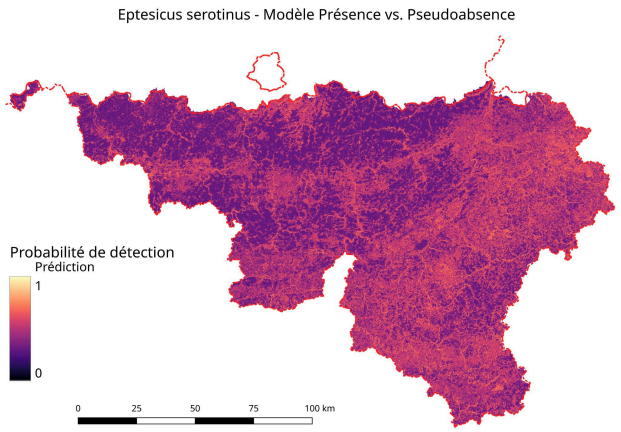
Méthode 1 : pseudoabsences



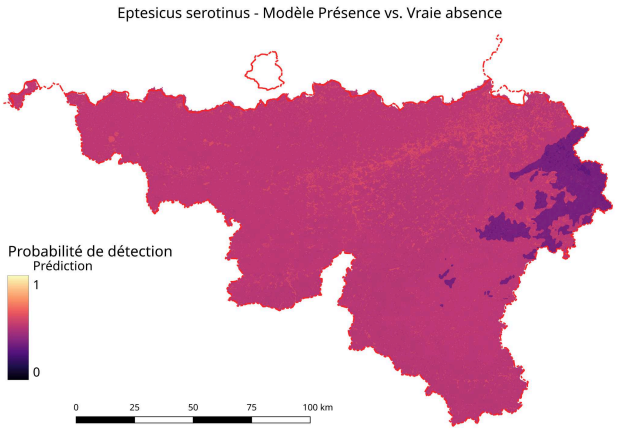
Méthode 2 : vraies absences



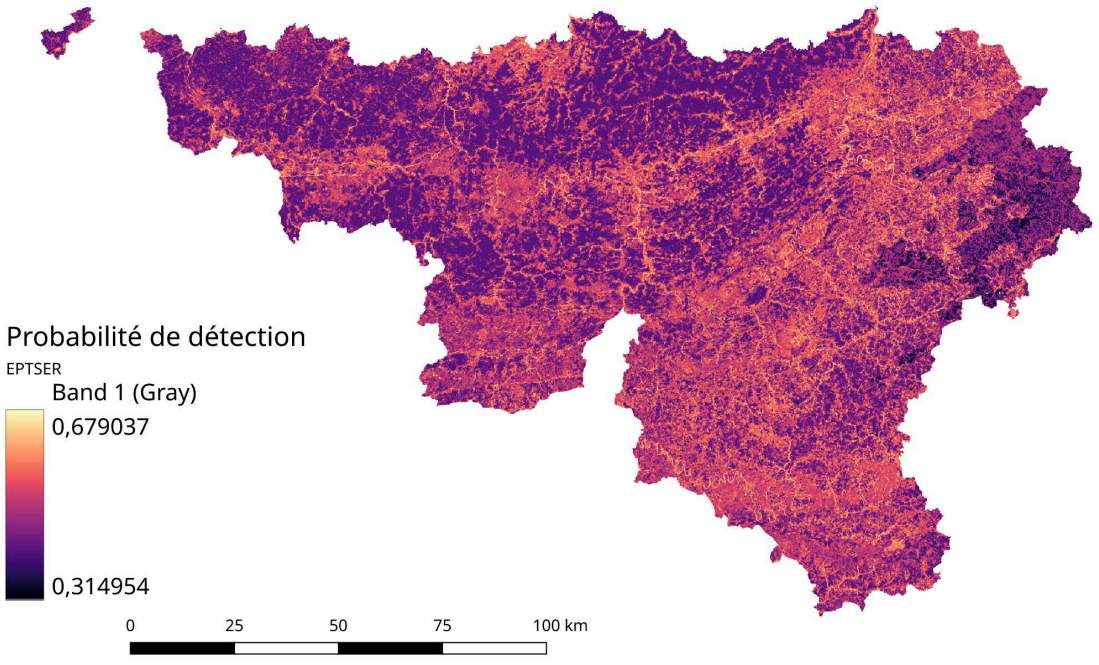
Méthode 1 : pseudoabsences



Méthode 2 : vraies absences



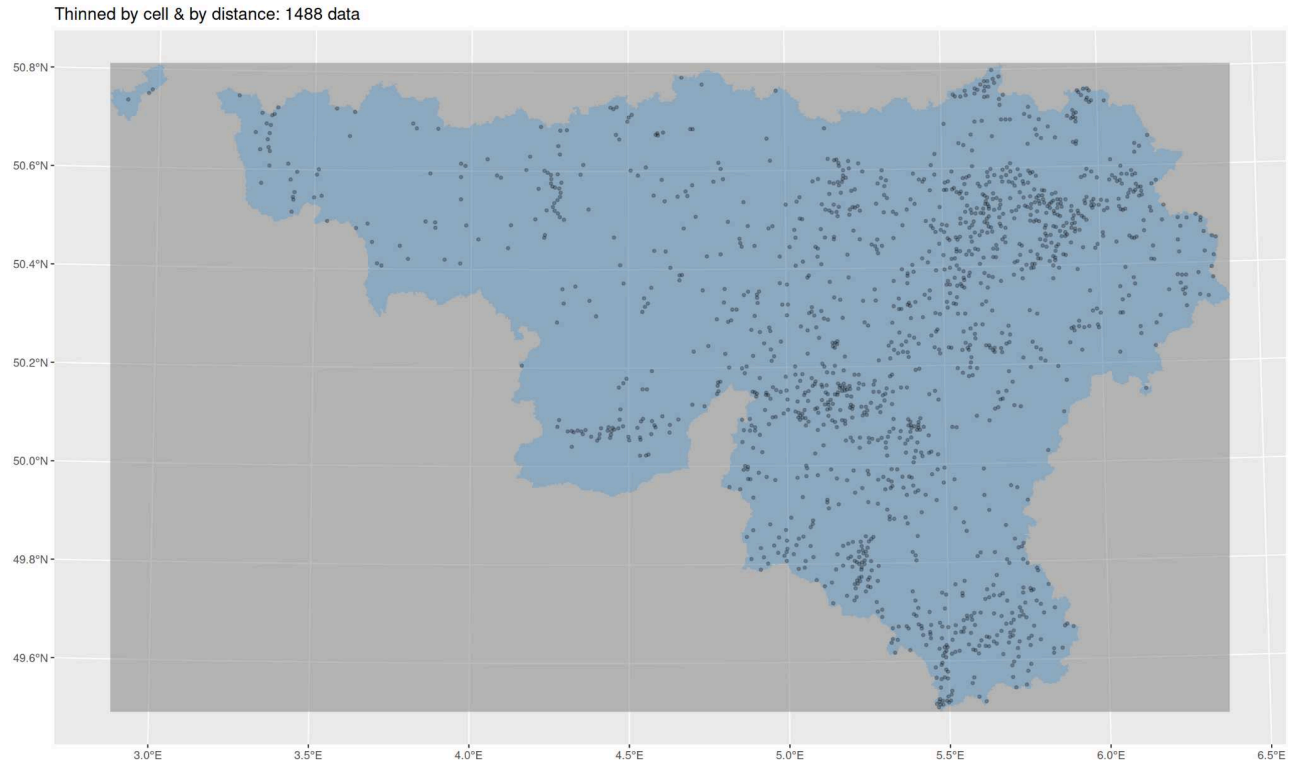
Eptesicus serotinus - Moyenne des modèles
(Méthode 1 + Méthode 2) / 2



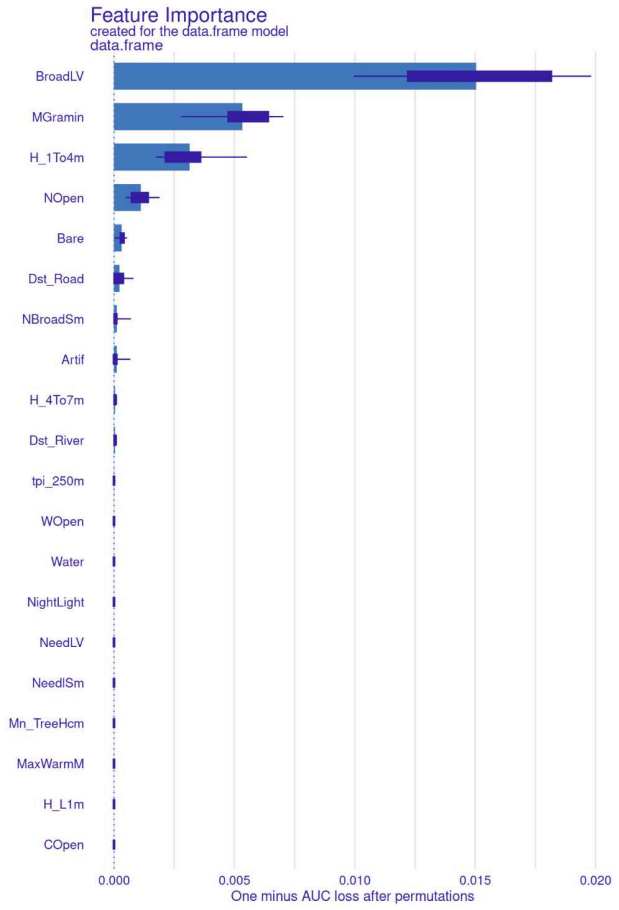
Pipistrelle commune (*Pipistrellus pipistrellus*)

Nombre de données de présence **avant** affinage : 12404

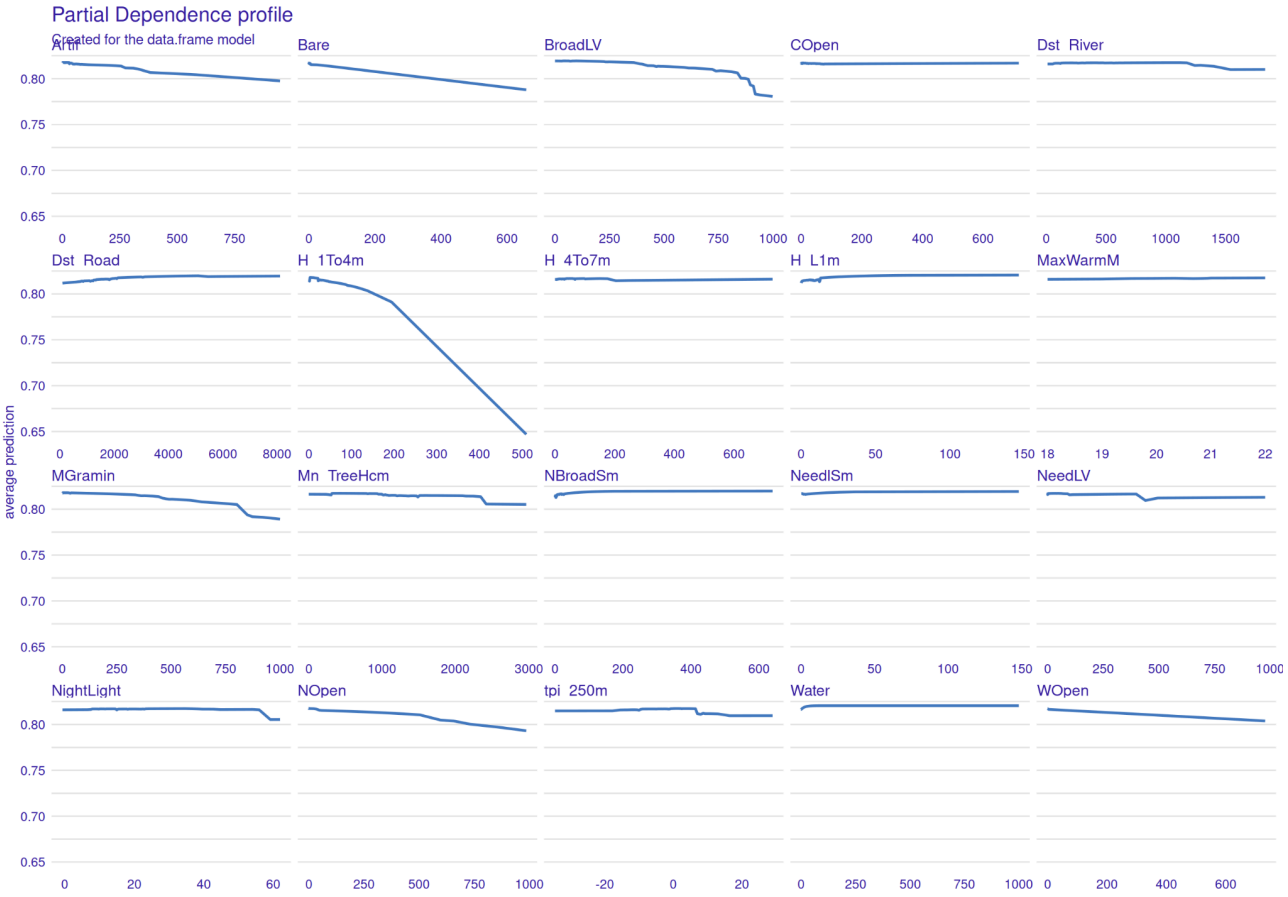
Nombre de données de présence **après** affinage : 1488



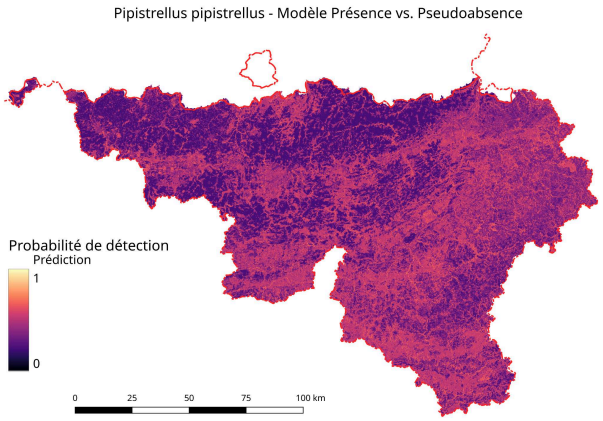
Méthode 1 : pseudoabsences (NB. la modélisation de la seconde méthode n'a pas pu être calculée)



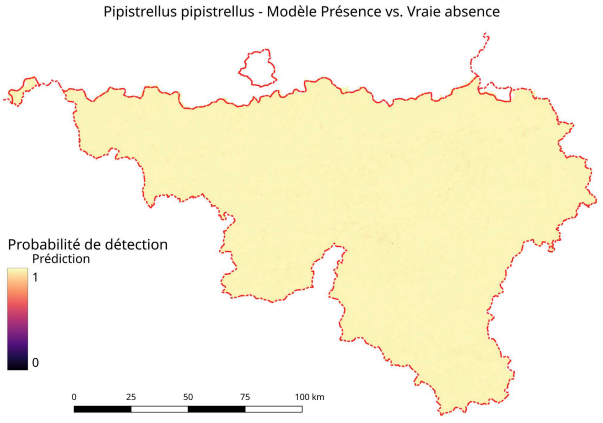
Méthode 1 : pseudoabsences



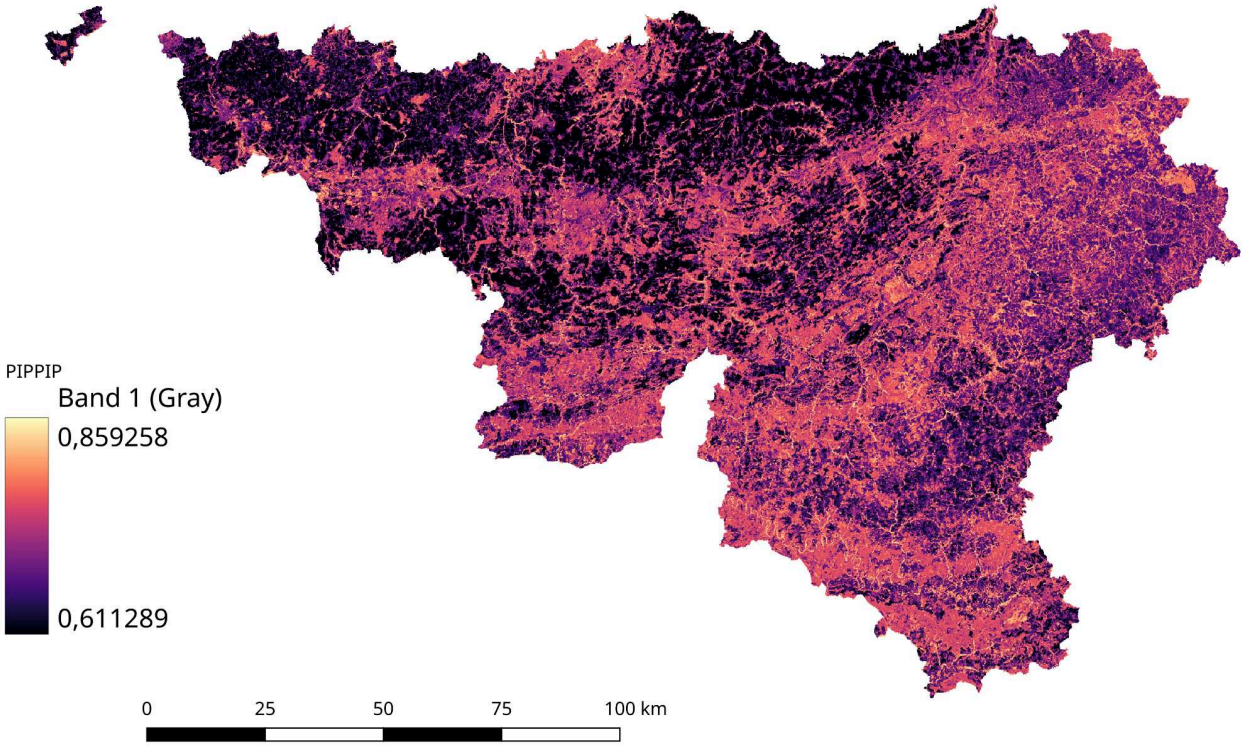
Méthode 1 : pseudoabsences



Méthode 2 : vraies absences



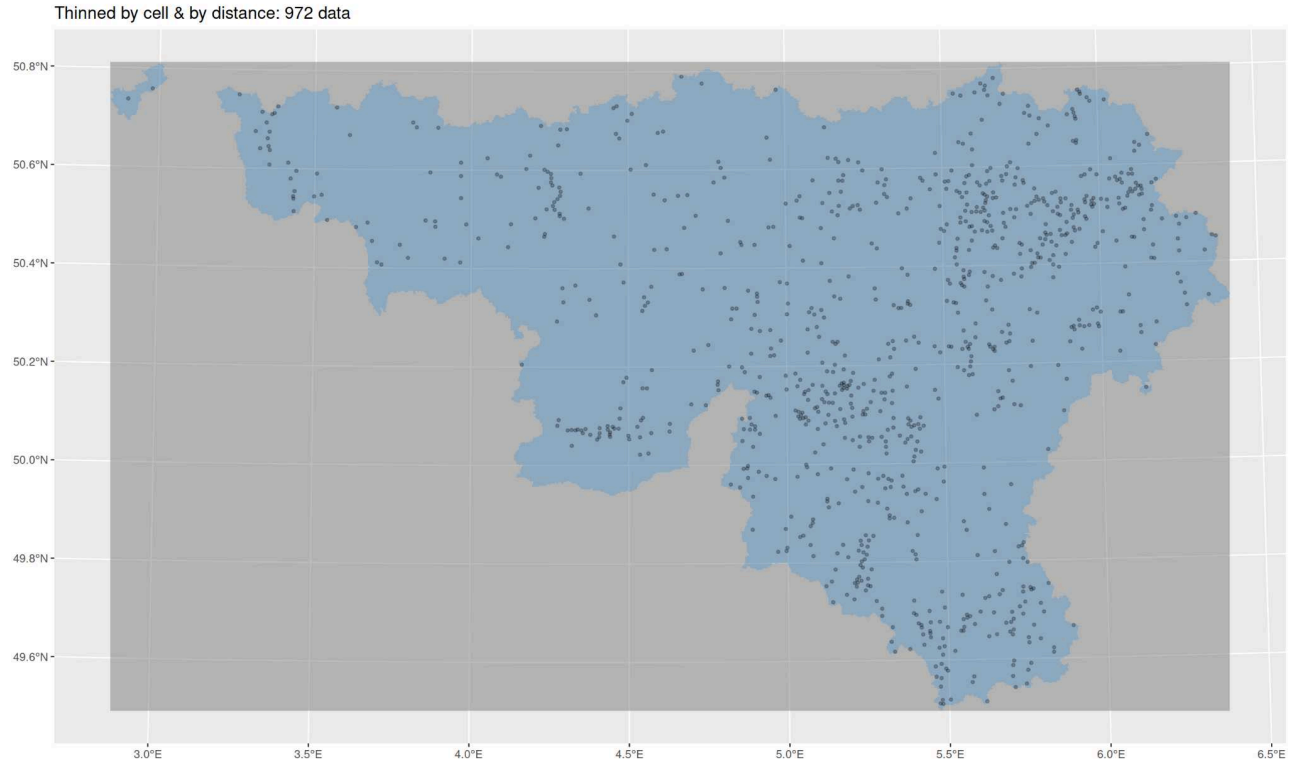
Pipistrellus pipistrellus - Moyenne des modèles
(Méthode 1 + Méthode 2) / 2



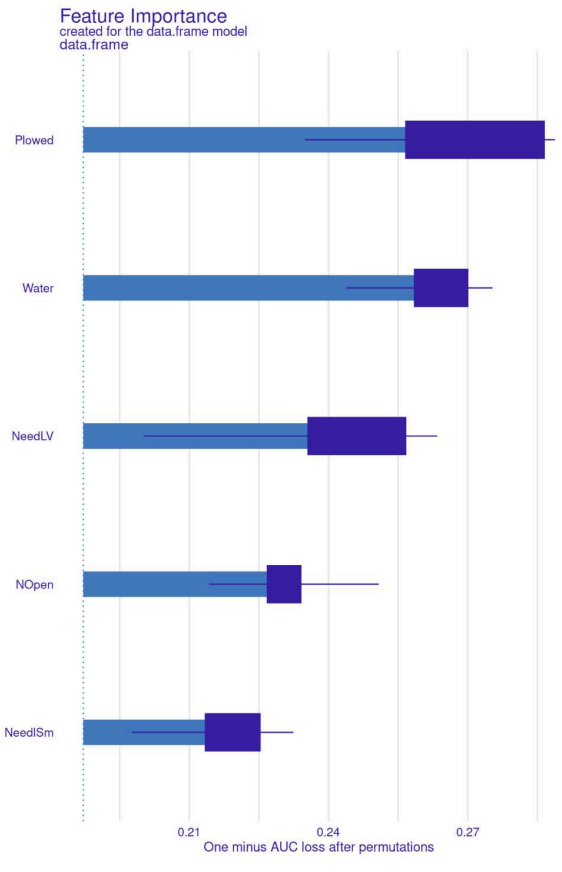
Pipistrelle de Nathusius (*Pipistrellus nathusii*)

Nombre de données de présence **avant** affinage : 7036

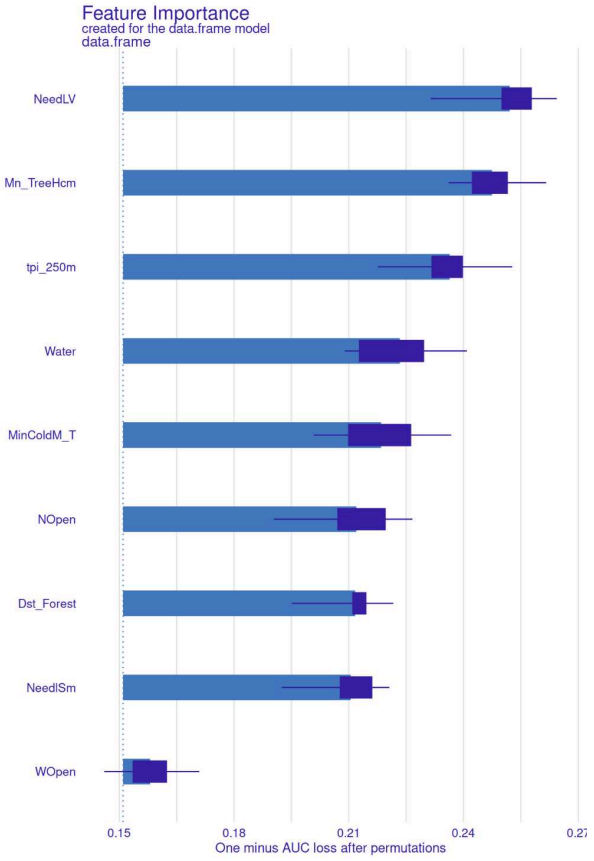
Nombre de données de présence **après** affinage : 972



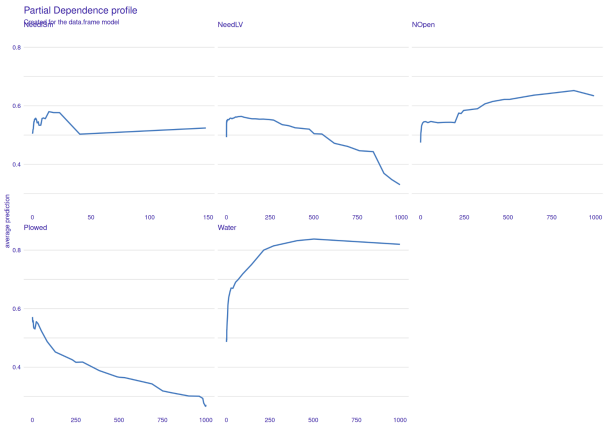
Méthode 1 : pseudoabsences



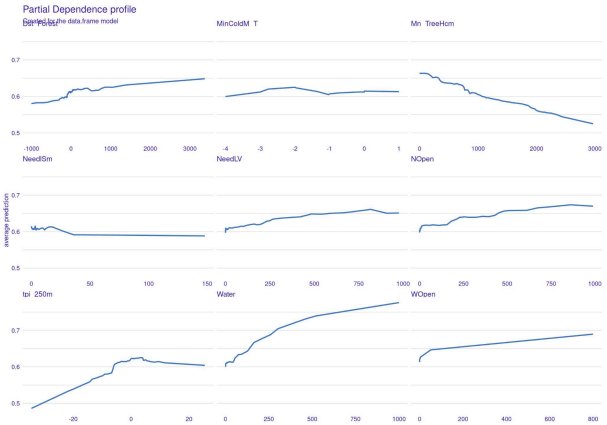
Méthode 2 : vraies absences



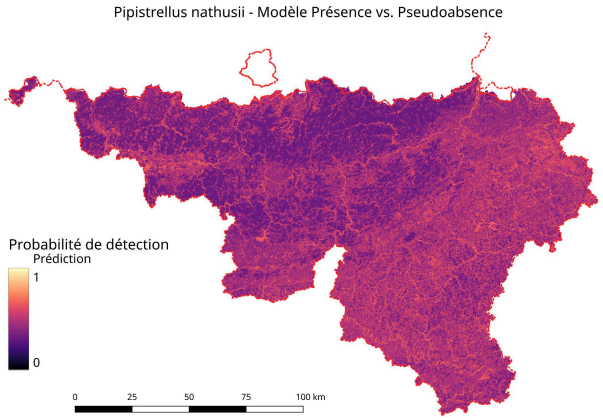
Méthode 1 : pseudoabsences



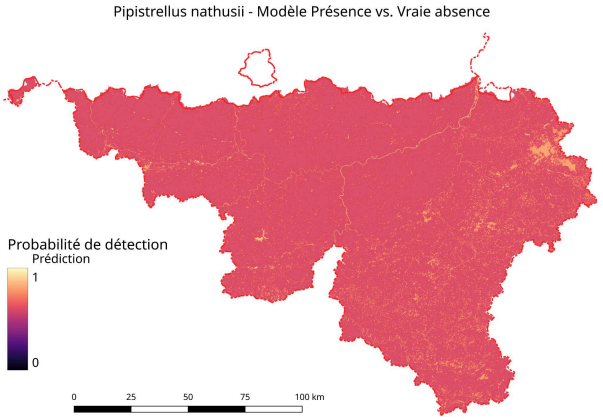
Méthode 2 : vraies absences



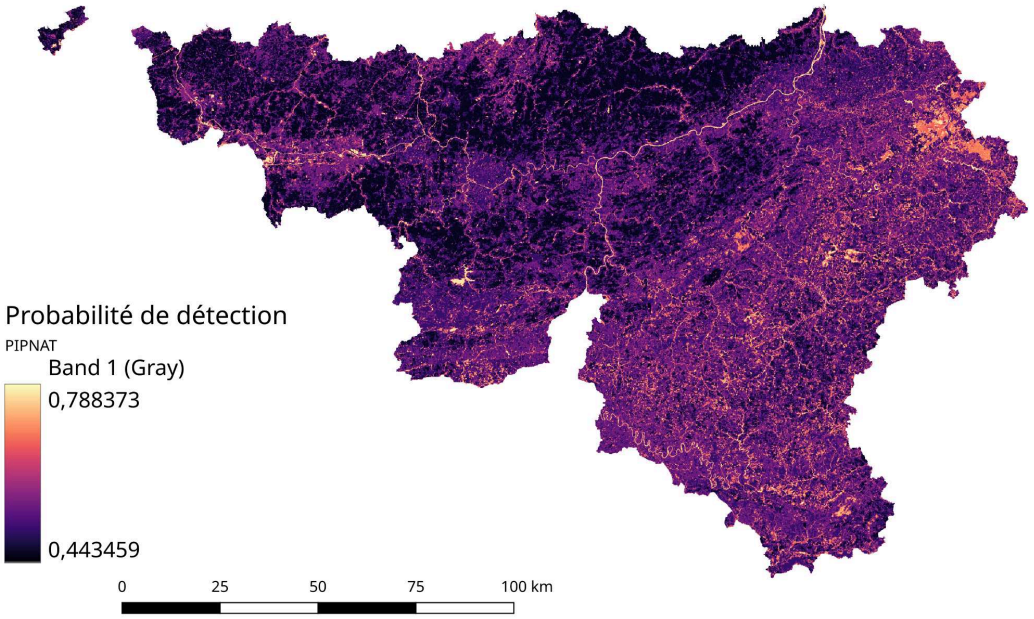
Méthode 1 : pseudoabsences



Méthode 2 : vraies absences



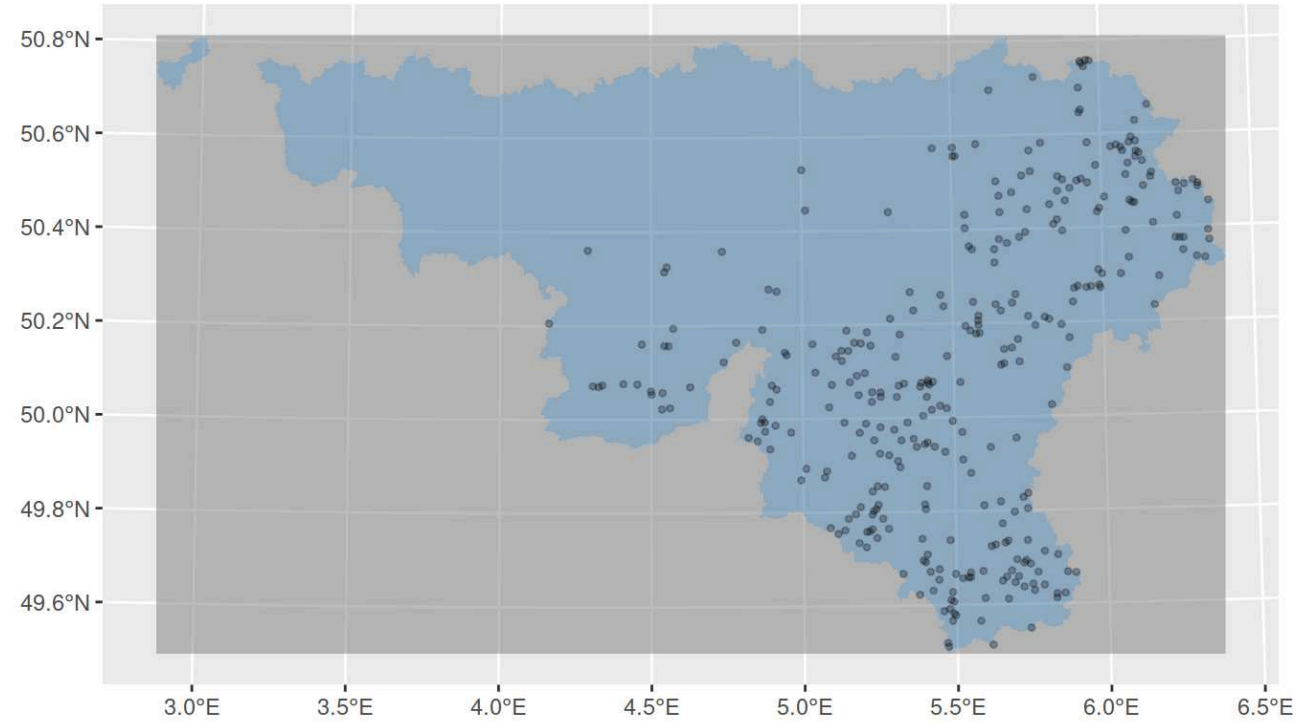
Pipistrellus nathusii - Moyenne des modèles (Méthode 1 + Méthode 2) / 2



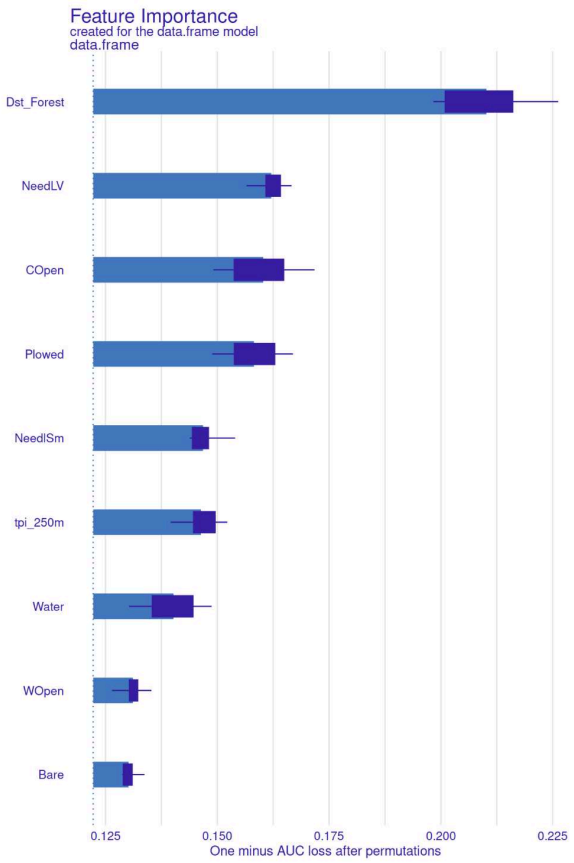
Grand murin (*Myotis myotis*)

Nombre de données de présence **avant** affinage : 938

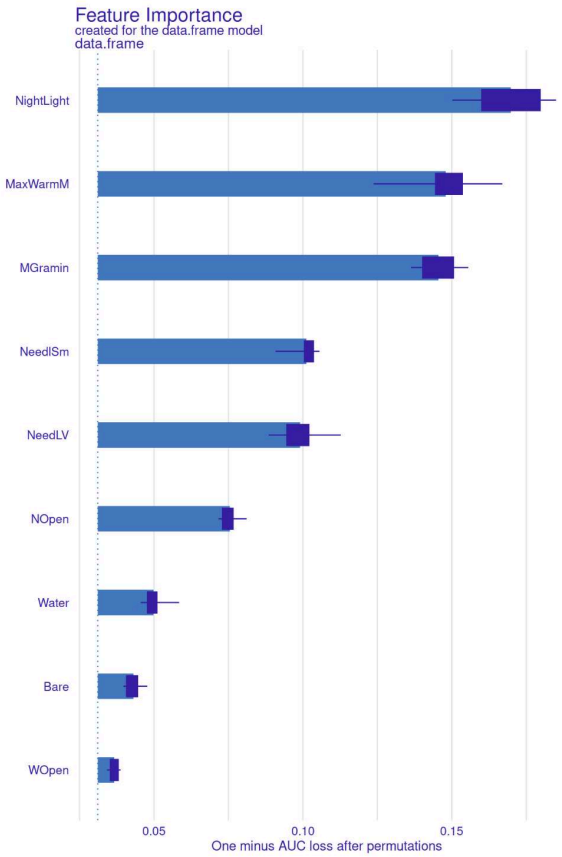
Nombre de données de présence **après** affinage : 333



Méthode 1 : pseudoabsences

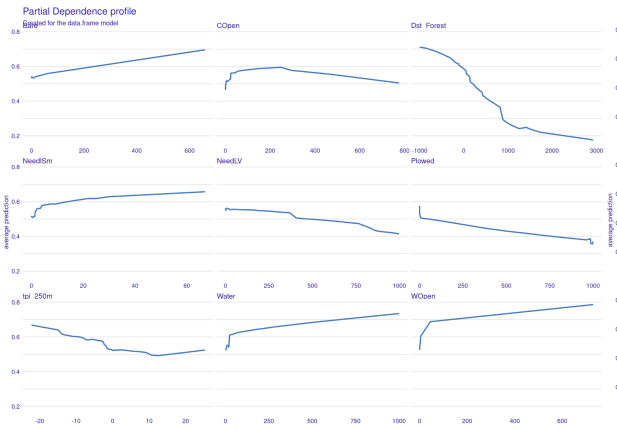


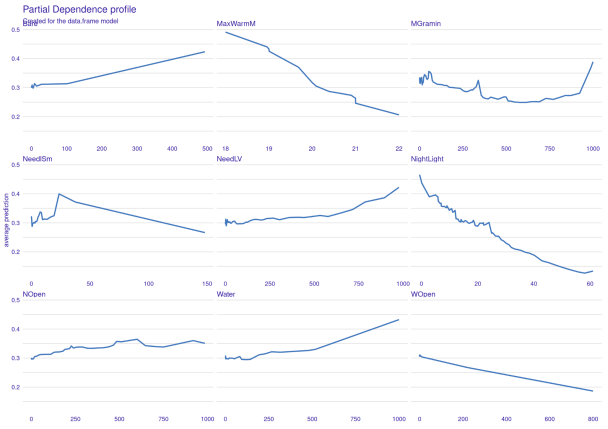
Méthode 2 : vraies absences



Méthode 1 : pseudoabsences

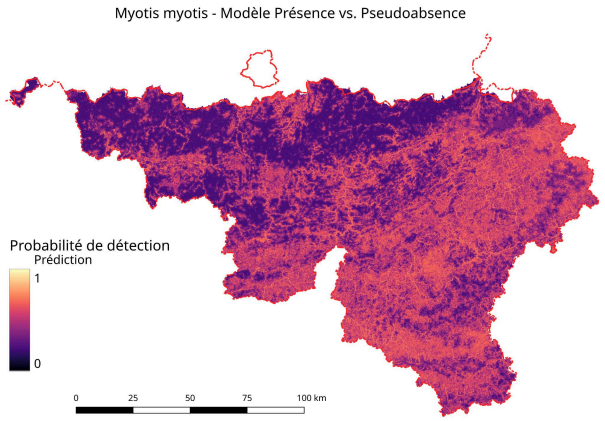
Méthode 2 : vraies absences

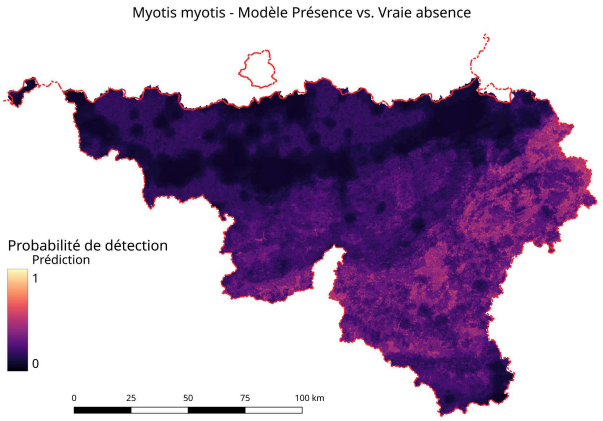




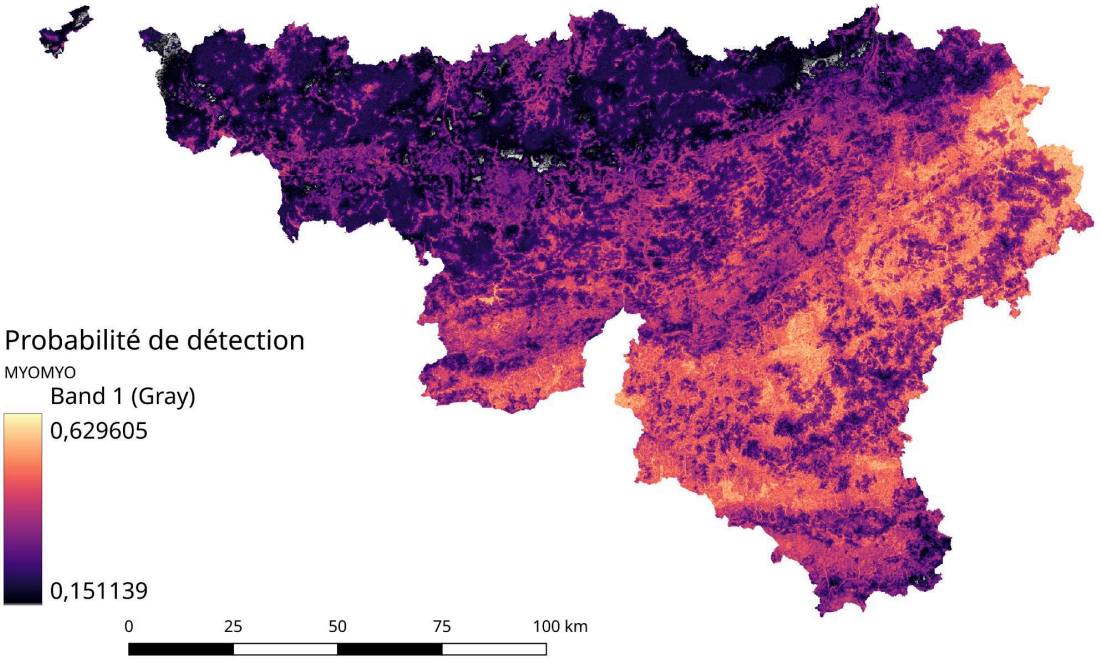
Méthode 1 : pseudoabsences

Méthode 2 : vraies absences





Myotis myotis - Moyenne des modèles
(Méthode 1 + Méthode 2) / 2



4.2 Discussion

Pour chaque espèce modélisée, on a intégré dans ce rapport les éléments clé d'analyse :

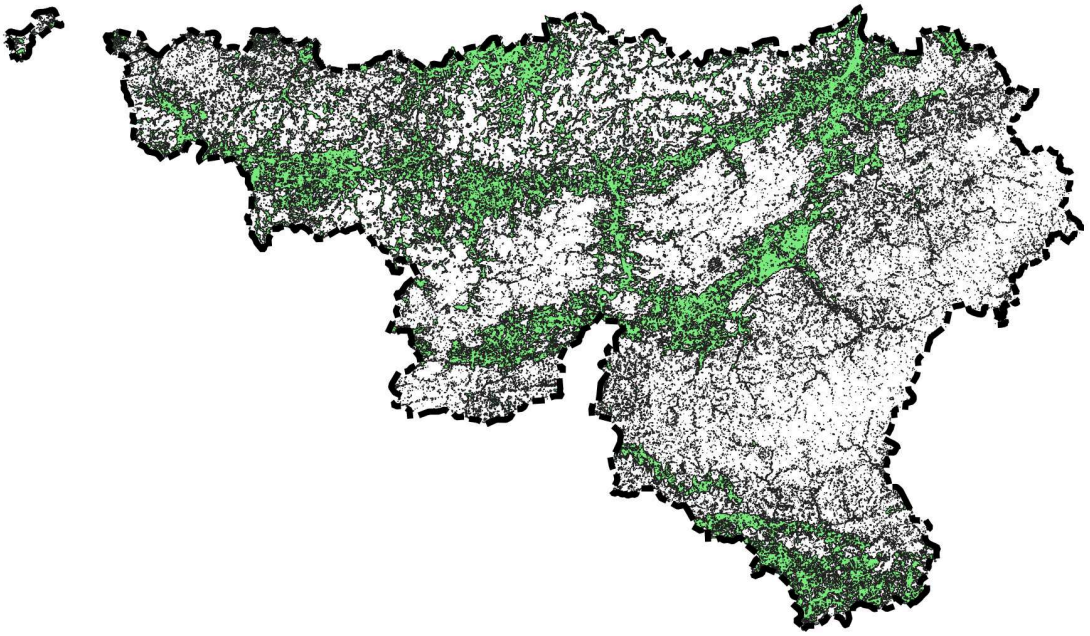
- description succincte des données source (nombre et localisation)
- poids des paramètres environnementaux choisis par les modèles pour les deux méthodes comparées
- évolution des prédictions en fonction de ces paramètres environnementaux pour les deux méthodes comparées
- cartes donnant un aperçu des valeurs prédites sur l'ensemble de la Région wallonne pour les deux méthodes comparées
- proposition d'un consensus des prédictions issues des deux méthodes

Lisibilité des cartes

Bien que tous les résultats bruts soient fournis en annexe numérique, nous avons choisi de les illustrer dans ce rapport sur base d'une échelle complète de valeurs possible s'étalant de 0 à 1. Le but est de donner une vision parfaitement comparable de chaque espèce pour les deux méthodes. Le désavantage est que cette échelle a tendance à gommer les contrastes puisqu'aucune espèce n'obtient réellement des valeurs allant de 0 à 1. C'est particulièrement visible pour la pipistrelle commune, qui obtient des prédictions proches de 1 sur quasi tout le territoire et c'est bien cohérent par rapport à la quasi ubiquité de cette espèce.
La troisième carte de "compromis" est, quant à elle, représentée espèce par espèce sur une échelle adaptée, qui s'étale de la valeur minimale à la valeur maximale des prédictions. Cette représentation met nettement mieux en évidence les contrastes de distribution.

Une dernière option, probablement plus intuitive, serait de transformer chaque raster en une couche vectorielle (polygones) qui ne conserve que les pixels du raster dont la prédiction est supérieure à un seuil déterminé. De cette manière, on transforme des gradients de probabilité en catégories : présent vs. absent.
La difficulté est alors de décider quelle valeur seuil traduit au mieux la réalité du terrain. Nous avons fait cet exercice en utilisant le seuil de probabilité de 0.5. Les couches vectorielles produites se trouvent aussi en annexe numérique.

Un exemple est illustré ci-dessous pour la noctule commune. Les surfaces en vert représentent les prédictions supérieures à 0.5. Étant donné la finesse de la maille des rasters (100x100m) le résultat est très découpé, ce qui pourrait être utile si on s'intéresse à une zone précise du territoire mais rend la lisibilité plus difficile à large échelle. Une option pourrait consister à d'abord ré-échantillonner les rasters, par exemple à une maille kilométrique avant de produire la couche vectorielle. On devrait aussi tester différents seuils car il n'est pas implicite que des prédictions inférieures à 0.5 soient pertinentes, du moins pour toutes les espèces. Des seuils tels que 0.25 ou 0.3 pourraient être aussi réalistes.



Représentation de la distribution de Nyctalus noctula *sous forme d'une couche vectorielle déduite du raster produit par la méthode des pseudoabsences.*

4.3 Perspectives

La base de données acoustique utilisée est très riche en données et couvre assez bien tout le territoire wallon. Il subsiste malgré tout des disparités d'effort d'échantillonnage dues aux choix des observateurs et à l'attractivité de certaines régions naturelles ou de sites protégés. Un autre écueil à l'homogénéité des données provient aussi des études ciblées telles que les EIE qui apportent des milliers de données sur quelques sites. Ces disparités du jeu de données pourraient être amoindries par des ré-échantillonnages (bootstrapping) par exemple.

Utilisation des données d'abondance
Nous avons exploré deux méthodes pour ce travail, celles des pseudoabondances "classiquement" utilisées pour des observations non planifiées et celles des vraies absences qui postule que les nuits complètes d'enregistrement donnent des absences fiables. Une troisième approche envisagée pourrait être de tenir compte des abondances relatives approximées par les nombres de contacts et de minutes positives. Cela semble envisageable étant donné la quantité de données présentes dans la base mais il y aurait certainement lieu d'effectuer des manipulations de standardisation ou de ré-échantillonnage pour minimiser les différences d'effort d'échantillonnage.

Compléter et nuancer les données sources
On pourrait essayer de réunir des données qui ne se trouvent pas encore dans la base acoustique actuelle (par exemple en faisant le tour des autres bureaux d'études qui ont potentiellement des données acoustiques, en allant trouver le secteur éolien qui possède beaucoup de données, en recontactant différents interlocuteurs dans le secteur associatif pour élargir la base de données avec jeux non transmis), et refaire tourner les scripts d'analyse pour apporter plus de précision aux modélisations. En début de mission, nous avions envisagé de réunir certains jeux de données identifiés (LIFE Elia et autres suivis

actuellement menés par Ecofirst, des données de bureaux d'études, secteur éolien, observateurs bénévoles, …) mais cela n'a pas été possible à cause du temps consacré à la préparation de la base existante.

Tenter de séparer les saisons dans l'analyse

On connaît de mieux en mieux l'effet de la saisonnalité sur la distribution des chauves-souris, notamment pour les espèces réputées migratrices (*P. nathusii, P. pygmaeus, N. noctula, N. leisleri*). Il y aurait ici aussi certainement moyen de gagner en qualité de la modélisation en différenciant les données des périodes migratoires des données estivales, à l'instar de ce qui est réalisé par l'équipe du MNHN (Bat migration routes in Europe).

Autres espèces

Nous avons préparé la base de données indifféremment pour toutes les espèces. Puisque les scripts d'analyse sont parfaitement reproductibles, on peut facilement exécuter les modélisations sur l'ensemble des espèces pour lesquelles on considère avoir assez d'informations suffisamment précises. On pourrait par exemple modéliser les espèces suivantes : *Rhinolophus hipposideros, R. ferrumequinum, Myotis daubentonii, Myotis bechsteinii, Myotis emarginatus, Myotis nattereri*, *Barbastella barbastellus*… ou rester au rang générique ou de taxon par exemple pour *Plecotus spp, Myotis mystacinus/brandtii/alcathoe*. Par contre ce ne serait pas pertinent pour les espèces rarissimes et/ou difficilement détectables (*Vespertilio murinus, Eptesicus nilssonii*). Ces choix pourraient être décidés par les validateurs de la base de données qui connaissent le mieux le niveau de fiabilité associé à chaque taxon.

5. Délivrables annexes à ce rapport

Structure des annexes numériques

- Analyses_R
    - Rapports_complets_analyses: sortie html des R-markdown. NB. plusieurs html n'ont pas abouti, probablement pour cause informatique!
    - Scripts: deux scripts au format R-markdown (.Rmd), correspondent aux deux méthodes d'analyse + script de nettoyage de la base de données
- SDM_Cartes_images (format png)
    - Compromis_Méthodes1-2
    - Méthode1_Pseudoabsences
    - Méthode2_Absences
- SDM_Rasters (format geotif)
    - Compromis_Méthodes1-2
    - Méthode1_Pseudoabsences
    - Méthode2_Absences
- SDM_Vectoriel: 6 couches vectorielles réunies dans un geopackage
- AllBats_RW.gpkg: Base de données nettoyée

Concernant la synthèse des résultats au format (RANGE) du rapportage pour la Directive européenne "Habitats" (article 17), le DEMNA nous confirme que les cartes de distribution et de Range sont calculées automatiquement via un outil standardisé fourni par la Commission Européenne, pour chaque espèce à partir des données de présence effective croisée avec une grille de référence 10 x10. Conformément aux instructions reçues par le DEMNA, nous ne nous en préoccupons pas ici.

Par contre, toujours dans l'idée d'alimenter les informations disponibles dans le cadre du rapportage pour le volet habitat, nous avons spécialement travaillé sur **une version vectorielle des couches de probabilité de présence** (voir discussion).