AGROCAMPUS OUEST

Année universitaire : 2019-2020. Spécialisation : Science des données	**Mémoire de fin d'études** ■ d'ingénieur de l'École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage (AGROCAMPUS OUEST), école interne de l'institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement ☐ de master de l'École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage (AGROCAMPUS OUEST), école interne de l'institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement ☐ d'un autre établissement (étudiant arrivé en M2)

Etudes de données issues de l'écologie participative : qualité et influence des protocoles

Par : Julien PRUNEAU

Soutenu à ***Rennes*** ***le*** *07/09/2020*

Devant le jury composé de :

Président :

Maître de stage : BENATEAU Simon

Enseignant référent : CAUSEUR David

Autres membres du jury (Nom, Qualité)

Sommaire :

Table des figures :

Contexte :

Vigie-nature est un programme de sciences participatives visant à étudier le devenir de la biodiversité face aux changements globaux à l'échelle française. Aujourd'hui, 15 programmes différents ont été développés, chacun visant à répondre à une question spécifique et basé sur des protocoles standardisés. Les programmes sont adaptés à différents types de publics : naturalistes qualifiés, agriculteurs, scolaires, gestionnaires d'espaces verts, et pour la moitié d'entre eux, le grand public. Chaque année, plus de 15 000 bénévoles participent à la collecte des données sur les espèces communes. Chaque programme se concentre sur un groupe particulier d'espèces : oiseaux, papillons, chauves-souris, mollusques, insectes pollinisateurs, libellules, plantes sauvages des villes, etc. Vigie-Nature poursuit un objectif scientifique avec la collecte d'une grande quantité de données standardisées permettant aux scientifiques de surveiller l'état de la biodiversité, et également un objectif éducatif de sensibilisation à la biodiversité par l'observation de la nature. Des enjeux importants sont au cœur du programme tels que l'effet du changement climatique (et sa mise en évidence), l'adaptation des organismes au milieu urbain ou l'impact sociologique de la participation, les protocoles ont donc été conçus afin de pouvoir traiter ces questions.

La portée de ces programmes ne cesse de grandir ce qui entraine une forte augmentation de la quantité de données disponibles (Houiller, 2016). Cette abondance permet à de nombreux articles de voir le jour sur des sujets variés (évolution temporelle des populations, état des paysages, comparaison des milieux…). C'est pourquoi il est intéressant de questionner la qualité des données recueillies. Les publics visés diffèrent ainsi que les protocoles associés, ajoutant une grande variabilité aux données. Une fusion de ces protocoles peut-elle permettre d'avoir une information plus complète ou si au contraire les différences sont trop importantes ou la qualité des données trop faible.

L'objectif de ce stage est de mesurer la qualité relative des données de Vigie-Nature en fonction des publics (scolaires, agriculteurs, grand public), en comparant des métriques obtenues à partir de données issues de divers observatoires pour un même taxon. Les résultats obtenus serviront ensuite de support pour communiquer avec les participants de ces programmes de manière simple et visuelle. Pour répondre à ces questions une démarche qui se veut générique pour pouvoir être appliquée à n'importe quel taxon doit être développée. Cette méthode devra être claire et compréhensible pour pouvoir être reproduite par d'autres personnes.

Données :

Ce stage porte sur des comptages de mollusques terrestres (escargots et limaces) pour trois types de participants : les agriculteurs (Observatoire Agricole de la Biodiversité), le milieu scolaire (Opération escargots – Vigie-Nature Ecole) et le grand public (Opération escargots – Vigie-Nature).

Trois jeux de données sont à disposition (un pour chacun des observatoires) pour lesquels la structure est similaire. Une ligne correspond à un relevé du participant (une observation) avec en colonne toutes les variables mesurées.

Observatoire Agricole de la Biodiversité (OAB) :

L'Observatoire Agricole de la Biodiversité (OAB) propose des protocoles d'observation de l'ensemble des espèces abondantes en milieu agricole aux exploitants intéressés en vue de mieux connaître cette biodiversité ordinaire et ses liens avec les pratiques. Quatre taxons sont proposés par cet observatoire mais seuls celui des mollusques terrestres est utilisé. Les agriculteurs participants aidés par des formateurs mettent en place sur leurs parcelles des planches de bois. Ils reviennent ensuite les soulever régulièrement pour compter l'effectif de chaque groupe morphologique de mollusques observé en étant le plus souvent accompagné par un membre de l'observatoire (protocole planche). Les données utilisées durant ce stage sont issues d'une sauvegarde de la base de données OAB. Ce jeu de données est constitué de 16 597 lignes correspondantes à autant d'observations et 127 variables (dont seulement 30 sont analysables). Quelques variables spécifiques sont disponibles pour ce jeu de données notamment les types de bordures de champ, la distance à des biomes (forêts, prairie…) ainsi que le type de culture et de conduite (biologique ou conventionnelle).

Vigie-Nature Ecole (VNE) :

Ce programme est destiné aux classes de tout niveau (allant du primaire au lycée). Les enseignants peuvent suivre une formation en direct ou se former en ligne grâce aux documents et outils de formation disponibles sur le site internet du programme et ensuite installer des planches en bois à proximité de l'école. Ils reviennent régulièrement les soulever avec leurs élèves afin de compter l'effectif de chacun des groupes morphologiques de mollusques présents (protocole planche). Le jeu de donnée utilisé pendant ce stage est issu d'une requête qui interroge la base de données VNE. Des données concernant le niveau des classes ainsi que des informations sur la disposition de la planche (contre un mur, à l'ombre d'un arbre…) sont disponibles pour cet observatoire. Ce sont 1386 observations et 117 variables (dont seulement 40 sont analysables) qui constituent ce jeu de données.

Opération Escargots – Vigie-Nature (GP) :

Ce dernier programme est destiné au grand public. Deux protocoles sont possibles. En plus du protocole planche similaire aux autres observatoires, le participant peut aussi réaliser ce comptage en marchant une vingtaine de minutes par temps pluvieux plusieurs fois par mois et compter l'effectif de chacun des groupes morphologiques de mollusques observés. C'est ce qui sera appelé protocole inventaire (en opposition avec le protocole planche précédemment décrit). Le jeu de données utilisé pendant ce stage est issu d'une requête qui interroge directement la base de données GP. Ce sont 73 variables (dont seulement 39 sont analysables), de 632 lignes pour le protocole planche et 4121 pour le protocole inventaire qui constituent ce jeu de données.

Les variables d'intérêt de ces jeux de données diffèrent mais peuvent être résumées en plusieurs catégories :

- Des données de comptage (une variable en colonne par groupe morphologique) avec le nombre de mollusques observés. Ces groupes morphologiques ont été harmonisés entre les observatoires. Cette synthèse a été la plus exhaustive possible (les groupes morphologiques finaux sont ceux de l'observatoire qui en avait le moins). Ce sont 11 groupes différents en plus des champs « autres escargots » et « autres limaces » qui sont sélectionnés. La diversité totale (nombre de groupes morphologiques différents

observés allant de 0 à 13) et l'abondance totale (nombre total de mollusques observés) sont calculées à partir de ces données de comptage. Un premier biais est présent au niveau de cette variable puisqu'il peut y avoir plusieurs espèces différentes dans le champ « autres escargots ».

- Des données de description du jardin/champ qui permettent d'avoir des informations détaillées sur la zone d'observation. Par exemple des booléens sont disponibles avec la présence/absence de variétés de plante ou d'infrastructures ou le type de cultures en place et la distance à des biomes.
- Des données sur les intrants. Ces données sont harmonisées entre les jeux de données pour correspondre à un indice de présence/absence pour chacun des types de traitement. Cependant elles sont souvent manquantes.
- Des données spatio-temporelles qui permettent de situer l'observation en France (position GPS pour certains observatoires, code postal pour d'autres) mais aussi dans le temps avec la date de saisie des données.

Chaque observatoire dispose aussi de quelques variables spécifiques liées au protocole ou bien au public visé. Par exemple des observations issues du protocole planche seront associées à des informations sur les dimensions de cette dernière.

Problématique :

Quel est l'impact des publics et des protocoles sur la qualité des observations de données issues de l'écologie participative ?

Analyse :

Cette analyse reprend la réflexion menée durant ce stage. Elle permet de mettre en place une méthode qui se veut générique et qui est constituée de plusieurs analyses liées à chaque fois à une problématique différente (ici écrite en bleu soulignée).

Quelles sont les variables analysables ?

La première étape de notre analyse a été d'enlever les variables non analysables. Deux types sont différenciés :

Les variables ayant une majorité de données manquantes. En effet les données issues des sciences participatives sont récoltées à partir de formulaires papiers ou en ligne. Certains champs ne sont pas obligatoires ce qui entraine parfois une proportion importante de données manquantes. Après discussion l'utilisation de méthodes statistiques (machine learning, imputation ou autre) semblait être une mauvaise idée pour plusieurs raisons : nos jeux de données n'étant pour certains pas assez conséquents et la volonté de rester au plus proche des données récoltées sans modifier ou renforcer les tendances existantes en créant des données. Les variables enlevées du jeu de données sont celles dépassant le seuil de 50% de valeurs manquantes. L'information contenue dans cette variable est alors considérée comme trop faible ou biaisée.

Les variables portant la même information. Certains champs des formulaires sont très similaires ou expliquent la même information. Des tests de corrélations pour chacune de nos variables ont donc été effectués. Dans le cas d'une corrélation forte (>|0,75|) après discussion la variable conservée est la plus fiable et interprétable. Cette décision prend fortement en compte la construction du formulaire de saisie (certains champs ne sont pas très clairs) et les connaissances biologiques à disposition (Barker, 2001 ; Kerney, 2015).

Quels sont les indices écologiques que nous souhaitons étudier ?

L'un des objectifs premiers des sciences participatives est de réaliser un état des lieux de la biodiversité (Houiller, 2016). C'est pourquoi l'information sur laquelle se base les études est l'abondance de groupes morphologiques observés. Des indices plus représentatifs de la présence de ce taxon peuvent être mis en place. La **diversité totale** (nombre de groupes morphologiques observés, ici de mollusques) et l'**abondance totale** ont été choisies.

Notre objectif est dans un premier temps d'évaluer la qualité de ce comptage ainsi que la capacité des participants à différencier les espèces à l'aide de clés de détermination et ce pour chacun des publics (GP, OAB, VNE).

Toutes les observations des participants sont-elles présentes dans nos jeux de données ?

Les relevés d'abondance en sciences participatives peuvent être soumis à un biais issu de l'interprétation que fait le participant du protocole. En effet il arrive souvent qu'une séance de comptage qui se solde par une abondance nulle soit interprétée comme inutile par la personne qui ne va alors pas enregistrer son comptage sur le site (Charonnet, 2019). Cela entraine une proportion beaucoup plus faible d'observations sans mollusques.

Pour étudier ce biais la proportion d'observation sans mollusques dans chacun des jeux de données est calculée. De plus la structure des formulaires de saisies est étudiée.

L'OAB ne demande pas uniquement aux participants le nombre de mollusques mais aussi celui d'autres taxons (insectes, araignées…). Cela permet d'avoir toutes les observations y compris celle où l'abondance est nulle. La possibilité de ne rien avoir sous la planche est alors très faible puisque ces taxons y sont souvent présents (en tout cas plus que les mollusques). De plus les agriculteurs sont souvent accompagnés au moins pour leurs premières observations. Pour VNE ce n'est pas le cas mais cette lacune actuelle est compensée par une formation détaillée notamment à ce propos. Cependant ce biais est présent pour le Grand Public qui ne met aucune de ces solutions en place pour le moment. Cela se répercute directement sur les données avec une proportion d'uniquement 21,4% de valeurs d'abondance nulle. Cette valeur devrait être très proche de VNE puisque les observations se font dans des zones d'observations similaires (voir section sur les classifications).

Jeux de données	GP		OAB	VNE
Protocole	Inventaire	Planche	Planche	Planche
Proportion d'observations sans mollusques	5,3%	21,4%	27%	34,3%

Figure 1: Proportion d'observations sans mollusques

Des observations sans mollusques sont donc manquantes pour les données GP mais ce biais est maintenant connu par Vigie Nature qui est en train de développer un nouveau formulaire de saisie pour Vigie Nature Ecole et le Grand Public demandant plusieurs taxons plus communs aux participants. Le protocole inventaire est moins touché par ce biais puisque le nombre d'observations par mois n'est pas limité : un participant n'ayant rien vu lors d'un de ses relevés en effectuera un autre pour compenser.

Comment représenter les données ?

Un des premiers objectifs de ce stage était de mettre en place des outils de visualisation de données. Ils doivent permettre de manière simple de générer des cartes lisibles et pertinentes sur les données issues d'écologie participative. De plus ces figures doivent pouvoir être enregistré en très bonne qualité dans le but d'être intégrées dans des articles à destination des participants mais aussi au sein de Vigie-Nature.

A l'aide du package tmap de R, j'ai réalisé une fonction générique qui représente des données quantitatives en fonction du département. Cette fonction prend en entrée un vecteur de données quantitatives (variable d'intérêt pour laquelle une représentation est souhaitée) et un vecteur des départements associés. Une carte de France où la valeur moyenne pour chaque département de la variable d'intérêt est affichée. Elle est associée à un zoom de la région parisienne. Pour évaluer la qualité de cette représentation l'utilisateur peut choisir d'afficher deux cartes complémentaires. La première représente le nombre de valeurs pour la variable d'intérêt par département et la deuxième la variance associée à ces valeurs. Finalement il peut choisir d'enregistrer ces cartes dans un format jpeg de bonne résolution. Cette fonction permet ainsi une représentation synthétique de l'information, un moyen rapide d'évaluer la précision des données représentées ainsi qu'une visibilité détaillée de la région parisienne. Cette fonction sera utilisée pour représenter les données de comptage.

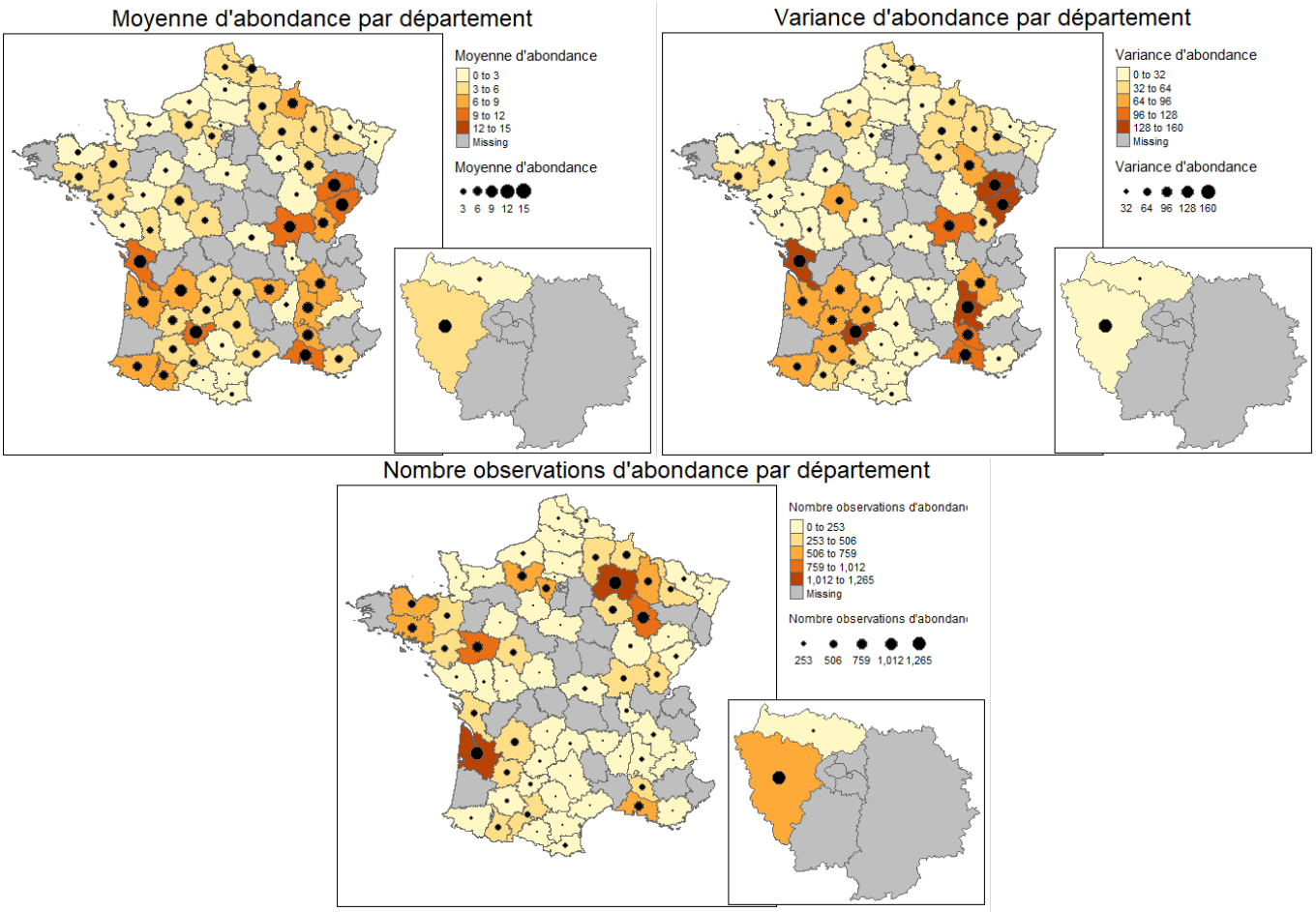

Figure 2 : Cartes issues de notre fonction à partir des données OAB

Les indices écologiques choisis sont-ils en phase avec nos connaissances ?

L'abondance de mollusques est-elle bien observée ?

Il existe des différences d'abondance pour les mollusques terrestres en France notamment en fonction de la nature de la roche mère (Barker, 2001). Des zones plus calcaires sont souvent associées à une plus forte concentration de mollusques qui utilisent cet élément pour leur coquille. Les différences d'abondance seront étudiées au niveau géographique et la proportion locale de calcaire sera utilisée dans nos futurs modèles.

Le nombre trop faible de nos observations ne permet pas d'afficher précisément l'abondance totale de mollusques à un niveau plus précis que celui du département (trop peu d'observations par commune par exemple). Une carte simplifiée des calcaires de France par département est générée afin d'être comparée aux cartes générées par notre fonction. Chaque département de France est associé à sa proportion de zone calcaire à partir de shapefiles des départements (Données publiques françaises, s. d.), de shapefiles des roches mères calcaires (BRGM, s. d.) et des fonctions st_intersection et st_area du package sf. Les zones ayant plus de 50% de calcaire sont considérées comme calcaires. Afin de valider cette méthode la perte de précision de cette simplification est calculée. Pour chaque point de grilles régulières de tailles variables (recouvrant la France) le taux d'erreur (zone prédite calcaire avec notre simplification alors qu'elle ne l'est pas au niveau GPS) est calculé. Pour une résolution de 500x500 points ce taux est de 11%. Cela reste élevé mais étant le meilleur niveau de précision pour ces données cette méthode est conservée.

Simplification par commune	92%
Simplification par code postal	89%
Simplification par département	80%

Figure 3 : Précision de l'information calcaire après simplification (résolution 500)

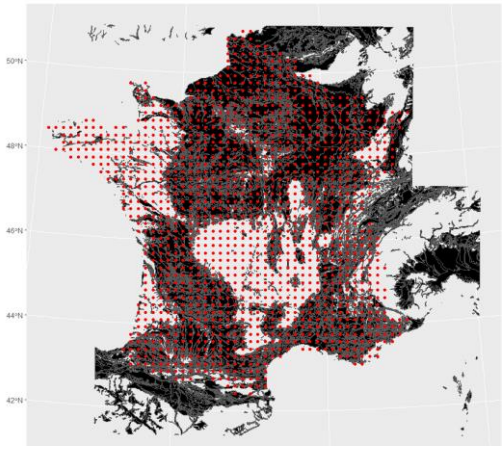

Figure 4 : Grille de résolution 100 sur la carte de France

Obtenant par la suite les coordonnées GPS pour chacune des observations une analyse de variance est mise en place. Pour associer une observation avec la nature de la roche mère (calcaire ou non) une fonction qui, à partir de coordonnées GPS, vérifie s'ils sont ou non présents sur un shapefile correspondant aux zones calcaires de France à l'aide des fonctions st_join du package sf) est codée. Une analyse de variance pour chacun des jeux de données et protocoles essayant de prédire l'abondance (par unité de surface pour le protocole planche) en fonction de la nature de la roche mère est ensuite mise en place. Le passage en densité n'affecte ni la tendance ni la significativité.

Jeux de données	GP	OAB	VNE
Signe du coefficient associé à une roche calcaire	+	+	-
p_value correspondante	0.130345	<2e-16	0.433

Figure 5 : Effet de la roche mère (calcaire ou non) sur la densité de mollusques

Les tendances sont conformes à nos attentes pour les données issues du Grand Public et de l'Observatoire Agricole pour la Biodiversité mais l'effet n'est significatif que pour ce dernier. Etrangement la tendance semble être inversée pour les données de Vigie-Nature Ecole même si l'effet n'est pas significatif.

Ces résultats montrent que les prédictions de cet indice écologique seront sans doute difficiles notamment pour les données VNE. A ce stade l'abondance de mollusques semble être mieux mesurée avec les données issues de l'OAB.

Les participants sont-ils capables de différencier les groupes morphologiques d'escargots ?

Dans un second temps la capacité des participants à différencier les groupes morphologiques observés à l'aide de la clé de détermination fournie dans les outils de Vigie-Nature (informations sur le site web, poster, livrets…) est évaluée. A partir de cartes de références (Kerney, 2015 ; INPH, s. d.) et des connaissances d'un expert (B. Fontaine - malacologue) les groupes morphologiques ayant une aire de répartition particulière en France sont sélectionnés. Le taxon *Elona quimperiana* (escargot de Quimper) par exemple se trouve uniquement en Bretagne et dans les Pyrénées Atlantiques. Ces cartes de référence sont ensuite comparées avec celles générées par notre fonction.

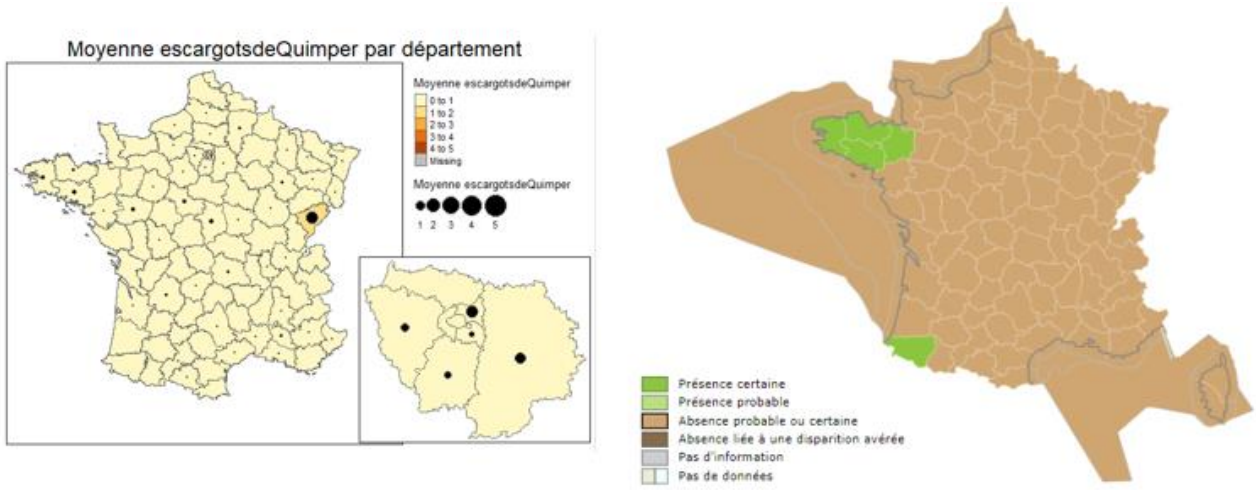

Figure 6 : Comparaison de la répartition d'une espèce de référence : l'escargot de Quimper avec les données Grand Public

La comparaison de ces cartes montre que les abondances enregistrées par les participants ne sont pas en phase avec la réalité. Cette comparaison est complétée avec des tests de moyennes entre les abondances dans la zone de répartition de l'espèce et celles dans le reste de la France. Cette différence d'abondance n'est significative dans aucun de nos jeux de données.

Les observateurs dans la plupart des cas ne peuvent associer un escargot avec son groupe morphologique correctement. Une étude détaillée pour chacun de ces groupes n'est donc pas envisageable. Ce résultat est très important et sera au cœur des améliorations du programme.

Pour corriger ces erreurs d'identification, Vigie-Nature va permettre à ses participant de soumettre les photos de leurs observations afin de bénéficier d'une vérification par des pairs et des experts. L'objectif est d'aider les observateurs à se perfectionner. Ce système est implémenté et en cours de test avec le Suivi photographique des insectes pollinisateurs, un autre observatoire de Vigie-Nature.

Nous partons cependant du principe que les observateurs sont capables de compter le nombre de groupes morphologiques différents observés mais ne peut pas dire avec certitudes lesquels il a observé. Cette hypothèse forte permet pour ce taxon de pouvoir étudier l'impact du protocole sur les résultats (le protocole inventaire permet d'obtenir uniquement la diversité totale).

Les zones d'observation sont-elles similaires entre nos jeux de données ?

Cette partie s'intéresse à la description des lieux d'observations par les participants. Les variables utilisées pour décrire ces zones sont-elles similaires entre observatoires ?

Cette étude concerne dans notre cas uniquement les données de GP et VNE puisque les observations de l'OAB sont réalisées uniquement dans des champs. La description des zones d'observation se présente sous la forme de présence ou absence de certaines variétés de plantes ou aménagements (booléens). L'utilisateur doit cocher ou non les cases correspondantes. Pour ces deux jeux de données des informations sur les mêmes éléments de la zone (les formulaires de saisie pour la partie description de la zone sont identiques) sont disponibles.

Dans un premier temps l'impact de certaines plantes sur les marqueurs biologiques est étudié. Pour cela un modèle linéaire simple prenant toutes les variables de description de la zone d'observation pour expliquer l'abondance et la diversité totale est ajusté. Ces résultats seront étudiés en dernière partie.

Une analyse en composantes principales à partir de ces données (les booléens étant transformés en variables quantitatives) pour chaque observatoire est ensuite effectuée. L'ACP est choisie plutôt que l'Analyse en Composantes Multiples par choix personnel d'interprétabilité des variables après avoir essayé les deux méthodes, l'espace sélectionné étant le même (à une symétrie près). L'objectif de cette analyse est de voir si l'espace représenté est le même pour les deux observatoires mais surtout le lien entre les variables et notamment si certaines plantes seraient en lien avec l'abondance ou la diversité totale. Pour ce faire ces deux indices écologiques sont ajoutés en variables explicatives. D'autres variables peuvent permettre une meilleure interprétabilité telles que la naturalité (somme d'éléments représentatifs de la nature : lierre, ronces, orties, espaces non entretenus/friches) et la distance aux environnements proches connus (bois, champ cultivé et prairie). Elles sont donc ajoutées en illustratif.

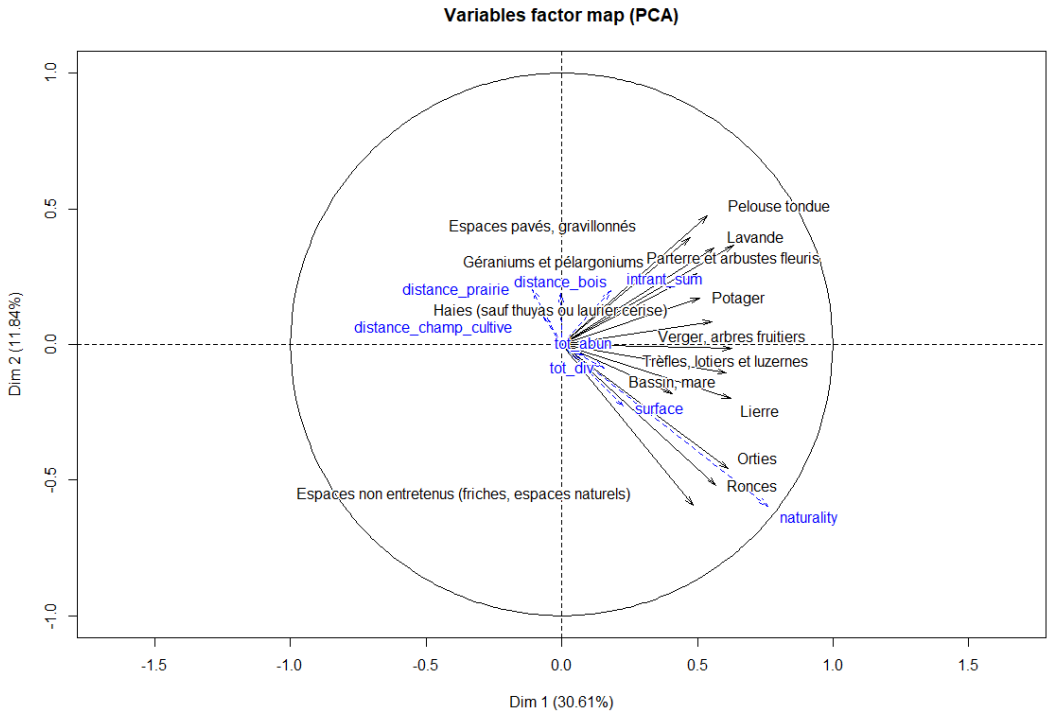

Figure 7 : ACP composition des jardins - données GP

Les 2 premiers axes de l'ACP expriment 42,44% de l'inertie totale du jeu de données ce qui explique en partie le fait que les variables ne soient pas parfaitement représentées sur cette figure mais qui permet d'avoir une explication pertinente de la variabilité des individus. Un axe de naturalité sur les deux premières dimensions se démarque nettement avec des jardins ayant plus de plantes sauvages positionnés en bas à droite (valeurs fortes sur l'axe 1 et faibles sur l'axe 2). A l'opposé un axe lié à l'anthropisation du jardin avec la pelouse tondue, les espaces gravillonnés et les arbustes fleuris est visible. Les jardins avec des aménagements se situent alors en haut à droite (valeurs fortes sur l'axe 1 et 2).

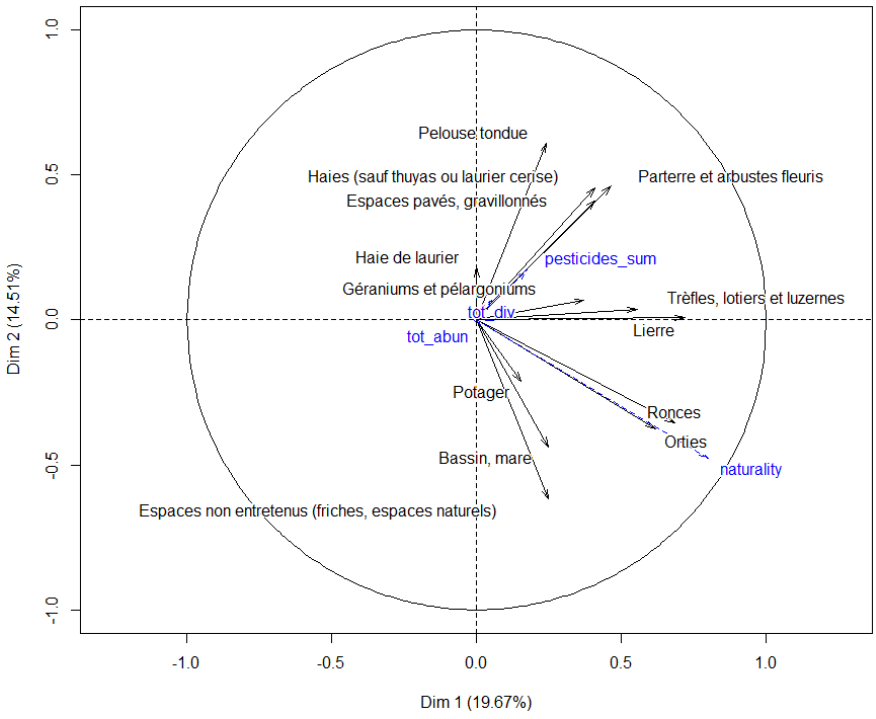

Figure 8 : ACP composition des jardins - données VNE

Les 2 premiers axes de l'ACP expriment 34,18% de l'inertie totale du jeu de données. Cette différence avec l'ACP de GP peut s'expliquer notamment par le plus faible nombre de jardins. Cependant une structure similaire avec le même axe de naturalité et d'aménagement est observée. Les indices écologiques d'abondance et de diversité sont ici aussi mal représentés. La position de la variable « bassin, mare » démarque ici puisqu'elle est plutôt associée à la naturalité qu'aux aménagements. Il y aurait sans doute plus de mares que de bassins dans les zones d'observation VNE contrairement aux jardins GP. En effet, dans le cadre de l'enseignement, la mise en place de mares pédagogiques est une activité courante.

Les descriptions des zones d'observations semblent être très similaires entre ces deux publics avec des analyses identiques. Certaines variables telles que les Orties, Ronces et Lierre sont plus associées à un jardin avec des zones sauvages. « Pelouse tondue », « Arbustes Fleuris », « Espaces Pavés » témoignent de l'entretien du jardin et de son anthropisation. Contrairement aux résultats attendus la présence de ces plantes sauvages ne semble pas être incompatible à un jardin très entretenu (pas de corrélation).

Est-il possible de synthétiser les informations de description de la zone d'observation ?

Est-il possible d'obtenir des « jardins types » à partir de la description des zones d'observation ? L'objectif de cette synthèse serait double : dans un premier temps de pouvoir communiquer au public ces archétypes de jardins afin de les aider dans la description des zones d'observation mais aussi de pouvoir avoir une variable synthétique de cette composition pour nos analyses.

Pour ce faire une classification ascendante hiérarchique est effectuée à partir des ACPs précédentes à l'aide de la fonction HCPC du package FactoMineR. Le nombre de classes choisi est celui conseillé par la fonction.

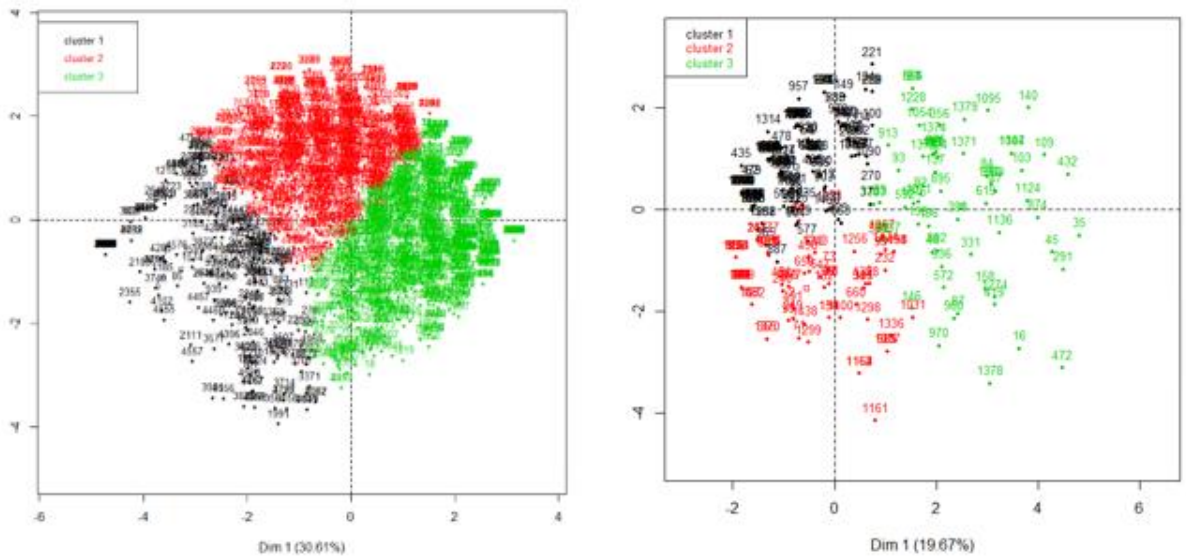

Figure 9 : Classification des jardins (GP à gauche et VNE à droite) sur les axes 1 et 2 de leurs ACP respectives

Les jardins sont divisés en trois classes. Une des classes (cluster 2 en rouge pour GP et cluster 1 en noir pour VNE) est constituée de jardins très entretenus caractérisés par une naturalité très faible et donc l'absence de plantes sauvages. Ces jardins sont aussi caractérisés par la présence d'aménagements tels que des espaces pavés ou des parterres de fleurs. Ces jardins correspondraient donc à des jardins d'ornement avec peu d'espaces naturels.

Le deuxième groupe de jardin (cluster 3 en vert pour GP et VNE) au contraire est caractérisé par une naturalité plus forte et donc la présence de plantes sauvages (lierre, orties, ronces…). Les variables correspondantes à l'aménagement sont aussi présentes. Ces zones d'observations peuvent donc être considérées comme des jardins plus sauvages.

La dernière classe (cluster 1 en noir pour GP et cluster 2 en rouge pour VNE) quant à elle est plus difficile à interpréter. Les jardins qui la composent ne sont pas différenciés par l'axe de la naturalité mais ils sont au contraire caractérisés par l'absence d'aménagement. Ils sont aussi caractérisés par une surface plus petite que la moyenne. Ces zones d'observations pourraient alors correspondre à des petits jardins urbains.

Pour les deux jeux de données la classification est similaire. Dans les deux cas ce sont trois classes qui sont retenues et qui s'interprètent de la même manière. Les données issues de la description de la zone d'observation sont donc très similaires et une structure de jardins similaire qui a du sens ressort.

Ces archétypes de jardins sont réutilisés dans un article de vulgarisation à destination des participants aux programmes. Cette information leur permettra à terme de les aider dans la description de leur zone d'observation.

Nos indices écologiques sont-ils expliqués par les mêmes facteurs ?

Des variables en lien avec ces indices peuvent-elles être ajoutées à partir de nos données ?

Notre objectif est maintenant de savoir si l'abondance et la diversité totale peuvent être expliquées de manière fiable à partir des informations à notre disposition. Pour ce faire des variables qui pourraient avoir un effet sur ces indices écologiques sont d'abord ajoutées à nos jeux de données. Comme expliqué précédemment la nature de la roche mère (calcaire ou non) est importante et a donc été ajoutée ainsi que d'autres facteurs d'importance non présents initialement.

L'**occupation des sols de la commune** (extraite à partir des données Corinne Land Cover - CLC) peut être représentative de l'environnement à une plus grande échelle que la zone d'observation (Kernay, 2015). Cette information est issue d'observations satellite converties en images vectorielles. Ces images sont associées à des polygones référencés dans l'espace correspondant chacun à un type d'occupation des sols. Il existe cinq types correspondant donc à cinq variables contenant respectivement le pourcentage de sol occupé par des « Terrains artificialisés » correspondant aux zones urbaines (CLC1), « Territoires agricoles » (CLC2), « Forêts et milieux semi-naturels » (CLC3), « Zones humides » (CLC4) et « Surfaces en eau » correspondant aux lacs et océans (CLC5). Les deux premiers indices étant très corrélés (>|0.8|) seul le pourcentage de zone urbaine de la commune (CLC1) est conservé.

A l'aide de la fonction intersection du package sf et des coordonnées des observations chaque participation est associée à une commune (shapefiles : Données publiques françaises, s. d). Une table de correspondance entre la commune et les CLC (Données publiques françaises, s. d.) permet d'ajouter l'occupation des sols à nos jeux de données.

L**'état de la planche** peut aussi avoir une influence sur son attractivité envers les mollusques. Une planche en très bon état faite de bois neuf n'a pas la même influence sur les observations qu'une planche plus abimée notamment par l'humidité (Barker, 2001). Cette variation d'état de la planche peut être approchée par son âge. C'est-à-dire depuis combien de temps elle est placée dans la zone d'observation. L'attractivité de la planche augmenterait pendant les premières semaines jusqu'à un plateau correspondant à une planche suffisamment modifiée par les conditions du milieu. Les protocoles de Vigie Nature stipulent que les premières observations doivent être effectuées quelques semaines après avoir placé la planche. Une vérification du respect de cette indication est effectuée. L'âge de la planche n'est cependant pas présente dans nos jeux de données.

Pour l'approcher la première observation est considérée comme étant l'âge 0 de la planche. Pour chacune des observations suivantes la différence de temps entre cette observation et la première est considérée comme étant l'âge de la planche. Des modèles linéaires et logarithmiques expliquant les indices écologiques en fonction de l'âge de la planche ainsi calculé sont ensuite ajustés.

Pour les trois jeux de données l'effet de cette variable n'a pas été significatif. Nous sommes cependant conscients de la limite de cette variable reconstituée puisqu'il n'y a pas toujours le même laps de temps entre la date de pose de la planche et celle de la première observation. Dans les futurs formulaires de saisie cette information devrait être ajoutée. Cette variable étant biaisée et son impact n'étant pas significatif elle a été exclue de l'analyse.

Mise en place de modèles de prédiction

L'objectif de cette partie est de mettre en place des modèles de prédiction de nos indices écologiques afin de voir si les publics, les protocoles ou l'indice prédit ont un impact sur les variables sélectionnées.

Dans un premier temps une réflexion est menée autour du type de modèle à utiliser pour prédire ces indices écologiques. Les variables d'intérêts sont issues de donnée de comptage ce qui oriente vers l'utilisation de la loi de poisson très souvent utilisée pour ce type de données. De plus la distribution de nos indices écologiques se rapproche fortement d'une loi de poisson et ce pour plusieurs de nos jeux de données. Le modèle linéaire sera utilisé comme référence de comparaison. L'utilisation d'autres types de modèles pourrait être envisagée pour expliquer nos indices écologiques mais ne sera pas abordée lors de ce stage par faute de temps.

Distribution des indices écologiques

Afin de ne garder que les variables ayant une influence sur les indices écologiques une sélection de variable est mise en place au sein de nos modèles. Cela permet de ne faire aucun *a priori* sur l'influence de certains facteurs tout en évitant de conserver des variables n'ayant aucun rapport avec nos indices. La première étape de cette sélection est d'enlever les variables corrélées entre elles. Déjà effectuée pour les données d'origine, cette démarche est réeffectuée avec nos nouvelles variables. Par exemple CLC1 et CLC2 qui correspondent respectivement à la part d'espace urbain et agricole dans le paysage de la commune d'observation sont fortement corrélés (>|0,7|). Le fait de ne pas laisser le modèle choisir laquelle des deux variables corrélées qui sera sélectionnée permet de garder le facteur contenant l'information la plus précise (en fonction de la formulation de la question sur le site) ou le plus interprétable d'un point de vue écologique. Dans un deuxième temps une sélection pas à pas basée sur le critère d'Akaike et la direction backward est mise en place. Ces deux critères permettent de sélectionner des modèles ayant plus de variables qu'avec le critère Bayésien et donc d'avoir des informations sur un plus

grand nombre de variables. Un compromis entre la qualité du modèle et le nombre de variables sélectionnées est donc adopté.

Tous ces modèles seront ajustés avec la fonction glm du logiciel R. Pour avoir une idée précise de la qualité des modèles tout en évitant le surajustement, une validation croisée est mise en place. Dix modèles sont ajustés en prenant à chaque fois neuf dixième des observations du jeu de données pour ajuster un modèle. Chacun des modèles est utilisé pour prédire les indices écologiques du dixième d'observations restantes (méthode 10_fold). Cette démarche est effectuée à l'aide des fonctions du package caret de R.

L'objectif est alors de comparer la qualité d'ajustement de nos modèles et de déterminer l'influence de plusieurs éléments sur la prédiction des indices écologiques. La simplification de la zone d'observation en trois classes diminue-t-elle la qualité de prédiction ? Est-ce que le modèle de poisson est plus adapté qu'un modèle d'analyse de covariance pour nos données ? Y a-t-il des différences sur ces résultats en fonction de l'indice écologique prédit (abondance ou diversité), du protocole ou du type de participant ?

Pour répondre à ces questions plusieurs indices de précision des modèles sont utilisés. La comparaison des modèles linéaires ou issus de la loi de poisson entre eux peut être effectuée en utilisant le coefficient de détermination ou la valeur d'AIC cependant ceux-ci n'étant pas calculés de la même manière pour ces deux types de modèles un indicateur identique pour ces deux familles était nécessaire. L'erreur absolue moyenne (MAE) qui correspond à la somme de la valeur absolue de chacun des résidus du modèle ainsi que l'erreur quadratique moyenne (RMSE) qui correspond à la racine carrée de la MCE (moyenne du carré des écarts) sont donc choisis. Ces deux indices sont très représentatifs de la qualité de précision du modèle tout en évitant le surajustement puisque les prédictions sont issues de notre validation croisée (méthode des 10 segments).

- <u>Protocole Inventaire</u>

Dans un premier temps le protocole inventaire est étudié. Il ne concerne que les données Grand Public. Plusieurs fois par mois les participants marchent dans leur jardin par temps de pluie et notent l'effectif de chacun des groupes morphologiques de mollusques rencontrés. A la fin du mois s'ils ont réalisé cet inventaire au moins une fois ils doivent remplir le formulaire de saisie en ligne comprenant en plus de ces indices écologiques des données de description de la zone d'étude. Une observation correspond donc à un ensemble d'inventaire réalisé par le participant en un mois. Ce nombre n'est pas demandé dans les formulaires de saisies c'est pourquoi l'abondance de mollusque n'est pas du tout homogène entre les observations. Les modèles ajustés dans cette partie ne concerneront donc que la diversité totale observée. L'hypothèse que le nombre d'inventaire réalisé dans le mois n'influence que très peu le nombre d'espèces observées est émise. Elle semble raisonnable d'un point de vue écologique. De plus aux vues des abondances de ce jeu de données le nombre d'inventaire réalisé par mois semble être en moyenne de 1 ou 2. Ces modèles se basent sur un total de 4121 observations.

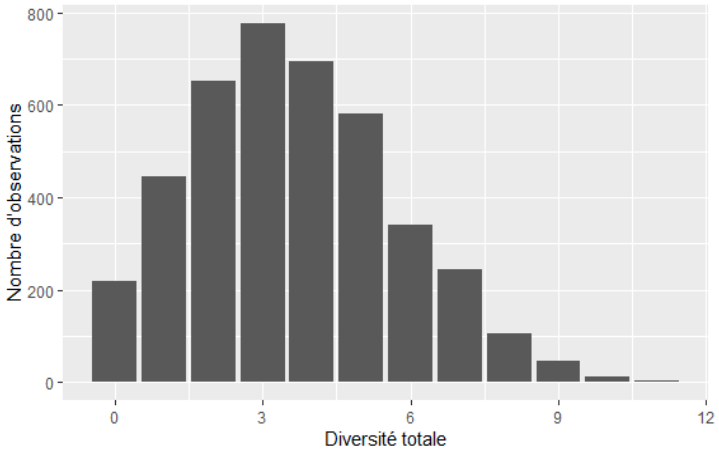

Figure 10 : Répartition de la diversité totale – Grand Public – Protocole Inventaire

type_model	zone_observation	RMSE	Rsquared	MAE	AIC
Poisson	Classe de jardin	3.095769	0.1330157	2.521050	16828.95
Poisson	Description zone	3.091495	0.1541701	2.518061	16730.28
Linear	Classe de jardin	1.922822	0.1448265	1.543533	17082.83
Linear	Description zone	1.903367	0.1621819	1.521100	16988.47

Figure 11 : Précision des prédictions - Grand Public – Protocole Inventaire

Les indicateurs de précision indiquent une précision très faible (RMSE, MAE et AIC très élevés tandis que le coefficient de détermination est proche de 0) pour nos modèles. Ils ne permettent donc pas de prédire de manière fiable la diversité totale observée.

Les modèles de poissons ont tendance à prédire avec une plus grande variabilité ce qui entraine des erreurs moyennes plus fortes (différence de MAE de quasiment 1 et de 1,2 pour la RMSE). Pour ces données le modèle linéaire semble prédire avec une meilleure précision la diversité totale. Ces résultats sont en phase avec la distribution de notre variable d'intérêt qui s'approche d'une loi normale.

Pour les deux familles de modèle la synthèse de la zone d'observation en trois classes diminue légèrement la qualité du modèle (baisse du coefficient de détermination) mais pas de manière significative (à revoir explication).

- Protocole Planche

Dans un second temps le deuxième type d'observation de notre jeu de données : le protocole planche est étudié. Les observateurs placent une planche dans leur zone d'observation puis après quelques semaines, ils soulèvent la planche régulièrement pour compter l'abondance de tous les groupes morphologiques observés. Avant de pouvoir expliquer ces indices écologiques l'impact de la surface de la planche (qui ne concerne évidemment que les données issues de ce protocole) est questionné. Cette variable peut être prise en compte de différentes manières.

Aucune hypothèse ne peut tout simplement être faite et le choix de sélectionner ou non cette variable est laissée au modèle. Si c'est le cas la relation entre l'abondance et la surface n'est pas la même en fonction de la famille du modèle utilisé. Cependant dans beaucoup de publications en écologie l'abondance totale est convertie en densité ramenant ainsi le nombre d'escargots à la surface de la planche. L'hypothèse d'une relation de proportionnalité entre la surface de la planche et le nombre de mollusques observés (plus la planche est grande et plus le nombre d'escargots observés sera élevé) est alors émise. La régression de poisson nécessite

obligatoirement des valeurs entières. Pour garder la relation de proportionnalité une adaptation du modèle est nécessaire : la surface de la planche est remplacée par son logarithme auquel est associé un coefficient fixe de 1 (Schaeffer, 2012).

Notre démarche est de laisser ces deux possibilités et de comparer les modèles en résultant. Les variables des modèles étant les plus proches des données seront alors étudiées.

- Grand Public

Dans un premier temps les données issues du Grand Public sont étudiées. Les modèles issues de cet observatoire se basent sur 662 observations.

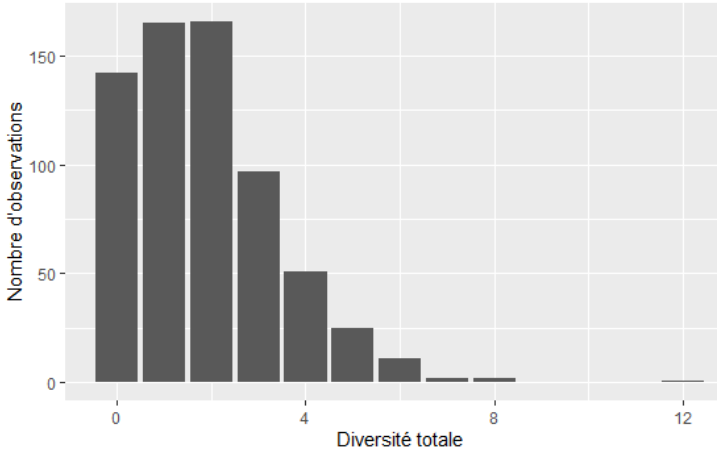

Figure 12 : Répartition de la diversité totale – Grand Public – Protocole Planche

type_model	zone_observation	variable_surface	RMSE	Rsquared	MAE	AIC
Poisson	Classe de jardin	imposée	2.424028	0.0014202055	1.809578	3059.903
Poisson	Classe de jardin	libre	2.009296	0.0207663991	1.518868	2299.683
Poisson	Description zone	imposée	2.407436	0.0006227917	1.815222	3001.037
Poisson	Description zone	libre	2.004275	0.0342487626	1.522199	2280.599
Linear	Classe de jardin	imposée	8.214189	0.0109442469	5.349262	4632.279
Linear	Classe de jardin	libre	2.009296	0.0207663991	1.518868	2299.683
Linear	Description zone	imposée	8.177753	0.0258522305	5.425194	4615.503
Linear	Description zone	libre	1.567448	0.0367276995	1.195061	2440.410

Figure 13 : Précision des prédictions - Grand Public – Protocole Planche

Les modèles ajustés ici sont eux aussi très loin de la réalité (RMSE, MAE et AIC très élevés tandis que le coefficient de détermination est proche de 0).

La diversité totale ne semble pas être influencée par la taille de la planche puisque les modèles prédisant cet indice écologique par unité de surface ont des précisions beaucoup plus faibles. De plus si la variable n'est pas imposée, elle est systématiquement supprimée au cours de la sélection de variables. Tout comme pour le protocole planche la description de la zone ne change pas significativement la qualité du modèle.

Pour les données Grand Public le protocole ne semble pas changer nos conclusions. Dans les deux cas, la prédiction de la diversité totale est très approximative. Le modèle linéaire semble être plus proche des données et la synthèse de la zone d'observation ne change pas la précision du modèle. La surface de la planche ne semble pas avoir d'impact sur la diversité totale. Ces données sont cependant soumises à plusieurs biais comme la présence plus faible d'observations sans escargots (diversité nulle) et l'absence du nombre d'inventaires réalisés.

- Vigie Nature Ecole

Dans un deuxième temps les données issues de Vigie Nature Ecole sont étudiées. Seul le protocole planche est utilisé pour ces observations et la zone d'observation peut être simplifiée à l'aide de notre classification des jardins. Dans ce jeu de données une observation correspond bien à un seul relevé de planche. L'abondance de mollusques en plus de la diversité peut alors être étudiée. Ces modèles sont ajustés sur 1386 observations.

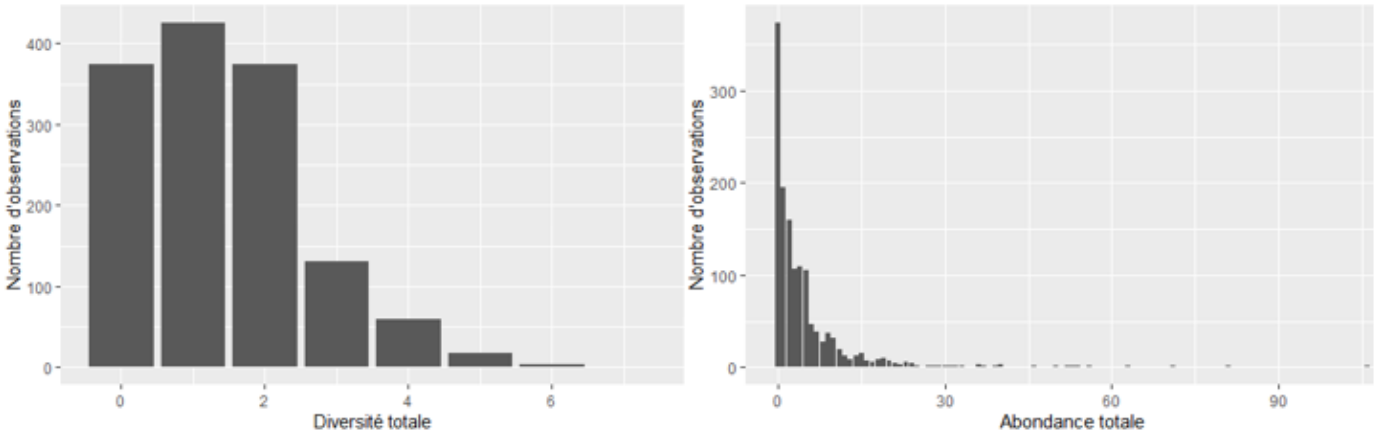

Figure 14 : Répartition des indices écologiques – VNE – Protocole Planche

type_model	zone_observation	variable_surface	RMSE	Rsquared	MAE	AIC
Poisson	Classe de jardin	imposée	1.579782	0.04923559	1.2093785	4040.883
Poisson	Classe de jardin	libre	1.599456	0.09022204	1.2331532	4041.662
Poisson	Description zone	imposée	1.616816	0.09280107	1.2545927	3982.566
Poisson	Description zone	libre	1.594375	0.13523852	1.2393775	3984.437
Linear	Classe de jardin	imposée	6.857481	0.05418964	4.9684443	9249.187
Linear	Classe de jardin	libre	1.153918	0.08800197	0.9119734	4309.110
Linear	Description zone	imposée	6.630539	0.11701893	4.8356927	9151.992
Linear	Description zone	libre	1.127006	0.13041607	0.8881579	4247.889

Figure 15 : Précision des prédictions de diversité - VNE – Protocole Planche

type_model	zone_observation	variable_surface	RMSE	Rsquared	MAE	AIC
Poisson	Classe de jardin	imposée	8.581947	0.03673437	4.137790	12620.842
Poisson	Classe de jardin	libre	8.536096	0.06928292	4.094081	12552.693
Poisson	Description zone	imposée	8.563961	0.03971974	4.131937	12315.409
Poisson	Description zone	libre	8.532423	0.07659441	4.092554	12157.933
Linear	Classe de jardin	imposée	36.994275	0.03899691	21.166302	13936.095
Linear	Classe de jardin	libre	7.794936	0.05697500	4.441209	9586.559
Linear	Description zone	imposée	36.778789	0.05497750	20.707860	13897.490
Linear	Description zone	libre	7.771627	0.06491437	4.393550	9564.484

Figure 16 : Précision des prédictions d'abondance - VNE – Protocole Planche

La qualité de prédiction des indices écologique est très faible. Que ce soit les modèles d'abondance ou de diversité le modèle de poisson entraine des résidus plus forts. La conversion de ces indices en densité entraine aussi une grosse perte de précision. Il n'y a pas de relation de proportionnalité entre nos indices et la surface de la planche. Cependant la surface de la planche est systématiquement sélectionnée dans les modèles. Finalement la simplification de la description de la zone d'observation en trois classes ne modifie pas significativement les qualités prédictives des modèles.

- Observatoire Agricole de la Biodiversité

Finalement les données issues de l'Observatoire Agricole de la Biodiversité sont étudiées. Elles concernent uniquement le protocole planche. Cependant les planches utilisées sont toutes standardisées et leur surface est donc toujours identique. Contrairement aux deux jeux de données précédent, la zone d'observation n'est pas un jardin et ne correspond donc pas à nos trois classes. Les modèles sont ici ajustés à partir de 16 597 observations.

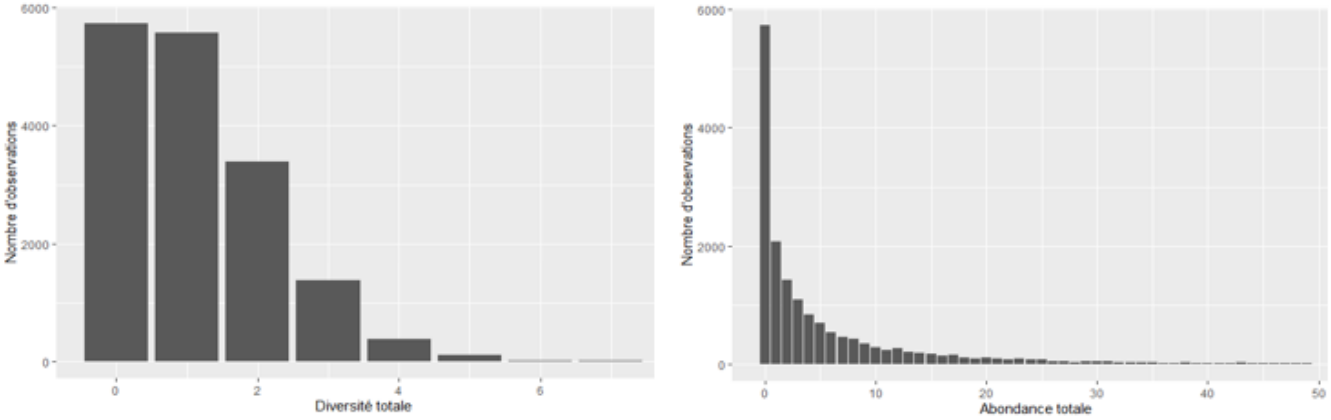

Figure 17 : Répartition des indices écologiques – OAB – Protocole Planche

type_model	RMSE	Rsquared	MAE	AIC
Poisson	1.489679	0.1476022	1.1516178	43580.70
Linear	1.016775	0.1501452	0.8038439	47638.44

Figure 18 : Précision des prédictions de diversité totale – OAB – Protocole Planche

type_model	RMSE	Rsquared	MAE	AIC
Poisson	8.773386	0.1112494	4.695750	169152.7
Linear	7.640441	0.1156476	5.094605	114590.8

Figure 19 : Précision des prédictions d'abondance totale – OAB – Protocole Planche

Tout comme pour les données issues du Grand Public et de Vigie Nature Ecole ces modèles ne sont pas capables de prédire correctement nos deux indices écologiques (abondance et diversité). Les modèles linéaires sont plus précis pour la prédiction de la diversité.

Tous les modèles ajustés ont beaucoup de mal à prédire nos indices écologiques. Cela rend nos conclusions sur l'impact de nos différents facteurs (protocole, famille…) très difficile. Cela peut s'expliquer en partie par la taille de notre jeu de données qui est plutôt faible mais aussi par des biais évoqués plus tôt. De plus ces données sont issues d'observations dans des environnements non standardisées ce qui implique sans doute l'absence d'informations importantes pour décrire ces zones d'observations.

La surface de la planche ne semble pas avoir de relation proportionnelle avec nos indices écologiques et les modèles de poissons ont des résidus globalement plus forts contrairement à nos attentes.

Finalement la classification des jardins semble pouvoir résumer la description des zones d'observation.

Dans un second temps l'impact des variables sélectionnées par nos modèles est étudié afin de savoir si elles ont un sens biologique. Si c'est le cas cette faible précision peut potentiellement être expliquée par le manque de variables d'intérêt. Sinon nous pourrons conclure que les variables utilisées n'ont pas de sens et sont donc potentiellement responsables de cette faible précision.

Quelles sont les variables sélectionnées par nos modèles ?

Les variables sélectionnées dans nos modèles sont maintenant étudiées. L'objectif est de déterminer quelles variables permettent d'expliquer les différences d'abondance et de diversité de mollusques et quels sont leurs effets. La présence ou non de différences sur ces résultats en fonction de l'indice écologique prédit, du protocole ou du type de participant sera aussi étudiée.

Les modèles linéaires et ceux issus de la loi de poisson sélectionnent les mêmes variables avec des effets similaires pour prédire nos indices écologiques et ce pour chacun des jeux de données. De plus les modèles linéaires étant légèrement plus proches de la réalité seules les variables issues de ces modèles seront regardées. Enfin pour les données issues du protocole planche aucunes hypothèses de départ ne seront émises laissant la surface de la planche libre d'être utilisée ou non à partir de la sélection de variable de nos modèles. Les effets des variables retenues pour chacun de nos jeux de données seront étudiées en comparant les résultats obtenus avec nos connaissances écologiques (Barker, 2001 ; Kernay, 2015)

Dans un premier temps le jeu de données Grand Public sera étudié en comparant les variables sélectionnées avec les protocoles planches et inventaire puis elles seront comparées à celles issues des observations OAB et VNE avec le même protocole : planche.

	2.5 %	97.5 %
(Intercept)	1.635612e+00	2.3436674183
distance_bois	-3.936337e-04	-0.0001288588
distance_prairie	1.714291e-07	0.0002932184
naturality	7.039043e-02	0.2442334135
month7	-5.562327e-01	-0.0115727663
month10	-6.083801e-01	-0.0117044763
CLC3	-1.136044e-02	-0.0003695163
CLC4	-1.813219e-01	-0.0252249079
CLC5	-6.093192e-02	0.0002186036

Figure 20 : Coefficients de l'analyse de covariance (intervalle à 95%) – GP – Protocole Planche – Prédiction de la diversité

Conformément à nos attentes la diversité de mollusques est expliquée par l'environnement. La présence de bois à proximité augmente la diversité tandis que des prairies la diminue. Une grande naturalité (présence de plantes sauvages : orties, ronces ou lierre) augmente elle aussi la diversité d'escargots. Les indices Corine Land Cover ont des effets plus difficiles à interpréter : une proportion plus forte de forêts sur la commune ou de zones humides ou de lacs diminueraient très légèrement la diversité. La nature calcaire de la roche mère est absente ce qui est assez étrange. De plus la synthèse de la zone d'observation en trois classes n'a pas été retenue.

Comme escompté la période d'observation a aussi un impact sur la diversité totale. Le printemps est plus propice à une observation de groupes morphologiques différents. L'année d'observation n'a cependant pas été sélectionnée (sans doute à cause du faible nombre d'observation comparé au nombre de modalités).

La surface de la planche n'a pas été sélectionnée pour ce modèle et ne semble pas avoir d'influence sur la diversité de mollusques ce qui n'est pas surprenant. Nous supposons en effet qu'elle n'a d'effet que sur l'abondance.

Les variables sélectionnées pour prédire l'abondance sont exactement les mêmes.

	2.5 %	97.5 %
(Intercept)	-1.477945e+00	1.084673e+00
distance_bois	-2.705606e-04	-1.325156e-04
distance_prairie	3.405869e-05	1.974656e-04
distance_champ_cultive	-1.716949e-04	-3.817811e-05
limestoneTRUE	2.573392e-01	5.288010e-01
naturality	5.619548e-02	2.017752e-01
month4	2.266533e+00	4.804568e+00
month7	2.240511e+00	4.779204e+00
month10	2.248913e+00	4.791399e+00
year2010	-3.376818e-01	8.247839e-02
year2011	-4.459811e-01	-3.403570e-02
year2012	-2.700003e-01	1.400941e-01
year2013	1.673718e-02	4.231501e-01
year2014	-1.261331e-01	2.531478e+00
year2015	-1.727761e+00	-1.203023e+00
year2016	-1.584094e+00	-1.081221e+00
year2017	-1.554561e+00	-1.000218e+00
year2018	-1.587916e+00	-9.548869e-01
CLC3	2.018952e-03	7.832428e-03
CLC4	-6.645973e-02	-7.646381e-03
garden_classJardins Sauvages	1.768555e-01	6.108257e-01
garden_classJardins Vides	-9.615279e-03	4.055401e-01

Figure 21 : Coefficients de l'analyse de covariance (intervalle à 95%) – GP – Protocole Inventaire – Prédiction de la diversité

Plusieurs différences sont à noter entre les variables sélectionnées par le protocole planche et inventaire. L'impact des variables d'environnement ne sont pas exactement les même et semblent être plus en phase avec la réalité pour le protocole inventaire. Comme pour le protocole planche la proximité avec les bois entraine une plus forte diversité mais cette information est cette fois en phase avec la proportion de bois dans la commune (CLC3). Les champs cultivés à proximité semblent aussi aller dans ce sens. Cette fois ci la nature de la roche mère est sélectionnée : une roche calcaire augmenterait le nombre de groupes morphologiques observés. C'est aussi le cas pour les zones d'observations classées comme jardins sauvages puisque la synthèse des jardins est sélectionnée. Les booléens de description sélectionnés pour la zone d'observation ont les mêmes effets pour les deux protocoles.

Tout comme pour le protocole planche la période d'observation joue un rôle important. Le mois est sélectionné avec beaucoup moins d'escargots en hiver (période d'observation disponible uniquement pour le protocole inventaire). L'année est aussi sélectionnée avec des variations diverses ce qui n'était pas forcément attendu.

Les variables sélectionnées par ces modèles sont globalement en phase avec nos attentes. Cependant celles issues des données du protocole inventaire semblent être plus détaillées et proches de la réalité notamment au niveau de l'impact de l'environnement. Cette différence est sans doute liée à la quantité d'observations nettement supérieure dans le cas du protocole inventaire.

Ces résultats pour le protocole planche vont maintenant être comparées avec nos deux autres jeux de données.

	2.5 %	97.5 %
Humiditesec	-0.3383773	-0.27408415
MAUTRECULTURE	-0.1200569	-0.03024194
MPRAIRIE	-0.1350187	-0.03507035
MBOIS	-0.1323645	0.01700026
METANG	-0.5553019	-0.12703132
MAUTRE	-0.1373850	-0.03183402
TYPEPARCELLELIBAUTRE-CULTURE-PERENNE	-0.2282170	-0.04157815
TYPEPARCELLELIBGRANDE-CULTURE	-0.5053073	-0.38793936
TYPEPARCELLELIBMARAICHAGE	-0.1153107	0.05832622
TYPEPARCELLELIBPRAIRIE	-0.2793730	-0.15624305
TYPEPARCELLELIBVITICULTURE	-0.5433105	-0.41490515
conduiteBIOLOGIQUE	-0.2313299	-0.12391832
conduiteCONVENTIONNELLE	-0.1605707	-0.06284699

Figure 22 : Coefficients spécifiques aux données OAB (intervalle à 95%) – Protocole Planche – Prédiction de la diversité

Tout comme pour les données Grand Public les effets de l'environnement sont importants avec une plus grande diversité pour une roche mère calcaire et à l'inverse une diversité plus faible dans les zones humides. Un des intérêts de ce jeu de données est la présence de plusieurs variables particulières utilisées pour décrire l'environnement de la parcelle. Le milieu limitrophe (M) est représenté par plusieurs booléens indépendants dont seulement quelque uns n'ont pas été sélectionnés (bois et zone urbaine) sans doute car ils portent la même information que les variables CLC1 et CLC3. La variable correspondant au type de parcelle correspond tout à fait à nos attentes avec le maraîchage, l'arboriculture (ici en référence) et les autres cultures pérennes (correspondant à des types de cultures alternatives) qui voient une diversité plus forte que les grandes cultures, prairie ou viticulture. Comme escompté le type de conduite est sélectionné et même si contrairement à nos attentes, la différence entre le conventionnel et le biologique ne sont pas significatives, les conduites alternatives sont associées à une forte diversité.

Une autre variable est intéressante ici : l'humidité. C'est la seule variable de cette étude qui concerne la pluviométrie et elle est sélectionnée avec une augmentation de la diversité en cas de plus forte humidité.

Les effets des variables sélectionnées pour prédire l'abondance de mollusques sont, comme prévu, très similaires. Les seules différences notables sont l'influence des villes et des bois. La présence de bois dans la commune ou à proximité semble impliquer une plus forte diversité mais une plus faible abondance, ce qui est un résultat connu (Kernay, 2015). Les variables correspondant aux zones urbaines sont quant à elles sélectionnées uniquement pour prédire l'abondance : le nombre d'escargots observés sous les planches de champs est plus élevé près des villes ce qui pourrait être expliqué par l'absence d'autres abris à proximité.

	2.5 %	97.5 %
(Intercept)	0.747651750	1.546394e+00
annee_scolaire2015	-0.152224839	5.098844e-01
annee_scolaire2016	-0.251085801	3.616574e-01
annee_scolaire2017	-0.055806414	6.124977e-01
annee_scolaire2018	-0.228098662	3.871730e-01
annee_scolaire2019	-0.671594102	-6.953021e-02
annee_scolaire2020	-0.121009064	5.225601e-01
surface_planche	0.822194719	1.674092e+00
`Posée sous un arbre/un buisson`	0.061177466	3.484686e-01
`Posée contre un mur`	0.088775056	4.203714e-01
`Posée contre la terre nue`	0.162764259	5.582966e-01
`Posée sur le gazon`	0.225786703	5.130067e-01
`Posée sur un terrain en friche`	0.126768959	5.183420e-01
limestoneTRUE	-0.634437997	-2.225415e-01
CLC3	-0.006629772	7.096927e-05
CLC4	0.100383920	8.472193e-01
CLC5	-0.051622564	1.975028e-03
garden_classJardins Sauvages	-0.266406246	5.522445e-02
garden_classJardins Urbains	-0.017799018	3.382844e-01

Figure 23 : Coefficients spécifiques aux données VNE (intervalle à 95%) – Protocole Planche – Prédiction de la diversité

La prédiction de la diversité à partir des données VNE semble être plus difficile avec très peu de variables associées à des coefficients dont le signe est significativement différent de 0. La surface de planche est cependant présente associée à un fort coefficient positif : une planche plus grande entraînerait une plus grande diversité. Contrairement à nos attentes la nature calcaire du sol entraînerait une baisse du nombre de groupes morphologiques observés. Les classes de jardins ne semblent pas permettre de différencier correctement la diversité. De plus contrairement aux autres jeux de données, la proportion de zones humide augmenterait la diversité. Cependant très peu de communes sont concernées pour ce jeu de données.

Les modèles de prédiction de l'abondance de mollusques pour les données VNE sont plus proche des autres jeux de données. Ils sélectionnent par exemple la période de l'année avec une abondance de mollusques plus faible en hiver. Une zone ayant une naturalité forte entraînerait aussi une abondance supérieure. Cependant certains résultats sont étranges comme l'influence négative des sols calcaires et des bois sur l'abondance.

Une grande partie des variables sélectionnées dans nos modèles ont des effets conformes à nos attentes et donc en phase avec nos connaissances biologiques. Cette démarche est donc prometteuse. Cependant certains résultats sont très étranges surtout pour le jeu de données VNE qui est celui contenant le moins d'observation et qui est biaisé par une fort part d'écoles d'Île-de-France ou autres milieux urbains modifiant l'effet de la nature calcaire de la roche mère. C'est sans doute lié à la précision actuelle de nos modèles qui ne permet pas de conclusions définitives et invite à rechercher plus d'information. Ces informations manquantes sont sans doute contenues par certaines variables importantes. Cela peut être par exemple à des variables liées au climat (température, pluviométrie, ensoleillement…). De plus les données issues du protocole inventaire sont soumises à de gros biais au niveau du nombre différent d'observations par mois et de la trop faible part d'observations sans escargots. Certaines valeurs sont aussi très exagérées pour l'abondance totale ou la surface de la planche dans certains jeux de données et peuvent être la cause d'une augmentation significative de nos indices de précision (RMSE et MAE). Finalement l'utilisation de nombreuses variables inhérentes à certains jeux de données et la présence du protocole inventaire pour seulement un seul des jeux de données rends la comparaison difficile.

Conclusion :

Une méthode générique d'étude des données issues de l'écologie participative a été mise en place et appliquée pour un taxon donné : les mollusques terrestres. Cette méthode a pour but d'évaluer la qualité des données et de comparer les résultats obtenus en fonction de facteurs inhérents à cette discipline : l'indice écologique d'intérêt (abondance et diversité totales dans notre exemple), le protocole utilisé (inventaire ou planche pour les mollusques) et le public visé (grand public, milieu agricole, milieu scolaire pour ce taxon). Dans un premier temps les variables qui ne sont pas assez robustes (trop de données manquantes, corrélations, description peu précise dans les formulaires de saisie…) sont enlevées.

Plusieurs résultats marquants ressortent de l'étude de ce taxon. Tout d'abord une étude détaillée sur la diversité totale a montré que les participants ne sont pas capables de reconnaître les espèces. Ces erreurs peuvent intervenir au sein d'un même groupe (deux individus identiques notés différents) ou bien simplement dans le nom du groupe morphologique (deux individus identiques sont notés comme tels mais pas dans le bon groupe). Au contraire l'abondance totale semble être beaucoup plus robuste et donne des résultats cohérents. C'est pourquoi seuls les indicateurs fondés sur l'abondance totale et non sur des espèces en particulier devront être étudiés avec ces jeux de données. L'étude de plusieurs biais souvent retrouvés en écologie participative montrent ensuite que le niveau d'animation a un impact sur la qualité des données. En effet la proportion de données manquantes est plus élevée pour le grand public où les participants sont beaucoup moins accompagnés. La description des zones d'observation très détaillée a pu ensuite être synthétisée en trois classes identiques pour les zones non agricoles (ces dernières étant trop différentes). Ce qui montre une réelle similitude malgré la différence de public. De plus la synthèse de cette information peut permettre de gagner en compréhension et d'aider les participants dans leur description.

Finalement des modèles prédictifs de nos indices écologiques à partir de toutes les variables d'intérêt sont ajustés. La méthode de prédiction est étudiée (famille de modèle, sélection de variable, validation croisée) et permet d'évaluer les variables structurantes des jeux de données en lien avec les indices écologiques étudiés et de les comparer à nos connaissances biologiques. Quelques facteurs semblent avoir un impact important sur nos indices écologiques comme la zone d'observation avec notamment la naturalité et l'artificialisation qui sont à chaque fois sélectionnés ainsi que la nature calcaire ou non de la roche mère. Cette influence est aussi présente à une plus grande échelle avec les structures géographiques proches (CLC et distance aux champs, bois…). Finalement la période est souvent présente avec notamment le mois d'observation montrant une saisonnalité forte. Les publics sont similaires sur ces points malgré certains résultats étranges pour Vigie Nature Ecole. Cependant la comparaison des protocoles est compliquée du fait de la présence de la totalité d'entre eux pour un seul public.

Une harmonisation des protocoles, du niveau d'accompagnement des participants et des formulaires de saisie est conseillée et commence à être mise en place. Elle permettra des comparaisons plus robustes entre les différents publics.

Quelques points de cette étude peuvent être améliorés. Les extractions des bases de données étaient complexes et n'ont pas permis d'avoir toutes les variables sans traitement intermédiaire, lesquels n'étaient parfois pas reproductibles. De plus une étude plus détaillée sur les erreurs d'identification des participants doit être menée pour valider ou non l'utilisation de la diversité totale. Pour continuer ce travail nous pourrions donc imaginer l'utilisation d'autre taxons ou indices écologiques (présence ou absence de certaines espèces, dans certaines zones…). Un choix différent de modélisation dans la dernière partie pourrait aussi améliorer significativement nos modèles (utilisation d'autres familles, critère BIC…). L'utilisation de méthodes de traitement des données manquantes peut aussi constituer une voie d'amélioration.

Bibliographie :

BARKER, G. M., 2001. Gastropods on land: phylogeny, diversity and adaptive morphology. BARKER, G. M. (éd.), *The biology of terrestrial molluscs*. Wallingford : CABI. p. 146.

CHARONNET, Emmanuel, 2019. A la recherche des papillons perdus: Les naturalistes amateurs à l'épreuve des observatoires participatifs de la biodiversité. p. 731.

HOULLIER F., MERILHOU-GOUDARD J.-B., 2016. Les sciences participatives en France. État des lieux, bonnes pratiques et recommandations. Rapport élaboré à la demande des ministres en charge de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche

KERNEY, M.P., CAMERON, R.A.D., 2015. Escargots et limaces d'Europe [adaptation française Alain Bertrand]. Paris : Delachaux et Niestlé p.370.

SCHAEFFER, Mickaël, 2012. Standardisation dans un modèle de Poisson: modélisation et conséquences sur les prédictions. p. 61.

Sitographie :

Bureau de Recherches Géologiques et Minières, (s. d.), infoterre.brgm.fr

Données publiques françaises, (s. d.), data.gouv.fr

Inventaire National du Patrimoine Naturel, (s. d.), inph.mnhn.fr

Logiciels et packages :

Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. *https://CRAN.R-project.org/package=caret*

Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal, 10 (1), 439-446, *https://doi.org/10.32614/RJ-2018-009*

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL *https://www.R-project.org/*. R version 3.6.1 (2019-07-05).

Sebastien Le, Julie Josse, Francois Husson (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01

Tennekes M (2018). "tmap: Thematic Maps in R." _Journal of Statistical Software_, *84*(6), 1-39. doi:10.18637/jss.v084.i06 (URL: *https://doi.org/10.18637/jss.v084.i06*).

l'institut Agro agriculture • alimentation • environnement	Diplôme : Ingénieur Agronome Spécialité : Spécialisation / option : Science des données Enseignant référent : CAUSEUR David

Auteur(s) : PRUNEAU Julien Date de naissance : 13/09/1997	Organisme d'accueil : Muséum National d'Histoire Naturelle Adresse : 57 rue Cuvier 75005 Paris Maître de stage : BENATEAU Simon
Nb pages : 22 Annexe(s) : 0	
Année de soutenance : 2020	

Titre français : Etudes de données issues de l'écologie participative : qualité et influence des protocoles

Titre anglais : Data study from participatory ecology : quality and protocols influence

Résumé :

Les sciences participatives sont de plus en plus utilisées comme support de publications scientifiques. Cet intérêt grandissant est encouragé par l'enseignement qui met à l'honneur ces programmes. C'est pourquoi la qualité des données issues de la participation de non-scientifiques ainsi que l'impact des protocoles utilisés est un nouvel enjeu. Nous cherchons ici à mettre en place une méthode d'analyse de ce type de données. Cette démarche est appliquée aux programmes de suivi de la biodiversité portant sur les mollusques terrestres pour trois types de participants (grand public, écoles et agriculteurs). Des indices écologiques sont d'abord mis en évidence puis questionnés : l'abondance et le nombre de groupes morphologiques différents observés (diversité) pour notre taxon. La comparaison des résultats de ces comptages avec des données de référence issues de relevés scientifiques ou de connaissance d'experts montre que les participants sont capables de respecter un protocole mais que l'identification à partir de clés de détermination pour différencier les groupes morphologiques est plus difficile. Les biais les plus classiques liés à l'écologie participative sont ensuite étudiés comme le manque d'observations sans mollusques ou les erreurs de saisie. La comparaison de nos sources de données permet de réfléchir à des solutions : formations des participants ou présence d'autres taxons plus communs dans les formulaires. Les zones d'observations sont ensuite étudiées et semblent être similaires pour les observations dans les jardins. La dernière étape de notre analyse est la mise en place de modèles pour mettre en évidence les variables qui expliquent les indices écologiques choisis. Cette prédiction est compliquée pour le taxon étudié mais met cependant en valeur des tendances qui sont pour la plupart intéressantes et en phase avec nos connaissances écologiques.

Participatory sciences are used more and more as a support for scientific publications. This growing interest is encouraged by the education that highlight these programs. This is why the data quality from the participation of non-scientists as well as the used protocols impact is a new issue. We are trying here to set up a method to analyze this type of data. This approach is applied to biodiversity monitoring programs on terrestrial molluscs for three types of participants (general public, schools and farmers). Ecological indices are first highlighted and then questioned: the abundance and number of different morphological groups observed (diversity) for our taxon. These counts results comparison with reference data from scientific surveys or expert knowledge shows that the participants are able to respect a protocol but that the identification from determination keys to differentiate the morphological groups is more difficult. The most classic participatory ecology biases are then studied, such as the lack of snail-free observations or data entry errors. The comparison of our data sources allows us to think about solutions: participants training or other more common taxa presence in the forms. The observation areas are then studied and appear to be similar for gardens observations. The last step of our analysis is model establishment to highlight the variables that explain chosen ecological indices. This prediction is complicated for studied taxon, but nevertheless highlights trends which are for the most part interesting and in line with our ecological knowledge.

Mots-clés : Ecologie participative, Mollusques, Indices Ecologiques, Protocoles, Qualité des données

Key Words: Participatory sciences, Molluscs, Ecological indices, Protocols, Data quality