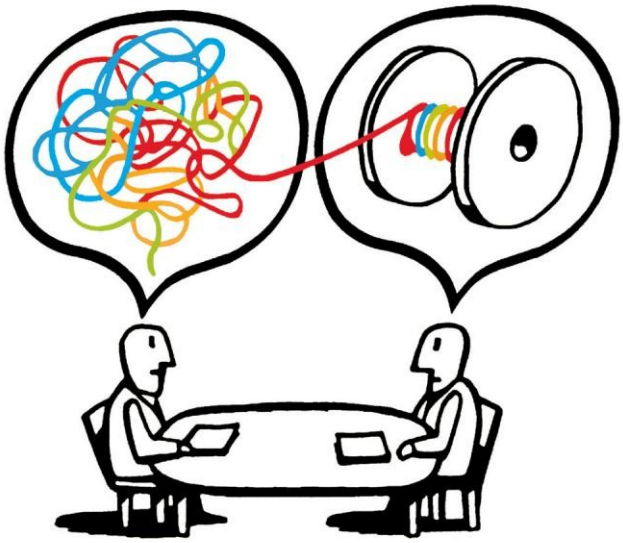
Analyse pratique de jeux de données massifs en écologie

Cours LBOE2121 - UCL - 2016

Gilles San Martin
(Centre Wallon de Recherches Agronomiques)



Shron 2014 - Thinking with data

Table des matières

Introduction

L'objectif de ce cours est la mise en pratique des méthodes statistiques vue dans d'autres cours (GLMs, Analyse multivariée, AIC) pour analyser un vrai jeu de données biologique de distribution d'espèces. On verra que l'analyse statistique proprement dite n'est qu'une petite partie du processus plus général d'analyse des données qui demande de pouvoir manipuler les données, les nettoyer, les explorer, en faire des représentations graphiques, évaluer la qualité des analyses, interpréter les résultats sur le plan statistique mais aussi et surtout biologique, etc... Le but est d'acquérir par la pratique de bons réflexes lorsqu'on se trouve face à une masse de données importante dont on veut extraire de l'information/de la connaissance et de savoir quels outils il faut mobiliser à quel moment.

Le jeu de données à analyser est composé de deux parties :

1) une série de jeux de données de distribution d'espèces pour la Belgique ou la Flandre disponibles en ligne sur la plate-forme GBIF. Pour vous faciliter la tâche ces jeux de données ont déjà été en partie nettoyés et mis en forme. Il vous faudra sélectionner parmi ces jeux de données un groupe taxonomique (coccinelles, orthoptères, oiseaux, plantes,...) et une ou plusieurs espèces de votre choix.
2) un jeu de données environnemental (~150 variables) décrivant le climat, l'occupation du sol, la pédologie, le relief,... créé sur base de données cartographiques disponibles publiquement pour la plupart (WorldClim et Corine Land Cover principalement)

Le but sera de trouver les variables environnementales qui permettent d'expliquer au moins en partie la distribution d'une espèce ou toute autre variable dérivée des jeux de données GBIF (pex la richesse spécifique). Le but prioritaire n'est donc pas de produire un modèle prédictif de la distribution mais bien de comprendre quelles sont les variables environnementales importantes et comment elles influencent la distribution des espèces. Les données de distribution et environnementales ont été regroupées selon une grille commune de carrés de 5km² dits "UTM" ou "MGRS" fréquemment utilisée pour les inventaires faunistiques.

L'analyse de la distribution d'espèces (Species Distribution Modelling) utilise aujourd'hui une très vaste gamme de méthodes et d'approches. Le but n'est pas de couvrir toutes ces méthodes. On utilisera une approche simple à base de modèles linéaires généralisés (GLMs). Cet outil statistique très largement utilisé a l'avantage d'être très flexible, relativement simple à utiliser et interpréter et d'être également adapté à de nombreuses autres questions en écologie notamment pour des approches plus expérimentales. Les distributions d'espèces ne sont ici qu'un prétexte pour mettre en pratique entre autres cet outil.

Pour ce travail vous devrez répondre aux questions présentes dans ce document (dernière section) mais aussi fournir l'ensemble des données et des informations (lignes de code ou autres) permettant de reproduire vos résultats sans que vous deviez donner d'explications supplémentaires ("reproducible research"). Ceci implique une grande rigueur et une description méticuleuse de toutes les étapes de votre analyse. Vous trouverez de nombreux exemples de cette manière de travailler dans les jeux de données qui vous sont fournis. Ceux-ci sont systématiquement accompagnés de fichiers décrivant leur contenu (fichiers de métadonnées README) mais également des scripts R permettant de reproduire ces jeux de données depuis les données brutes.

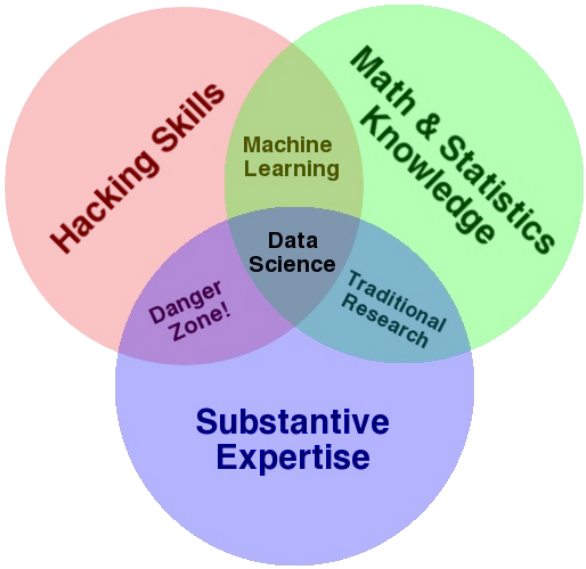
De "la statistique" à la "science des données"

Le terme "data science" est un terme très à la mode depuis une dizaine d'années en particulier dans les milieux du business où les jobs de "data scientists" sont très en vue. Pour certains il est simplement synonyme de "data analysis", "applied statistics", "data mining" ou encore "machine learning" saupoudré de "data visualization" de "big data" et d'une bonne dose de "buzz" et de "hype".
La "science des données" comme on peut la traduire en français est en effet souvent simplement définie comme un domaine interdisciplinaire visant à analyser des quantités massives de données brutes pour en extraire des nouvelles connaissances ou une meilleure compréhension du problème étudié.
L'accent est également souvent mis sur la capacité à communiquer les résultats en particulier via la visualisation des données et sur la valorisation financière possible des "connaissances produites" pour une entreprise ou au moins sur les applications pratiques de ces connaissances pour résoudre des problèmes concrets (orientation résolument "appliquée").

La définition donnée ci-dessus semble pouvoir décrire n'importe quelle démarche scientifique. Si on exclut l'aspect "business", la particularité ici est l'aspect "massif" et "brut" des données. En effet, quoiqu'il en soit de la nouveauté du terme ou de la discipline il dénote une évolution que personne ne nie : l'accumulation de jeux de données de plus en plus massifs, complexes et aussi de plus en plus accessibles à la fois grâce à la démocratisation des outils d'analyse de données et par une tendance à la mise à disposition publique des données ("open data") en tous cas dans le monde scientifique et le secteur public. Cette évolution ouvre des perspectives nouvelles dans de nombreux domaines et pas seulement le business. Elle implique aussi une évolution des outils et des connaissances pratiques nécessaires pour pratiquer l'analyse des données.

Drew Conway a produit un diagramme de Venn éclairant sur les différentes dimensions de la science des données. Elle sont au nombre de 3 :

- **Maths & Statistics** : compétences en statistiques traditionnelles
- Hacking skills : **compétences informatiques** orientées sur la résolution de problèmes
- Substantive expertise ou "**domain expertise**" : expertise dans un domaines de connaissance (pex l'écologie)



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

L'idée est que pour pouvoir extraire de la connaissance d'une masse de données on a aujourd'hui non seulement besoin des outils des statistiques classiques mais aussi de **compétences informatiques** afin de pouvoir manipuler ces grosses quantités de données souvent non structurées pour les nettoyer et les mettre dans une forme qui soit exploitable pour le problème étudié ("data tidying"). Les compétences informatiques permettent aussi de faciliter grandement l'application de méthodes statistiques et de visualisation ce qui permet de réduire le temps et le coût entre les idées que l'on veut explorer et leur mise à l'épreuve des données.

Il faut également des connaissances dans le domaine qui touche les données que l'on veut analyser ("**domain expertise**") par exemple l'écologie, l'agronomie, l'économie, etc... (voire des domaines plus restreints) pour pouvoir poser les bonnes questions et interpréter correctement les résultats.
Lorsque le domaine d'expertise est la biologie moléculaire, le "data scientist" est en général appelé "bioinformaticien". Il ne semble pas exister d'équivalent aussi répandu en biologie des organismes. Des compétences dans le domaine d'expertise sont donc importantes en amont (choisir les bonnes questions) et en aval (interprétation des résultats) de l'analyse de données mais aussi pendant l'analyse de données pour guider les nombreux choix qui doivent être faits sur base des connaissances déjà existantes sur le domaine étudié.

Sur le diagramme on distingue aussi 3 zones constituées par l'intersection de 2 des 3 domaines de connaissance.

Au croisement statistiques et expertise de domaine, on trouve la **recherche traditionnelle**. Elle correspond en général aux cas où le scientifique récolte lui-même une quantité de données souvent limitée (pex essais expérimentaux en labo ou en champ). C'est une approche qui a encore de beaux jours devant elle car elle permet de tester des hypothèses précises et notamment les liens de causalité entre plusieurs facteurs. Les besoins de connaissances informatiques sont ici limitées notamment à cause de la petite taille des jeux de données. Cependant en se cantonnant à ces approches on se prive d'explorer des questions scientifiques qui ne peuvent être explorées avec les outils de la recherche traditionnelle et on ignore une masse d'informations et de données qui sont à disposition de tous.

Au croisement des compétences informatiques et de l'expertise de domaine, on trouve une "**zone de danger**". En effet la facilité amenée par la maîtrise des outils informatique permet d'appliquer de nombreuses méthodes statistiques à un jeu de données. Le danger est de mal interpréter les résultats ou d'utiliser les outils à mauvais escient si on a pas le background nécessaire en statistiques...

Enfin au croisement des compétences informatiques et statistiques on trouve le "**Machine learning**". Il s'agit ici aussi d'un terme relativement vague qui recouvre en général une série de techniques algorithmiques et statistiques visant à traiter de manière semi automatique des données pour obtenir la prédiction la plus précise possible d'un phénomène, sans spécialement chercher à comprendre les mécanismes. Ces méthodes sont souvent vues comme des "black boxes". ex : prédiction de spams, propositions d'achats sur les boutiques en ligne, reconnaissance de patterns (OCR, images,...), télédétection, spectrométrie,... Il manque ici la dimension de production de "connaissances". On veut créer un outil qui peut prédire quoi, où et quand mais on ne veut pas spécialement comprendre le pourquoi. On ne veut par exemple pas spécialement savoir pourquoi telle longueur d'onde permet de distinguer une prairie d'une forêt ou quels sont les mots les plus importants pour identifier un spam. Les outils du "machine learning" peuvent cependant aussi être utilisés pour comprendre un phénomène (modèles explicatif plutôt que prédictif) si ils sont combinés avec l'expertise dans le domaine (sans quoi il s'agit aussi d'une zone de danger où on risque de mal interpréter les résultats...).

Au final ces 3 compétences peuvent se trouver chez des personnes différentes avec des profils différents d'informaticiens/programmeur, de statisticien, d'écologues, qui peuvent collaborer dans des équipes pluridisciplinaires. Il est en effet matériellement impossible pour une même personne d'approfondir tous les sujets. Il est cependant nécessaire que l'écologue ait un minimum de baguage à la fois en statistiques et en informatique et ce pour au moins trois raisons. 1) Pour pouvoir se parler dans un langage commun et donc collaborer efficacement, ces 3 personnes doivent avoir de bonnes notions dans les deux autres disciplines. Il faut donc idéalement que le statisticien et/ou l'informaticien du groupe aient aussi quelques bonnes notions d'écologie... 2) Les écologues ont tout simplement rarement "sous la main" des informaticiens et statisticiens qui peuvent se consacrer à l'analyse en profondeur de leurs données et encore moins des informaticiens et statisticiens qui ont de bonnes notions en écologie. L'écologue doit pouvoir se débrouiller dans la plupart des cas et aller chercher de l'aide pour les cas de figure plus complexes. 3) De nombreuses étapes dans l'analyse de données demandent en fait de mobiliser les 3 types de compétences en même temps et il serait peu efficace que ces compétences se trouvent entièrement séparées chez 3 personnes différentes...

La "science des données", pour qui ?

Il y a encore peu de temps, l'utilisation des statistiques était cantonnée principalement au domaine de la recherche. Aujourd'hui, avec le glissement vers la "science des données" et les jeux de données massifs, les besoin de traitement de données s'étendent de plus en plus en dehors du domaine strict de la recherche scientifique et au delà des simples" tests statistiques".

Aussi de nombreuses personnes avec un profil de macro-biologiste/écologiste peuvent avoir besoin des compétences liées à la science des données :

1) Le **chercheur** parfois frustré de ne pas pouvoir traiter efficacement les nombreuses données qu'il a récoltées sur un sujet qui le passionne. Les chercheurs ont aussi de plus en plus besoin de pouvoir exploiter les masses de données existant en ligne ou dans des bases de données privées et plus seulement les données qu'ils ont récoltées eux-mêmes.

2) Le/la **biologiste de terrain** (ONGs, administrations, bureaux d'études...) qui doit de plus en plus traiter des masses importantes de données. Il s'agit parfois de traitements statistiques proprement-dits (modèles, tests statistiques, etc...) mais aussi très souvent simplement d'être capable de manipuler et résumer de grosses quantités de données (parfois de manière récurrente). Quelques exemples :

- traitements statistiques pour évaluer les tendances de populations (listes rouges, bioindicateurs,...), pour déterminer les facteurs environnementaux importants pour la conservation d'une espèce,...
- traiter les récoltes de données automatiques : enregistreurs automatiques d'ultrasons émis par des chauves-souris ou des orthoptères (avec identification semi-automatique), metabarcoding (ea environnemental), suivi des déplacements d'animaux munis de balises GPS, ...
- utiliser des données cartographiques (GIS) : cadastre pour l'achat de terrain, définir des zones d'actions prioritaires (pex via cartes de qualités de l'habitat), utiliser des photos satellites/aériennes/radar pour repérer des zones d'action - pex : prédiction de taches plantes invasives le long de cours d'eau, prédiction d'arbres dépérissant, comptages de grands mammifères sur base de photos aériennes, anciens lits de rivière drainée à restaurer, aire de faulde sur base de couches LIDAR, etc...
- automatiser les rapports et les analyses récurrentes (typiquement on réévalue par exemple annuellement l'état de la qualité des eaux ou des populations d'oiseaux hivernants selon le même schéma)

3) L'**étudiant(e) en biologie** qui ne doit pas seulement se forger une culture générale sur l'état des connaissance en biologie mais qui doit être capable de comprendre comment les nouvelles connaissances sont créées (de façon par exemple à pouvoir lire de manière critique la littérature scientifique) et créer soi-même de nouvelles connaissances (travail de fin d'étude pex).

4) La **citoyenne/journaliste** qui veut vérifier par elle-même à la source si ce qu'on (politiciens, publicitaires, lobbyistes, autres citoyens, ...) lui dit est vrai ou pour comprendre la société dans laquelle on vit. La production de connaissance et d'information est de plus en plus prise en charge par des citoyens, des ONG et des médias participatifs. Les données brutes sont souvent disponibles mais elles ont besoin d'être traitées pour être transformées en information.

Alberto Cairo - 2016 - The Truthful Art

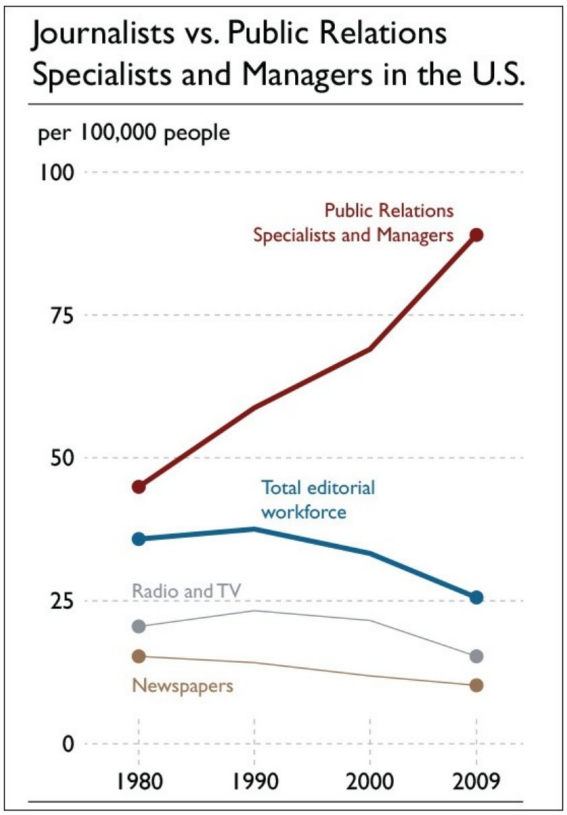
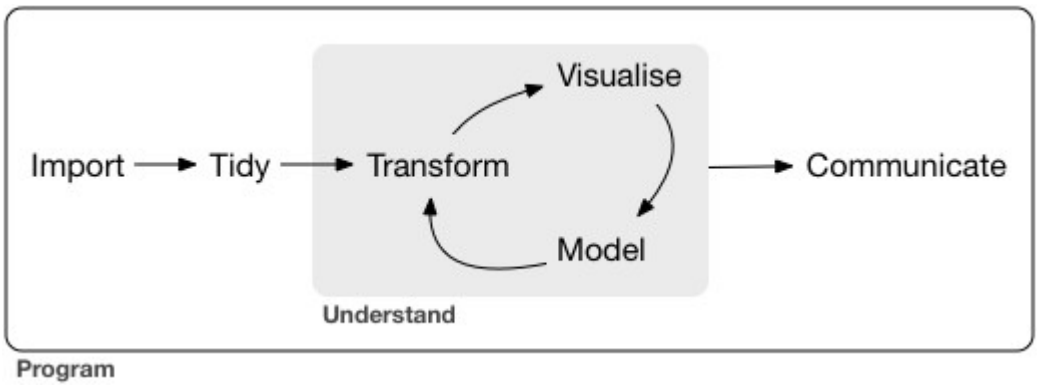


Figure I.10 The number of professionals working in public relations has expanded greatly in the past three decades, while the number of journalists has dropped. (Graph based on McChesney and Nichols, 2011.)

Les étapes typiques d'une analyse de données

L'analyse d'un jeu de données complexe se fait souvent au cas par cas et il est illusoire de présenter une recette unique. Cependant, il existe souvent une approche commune en particulier dans les premières étapes de l'analyse : description du contexte et définition des questions, récolte, vérification, nettoyage puis exploration des données (ea graphiques),... La création d'un modèle explicatif ou prédictif (pex GLMs) pour une ou plusieurs variables d'intérêt est également une étape importante dans de très nombreux cas (mais pas tous). C'est l'exemple que l'on prendra ici. Il faut ensuite interpréter les résultats et les communiquer.

Le processus présenté ici sous forme linéaire ne doit pas être vu comme une procédure figée. En particulier le processus n'est souvent pas aussi linéaire et il est fréquent de retourner à des étapes antérieures en cours d'analyse de données ou de précder de manière circulaire comme sur le graphique ci-dessous (par exemple pour corriger des erreurs que l'on avait manquées au départ ou pour rassembler de nouvelles données pour approfondir certains aspects au vu des premiers résultats, ...). Les étapes ne sont également souvent pas aussi clairement séparées (on explore déjà les données quand on est en train de les assembler...).
NB : à chaque étape du processus détaillé ci-dessous on indiquera l'importance relative des 3 domaines de la science des données : statistique, informatique ("hacking skills"), expertise de domaine ("Substantive expertise").



Grolemund & Whickam 2016 - R for data science

1) Contexte et question scientifique

La première étape consiste à définir clairement le **contexte de l'étude** et le **problème à résoudre** ou la **question scientifique**.

On ne saurait trop insister sur l'importance de cette étape pourtant souvent négligée. La ou les questions qu'on se pose vont déterminer la manière de récolter les données, le type de données à rassembler mais aussi par exemple vont être déterminantes pour décider de la qualité des données. En effet, un même biais/problème dans la récolte de données peut avoir un effet dramatique pour une question scientifique et pourra être négligé pour une autre question scientifique appliquée au même jeu de données (Ex : comptages chauves-souris vs temps ou vs habitat).

L'exploration des données disponibles et les premiers résultats peuvent parfois aider à raffiner les questions ou en amener de nouvelles (ex : abeilles + fongicides et intercultures, pies grièches et haies).

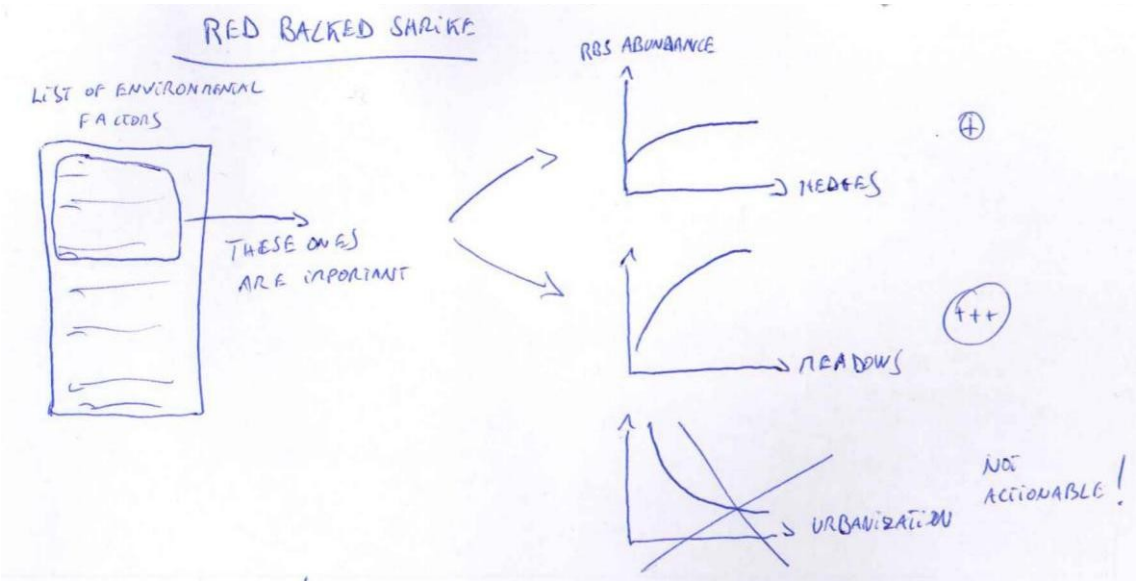
Il est aussi important que les questions posées ne soient pas trop générales (pex : Pourquoi les abeilles domestiques disparaissent ? Comment augmenter les ventes de nos produits?). De telles questions relèvent plutôt du contexte. Les questions doivent correspondre à des hypothèses plus précises ou des objectifs plus restreints qui peuvent être reliés à des données précises - pex : Est-ce que le risque de dépérissement des abeilles est plus élevé lorsqu'elles sont exposées à une combinaison de pesticides et de virus ? Comment identifier les femmes enceintes sur base de leurs achats en ligne pour leur envoyer des publicités ciblées ? (exemple célèbre en data science...).

Un problème fréquent, en particulier lors d'approches expérimentales, est qu'on a trop de questions différentes pour des moyens limités. A chaque question correspond souvent une manière optimale de collecter les données qui n'est pas compatibles avec les autres questions. Et plus vous avez de questions plus vous aurez besoin d'un grand nombre de répétitions ce qui est souvent difficile avec des approches expérimentales ou semi-expérimentales. Il faut donc pouvoir mettre des priorités sur vos questions et faire des choix si nécessaire.

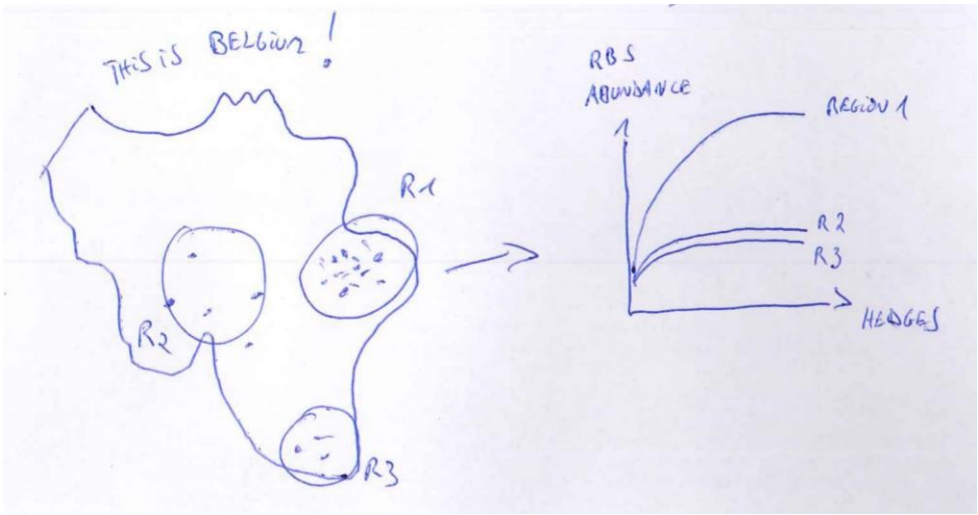
Il est particulièrement utile à se stade de se demander qui va utiliser le résultat final de l'analyse, dans quel but (quels sont les besoins des utilisateurs finaux) et en conséquence d'**imaginer quelle sera la forme idéale de l'output** (article scientifique, rapport, site web interactif , programme informatique, ...) et de son contenu (modèle prédictif, modèle descriptif brut, arbre de décision, représentation graphique des modèles, carte pex de qualité de l'habitat,...) ? Imaginer de cette manière le résultat final vous oblige à bien définir les questions. Cette étape de projection peut se faire avec un simple crayon et bout de papier (voir exemples ci-dessous)...

On peut déjà à ce stade très précoce (avant d'avoir les données) avoir une **idée approximative du type d'analyses statistiques** qui devront être mises en œuvre au moins en première approche.

Dans l'exemple ci-dessous on imagine avant même de récolter les données à quoi pourrait ressembler la présentation finale des résultats dans une étude visant à déterminer quelles sont les actions à mener en priorité pour la conservation de la pie grièche écorcheur. Le but est de voir si ce qu'on envisage comme analyse correspond aux besoins finaux et si les questions/problèmes ont été clairement définis. On peut aussi imaginer à quoi ressemblera le résultat exprimé avec des mots (et des chiffres complètement inventés à ce stade) : "Parmi les facteurs environnementaux étudiés 3 semblent liés à la l'abondance des pies grièches (en nombre de couples par km²) : les haies, les prés de fauche et l'urbanisation. Si la surface de prés de fauche passe de 1ha à 2ha par km² on s'attend à une augmentation de 150 % des populations de pie grièche. Si on passe d'un linéaire de haies de 100 à 600m par km² on s'attend seulement à une augmentation de 10 % des effectifs de pies grièche. L'urbanisation a un effet très négatif mais il est plus difficile de prendre des mesures de restauration concernant ce facteur."

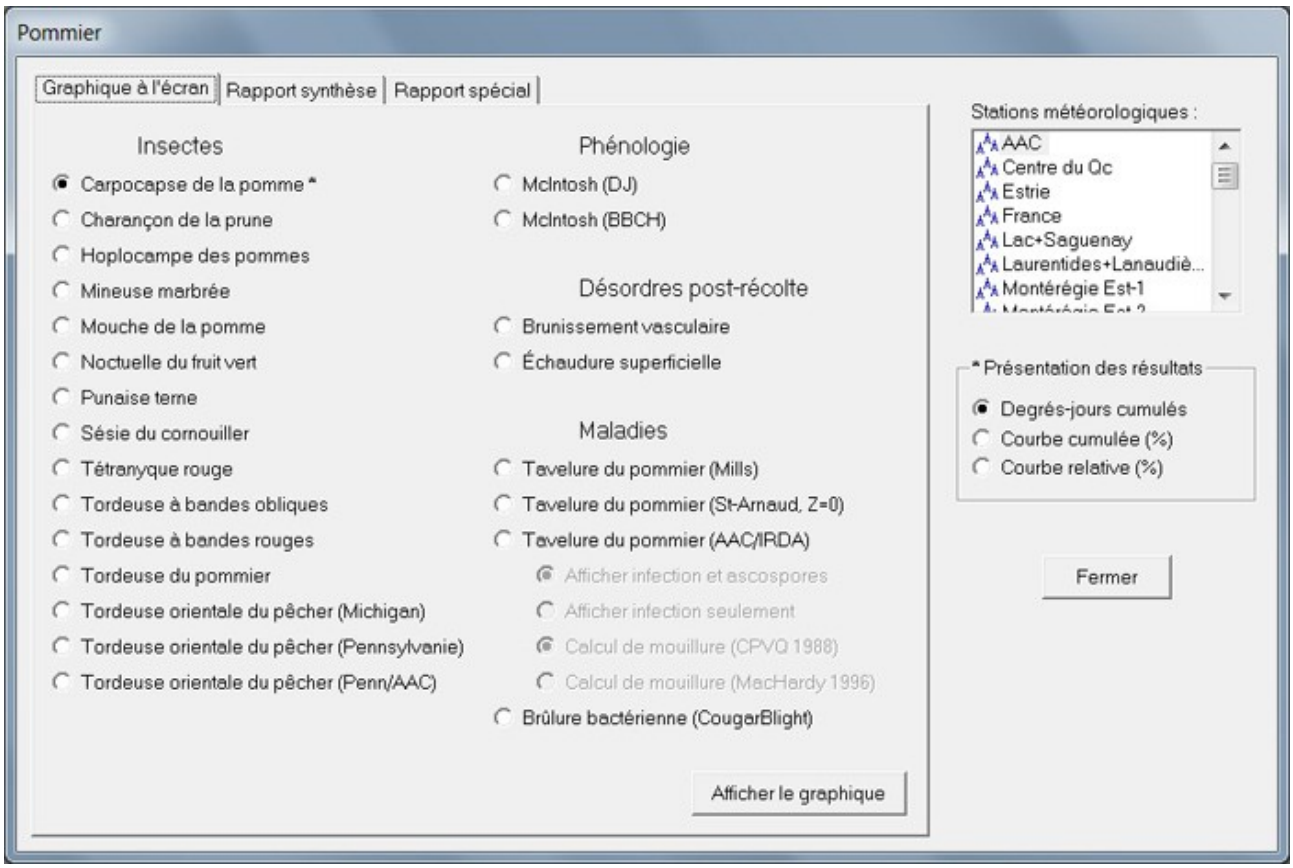


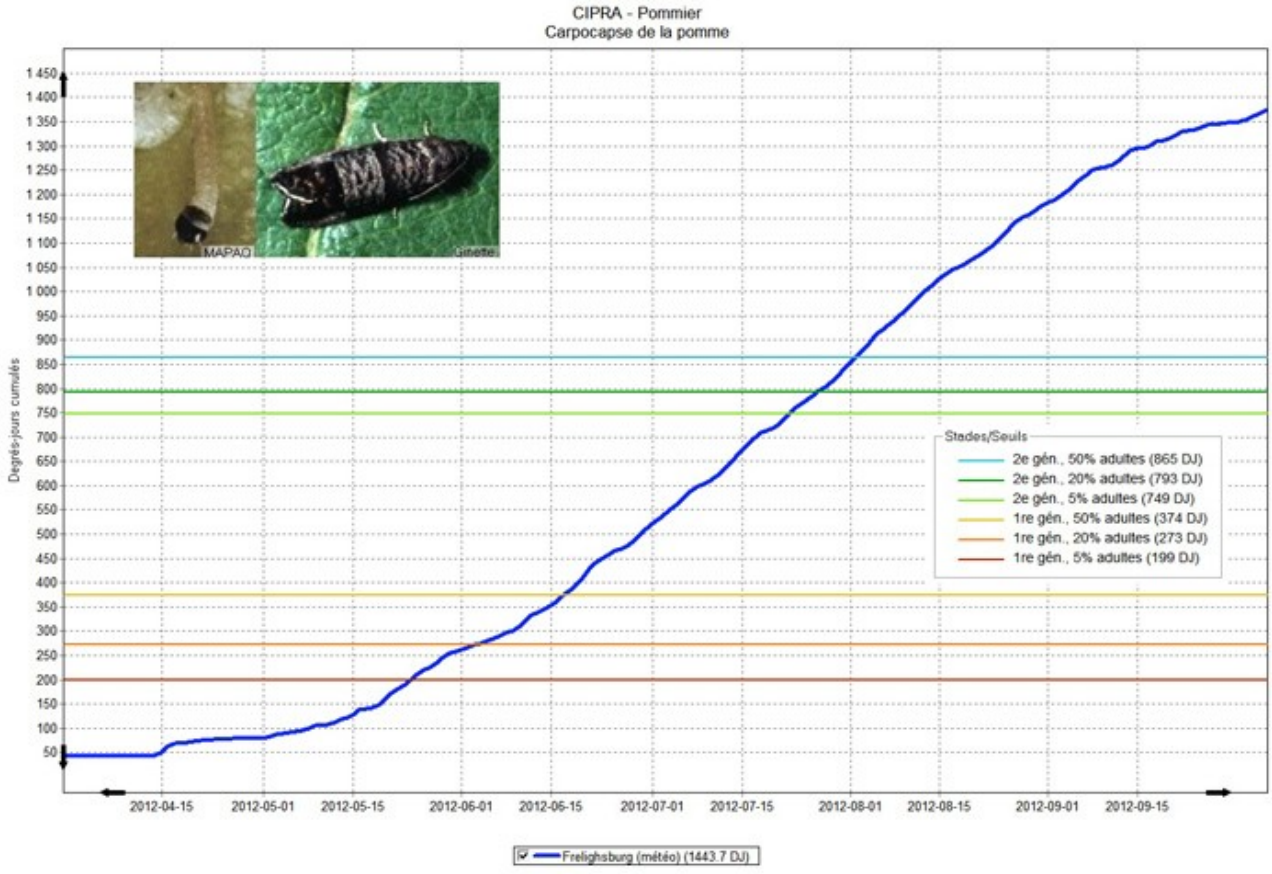
Parfois les premiers résultats amènent d'autres questions et permettent de raffiner l'analyse. Par exemple si on est surpris du faible effet des haies sur cet oiseau on peut se demander si cet absence de lien est présent partout. On pourrait alors constater qu'il y a un effet positif des haies dans une région mais pas dans les autres (sites de nidification alternatifs ? saturation du milieu en haies?). Attention que cette démarche de fouille dans les données ("data dredging" - "data mining") finira toujours par faire apparaître des patterns même si ils sont dus au hasard. Le but dans ce cas est surtout de générer des hypothèses plutôt que de les tester. Si on veut tester formellement l'hypothèse de l'effet des haies, il faudra vraisemblablement le faire sur un nouveau jeu de données indépendant ou voir si d'autres études trouvent des résultats convergents ...



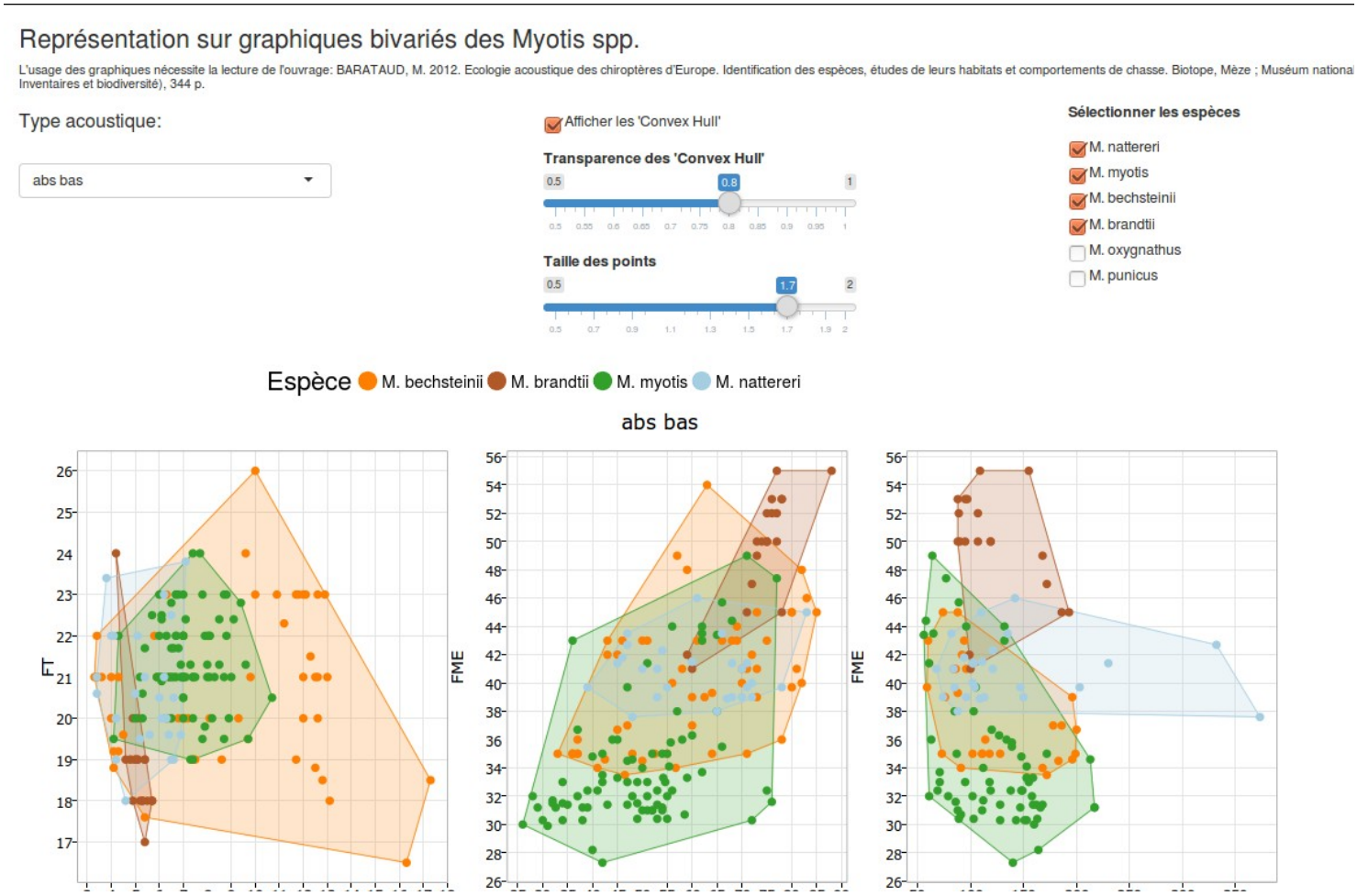
Si l'objectif de l'étude est de restaurer l'habitat d'un papillon dans une réserve naturelle précise les résultats finaux seront probablement présentés différemment même si ils sont basés sur le même genre d'approche analytique. Le but est ici de savoir où sont les zones défavorables sur lesquelles on pourrait agir directement. Une série de cartes est sans doute l'output le plus logique dans ce cas.

L'exemple ci-dessous montre des captures d'écran d'un logiciel canadien - CIPRA - permettant de prédire le développement de ravageurs et des cultures en se basant sur des données météorologiques locales. Le but est d'aider les agriculteurs à évaluer en temps réel les risques pour leurs cultures et favoriser des interventions raisonnées. La création d'un logiciel est un métier en soi mais il existe de plus en plus d'outils permettant de créer facilement des petites applications interactives en ligne (pex en R avec Shiny). Une autre approche fréquemment utilisée en agriculture sont des systèmes d'avertissement sous forme d'e-mails. Des scientifiques récoltent des données et les analysent en temps réel (tous les jours par exemple) et diffusent les résultats sous forme d'informations totalement pratiques (faut-il intervenir oui/non/probablement) et avec très peu de détails techniques. Si c'est l'objectif final, il faudra réfléchir dès le départ aux données nécessaires pour décider par exemple d'un seuil de nuisibilité (données binaires) plutôt que de prédire par exemple le nombre exact d'individus dans une population....


Pommier
Graphique à l'écran
Rapport synthèse
Rapport spécial
Insectes
Carpocapse de la pomme *
Charançon de la prune
Hoplocampe des pommes
Mineuse marbrée
Mouche de la pomme
Noctuelle du fruit vert
Punaise terne
Sésie du cornouiller
Tétranyque rouge
Tordeuse à bandes obliques
Tordeuse à bandes rouges
Tordeuse du pommier
Tordeuse orientale du pêcher (Michigan)
Tordeuse orientale du pêcher (Pennsylvanie)
Tordeuse orientale du pêcher (Penn/AAC)
Phénologie
McIntosh (DJ)
McIntosh (BBCH)
Désordres post-récolte
Brunissement vasculaire
Échaudure superficielle
Maladies
Tavelure du pommier (Mills)
Tavelure du pommier (St-Arnaud, Z=0)
Tavelure du pommier (AAC/IRDA)
Afficher infection et ascospores
Afficher infection seulement
Calcul de mouillure (CPVQ 1988)
Calcul de mouillure (MacHardy 1996)
Brûlure bactérienne (CougarBlight)
Afficher le graphique
Stations météorologiques :
AAC
Centre du Qc
Estrie
France
Lac+Saguenay
Laurentides+Lanaudiè...
Montérégie Est-1
* Présentation des résultats
Degrés-jours cumulés
Courbe cumulée (%)
Courbe relative (%)
Fermer


CIPRA - Pommier
Carpocapse de la pomme
MAPAQ
Degrés-jours cumulés
Stades/Seuils
2e gén., 50% adultes (865 DJ)
2e gén., 20% adultes (793 DJ)
2e gén., 5% adultes (749 DJ)
1re gén., 50% adultes (374 DJ)
1re gén., 20% adultes (273 DJ)
1re gén., 5% adultes (199 DJ)
2012-04-15
2012-05-01
2012-05-15
2012-06-01
2012-06-15
2012-07-01
2012-07-15
2012-08-01
2012-08-15
2012-09-01
2012-09-15
Frelighsburg (météo) (1443.7 DJ)

Exemple d'application interactive en ligne (créée avec R et shiny) pour faciliter l'exploration d'une base de données de sons de chauves-souris en vue de faciliter leur identification. Il ne s'agit donc pas ici d'une analyse statistique proprement dite mais d'un outil pour faciliter la consultation des données.
https://jeff37.shinyapps.io/Shiny1fileBarataud2016/
(Création de JF Godeau, sur base de données de M.Barataud)



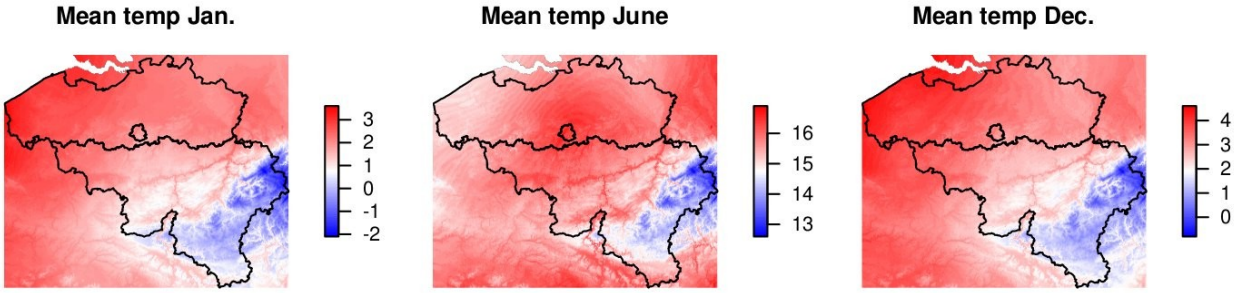
2) Récolte de données

On peut ensuite récolter de nouvelles données (sur le terrain ou en labo) ou rassembler des données déjà existantes. De plus en plus souvent on combine les deux approches...

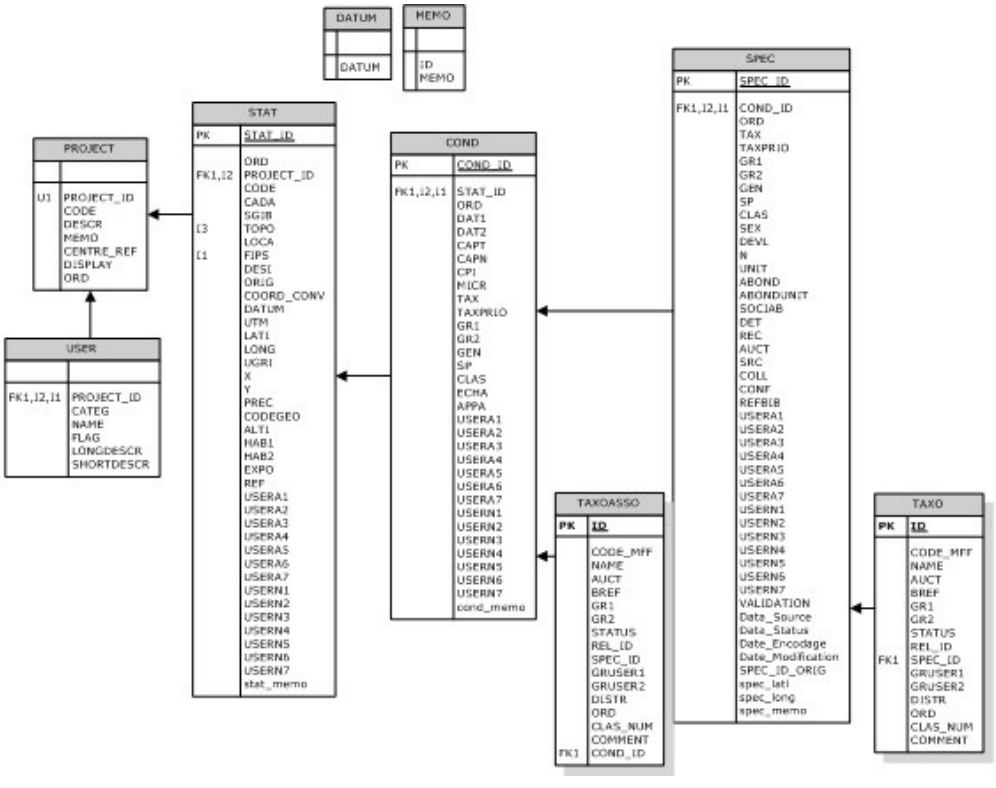
La récolte de nouvelles données implique une bonne connaissance du domaine d'expertise mais aussi dans les bonnes pratiques du design expérimental (compétences statistiques). Il faut idéalement choisir de bons contrôles, réfléchir à l'échantillonnage (aléatoire, aléatoire stratifié, cas/témoin, etc...), s'assurer qu'on a bien des réplicats indépendants (et pas majoritairement des pseudoréplicats, des mesures répétées, etc...), s'assurer d'avoir un nombre d'échantillons suffisants, etc...

Rassembler des données déjà existantes demande souvent plutôt des compétences informatiques dans un premier temps pour se connecter à des sources de données dans différents formats (csv, SQL based, xml, JSON, geoTiff, shp, ...) , extraire l'information pex de données cartographiques (GIS) ou perdue dans des pages web (web scrapping) et ensuite mettre en forme les données, les résumer si nécessaire, pour qu'elles soient exploitables.

Exemple de données largement disponibles : températures mensuelles de Belgique avec une précision de < 1km² disponible sous forme de données spatiales raster. Il faut être capable d'extraire et de résumer les données pour les zones qui nous intéressent dans un projet précis.



Structure d'une base de donnée fréquemment utilisée en Wallonie pour stocker des données biogéographiques de distribution d'espèces (Data Fauna Flora). Il faut être capable de comprendre la structure d'une telle base de données pour pouvoir l'exploiter et parfois d'écrire quelques commandes SQL simples pour pouvoir extraire les données.



3) Vérification et nettoyage des données

L'idée ici est d'abord de vérifier qu'il n'y a pas de données mal encodées (valeurs impossibles, fautes de frappes), de repérer les valeurs manquantes, de vérifier que les données ont été correctement importées dans le logiciel d'analyse, etc ...
Il faut aussi se poser la question de la qualité des données : comment les données ont été récoltées et quels problèmes/biais peuvent être présents dans les données en conséquence. On peut alors choisir d'éliminer une partie des données de plus mauvaise qualité ou qui ne correspondent pas aux questions posées.
Cette partie demande donc surtout des compétences informatiques pour pouvoir manipuler aisément les données et des compétences dans le domaine d'expertise pour pouvoir évaluer de manière critique la qualité des données et les biais potentiels.

Ex : fautes de frappe très fréquentes, espaces dans des champs numériques,...
Ex : on importe 32000 lignes d'un jeu de données qui en fait 70000 (à cause de caractères spéciaux). On importe des données numériques sous forme de texte (à cause du séparateur des décimales = ",")
Ex données GBIF: élimination données trop anciennes, élimination des données avec une précision géographique insuffisante, élimination des carrés où l'effort d'échantillonnage n'est pas assez important (pour minimiser les fausses absences),...

Dans R, un simple "summary" d'un jeu de données permet souvent de repérer déjà certaines erreurs ou risques potentiels pour l'analyse de données:

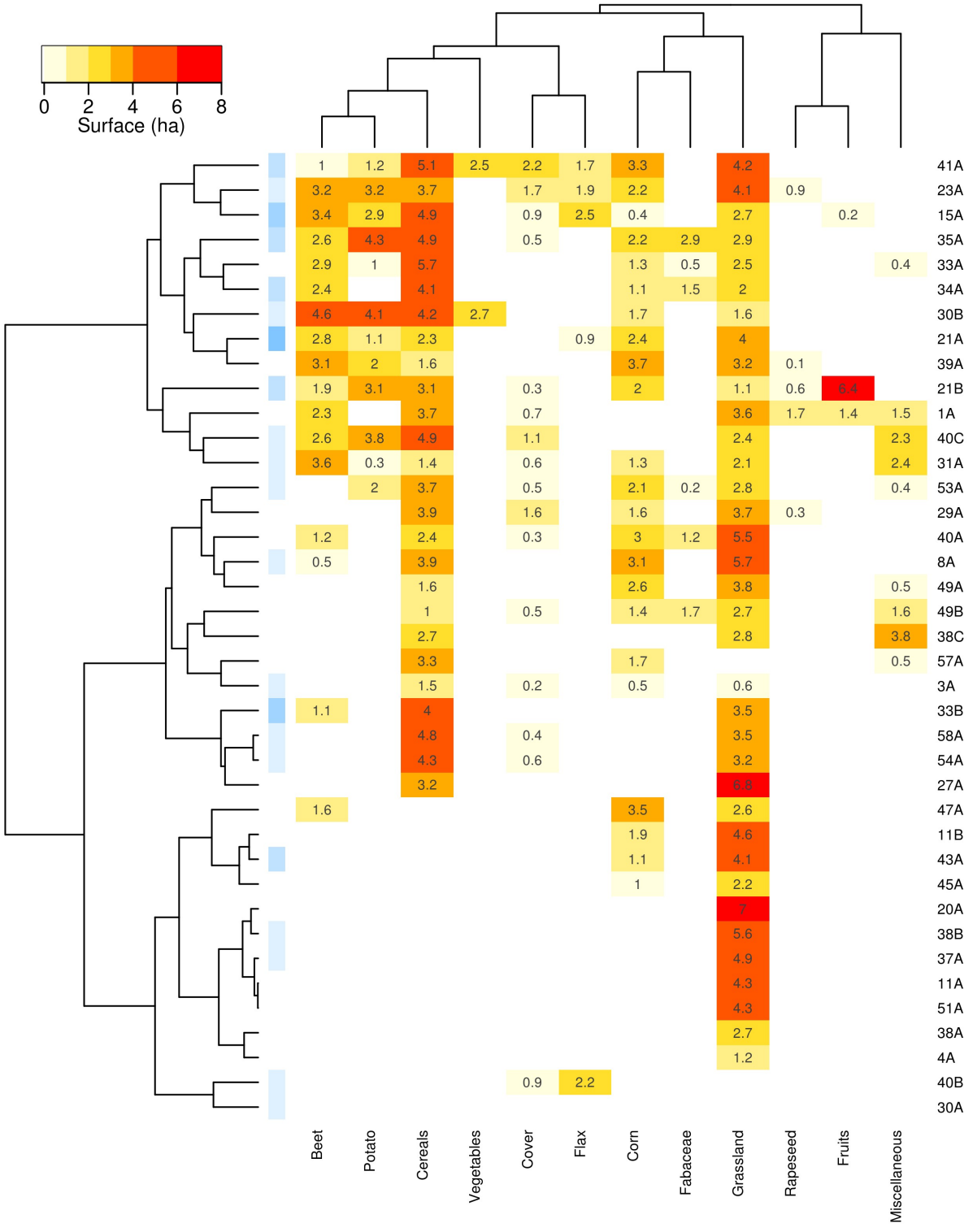
```
> head(d)
  taille poids site nid   sexe            memo
1    9.4   5.3    3   2   male
2    9.8   5.3    1   1 female
3   11.6   6.3    2   2   <NA>       Mal formé
4   10.1   5.7    4   2 female
5   12.0   3.7    4   1   male peson déréglé ?
6   11.7   6.7    1   1 female

> summary(d)
     taille          poids            site           nid           sexe                   memo
 Min.   : 7.70   Min.   :3.700   Min.   :1.000  Min.   :1.000  female:36                  :72
 1st Qu.: 9.40   1st Qu.:4.900   1st Qu.:2.250  1st Qu.:2.000  femle : 1   Mal formé      : 3
 Median :10.05   Median :5.500   Median :4.000  Median :3.000  male  :38   peson cassé    : 2
 Mean   :10.29   Mean   :5.471   Mean   :3.987  Mean   :2.821  NA's  : 3   peson déréglé ?: 1
 3rd Qu.:10.78   3rd Qu.:6.000   3rd Qu.:6.000  3rd Qu.:4.000
 Max.   :29.00   Max.   :6.900   Max.   :7.000  Max.   :5.000
                 NA's   :2
```

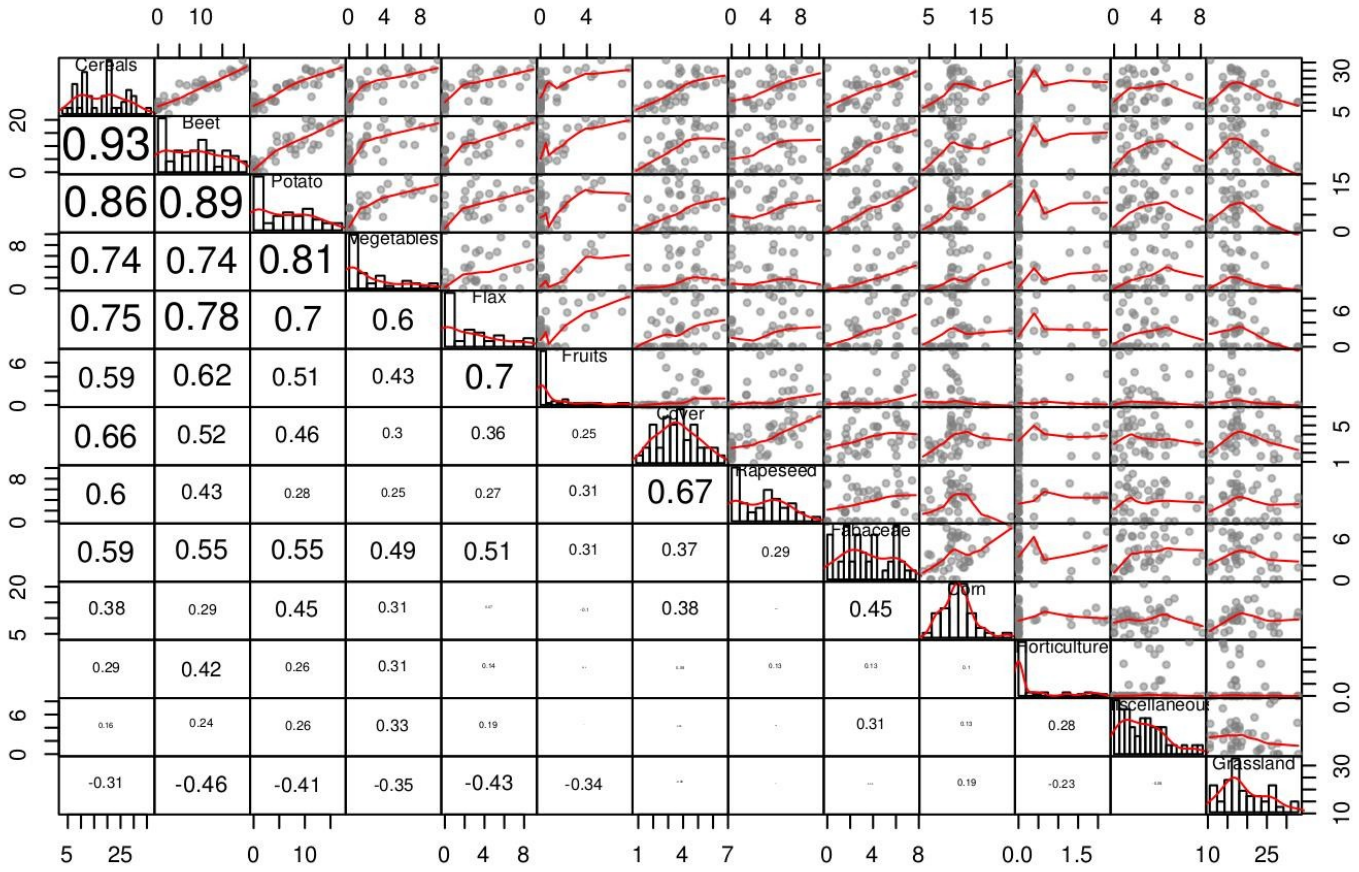
4) Exploration des données

Il s'agit ici non plus de vérifier la qualité des données mais bien d'explorer et de s'approprier leur contenu.
Le but est à la fois 1) d'acquérir une **connaissance approfondie des données** qui aideront à l'interprétation du résultat final et 2) de repérer les **problèmes potentiels pour les analyses** statistiques subséquentes (variables explicatives fortement corrélées entre elles, distributions fortement asymétriques, variables qualitatives avec des classes sous-représentées,...).
On utilise ici massivement des **représentations graphiques** et/ou des **statistiques descriptives** permettant de résumer un jeu de données complexe (pex clustering, ordinations, heatmaps,...). On explore en particulier les relations (pex corrélations) entre les future variables explicatives potentielles, leur distribution (uniforme ? symétrique ? fortement asymétrique?), etc... Si nécessaire on pourra déjà transformer certaines variables, modifier leur échelle, combiner plusieurs variables pour en former de nouvelles, etc...
Bien entendu il faut garder son esprit critique en éveil et ce n'est parfois que lors de cette phase d'exploration approfondies qu'on s'aperçoit que certaines parties du jeu de données ne sont pas correctes et doivent être éliminées ou corrigées.
Les compétences nécessaires ici s'orientent progressivement plus vers les statistiques tout en nécessitant toujours des compétences informatiques pour manipuler les données et les représenter graphiquement.

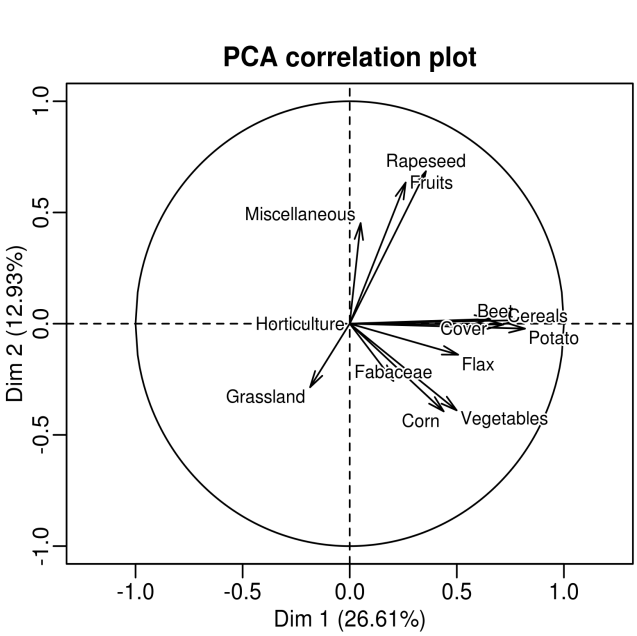
La combinaison d'analyses multivariées (clustering) et d'une représentation graphique (dendrogramme + heatmap) permet de visualiser les patterns qui auraient autrement été invisibles dans une matrice complète de données. Cette matrice représente les surfaces de différents types de cultures dans un rayon de 500m autours de ruchers d'abeille domestique. Le clustering est utilisé pour réorganiser les lignes et les colonnes du jeu de données afin de mieux visualiser les relations entre variables et entre sites. Le clustering des sites montre clairement trois groupes qui peuvent être facilement interprétés grâce à la carte de chaleur (heatmap) : un groupe de grande cultures intensives (betteraves, pomme de terre, céréales), un groupe avec des céréales mais sans ou avec très peu de betteraves/pdt et un groupe avec très peu de cultures et principalement des prairies. Sans la carte de chaleur les résultats du clustering seraient très difficiles à interpréter...

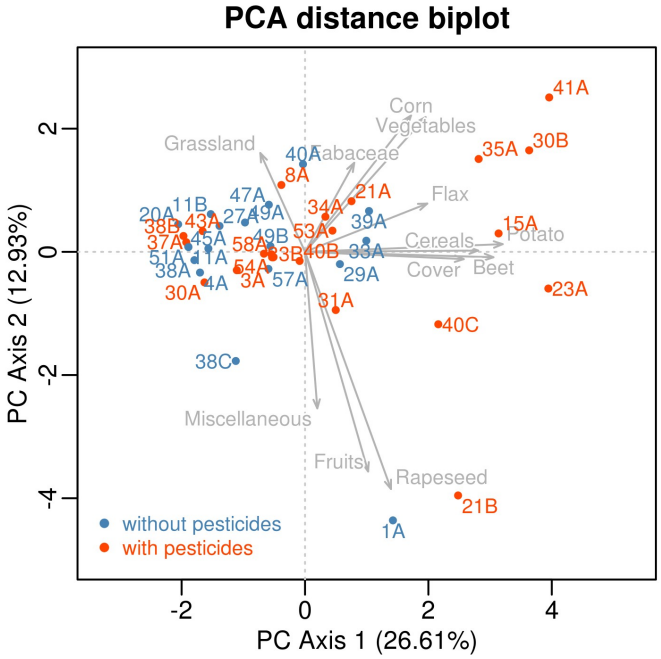


Un autre type de représentation graphique ("scatterplot matrix") permet de visualiser directement les corrélations entre les surfaces de différentes cultures et leur distribution (sur la diagonale). Ici aussi les variables ont été automatiquement réordonnées de façon à grouper les variables les plus corrélées. Ce genre de graphique très chargé n'est pas destiné à communiquer des résultats à d'autres personnes mais bien à explorer soi-même le contenu d'un jeu de données.

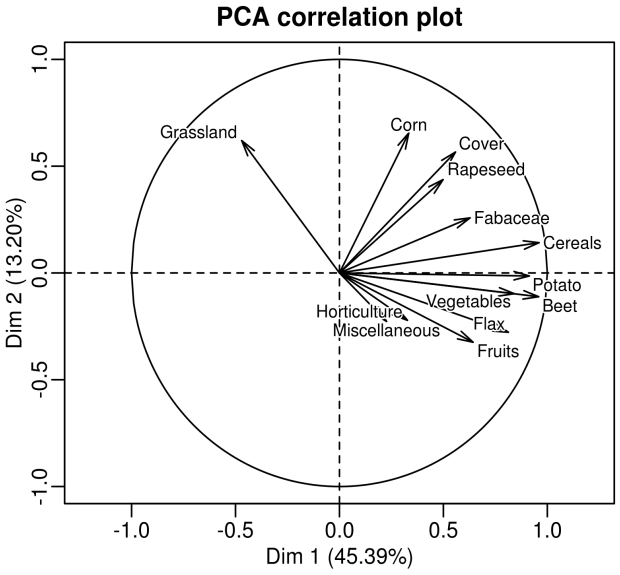


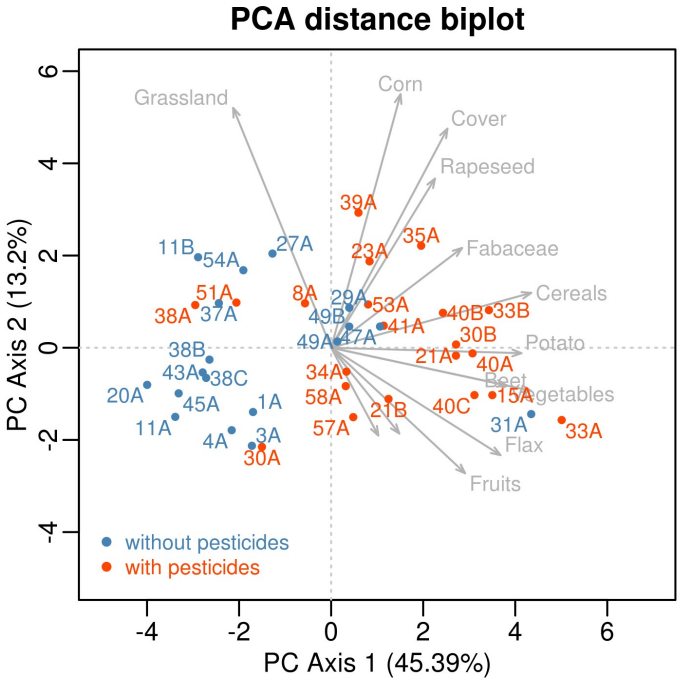
Une troisième représentation de la même matrice de données ne donne pas les mêmes informations. Ici la représentation est moins claire que les clustering + heatmap. C'est la combinaison des différentes approches qui permet de voir la structure "cachée" des données.





Par contre le même outil statistique fait apparaître des patterns intéressants pour la compréhension des données sur une matrice des surfaces agricoles dans un rayon de 3km autour des ruchers (au lieu de 500m). Mais ce genre de représentations très résumées ne sont pas faciles à interpréter pour le lecteur non averti et demande souvent quelques explications...
On voit ici clairement un gradient de surfaces de cultures sur le premier axe et de surfaces de prairies (et le maïs associé) sur l'axe 2. De plus on voit clairement que les échantillons contaminés par des pesticides se trouvent essentiellement à droite le long du premier axe là où les surfaces de cultures sont plus grandes.
Ici on ne se contente pas d'explorer les variables explicatives. On a déjà en tête l'objectif de l'étude : déterminer l'origine des contaminations.





5) Analyse statistique proprement dite - modélisation

Les compétences statistiques sont ici prépondérantes. Les connaissances informatiques nécessaire sont moindre que dans les étapes précédentes, les lignes de codes sont simples et souvent similaires d'un projet à l'autre.

a) Construction du modèle

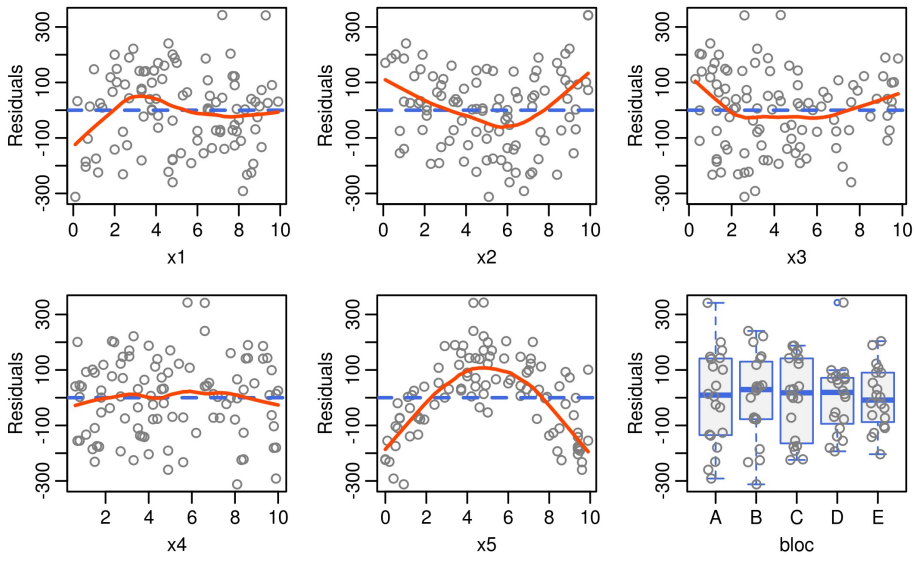
On peut ensuite construire un premier modèle prédictif/explicatif par exemple avec un GLM. On aura choisi les variables explicatives à la fois sur base de critères biologiques (variables potentiellement importantes pour expliquer le phénomène étudié) et statistiques (éviter les variables trop corrélées, éviter les points extrêmes,...).

Bien avant de regarder les résultats du modèle (coefficients, p-valeurs,...) il faut évaluer de manière critique la qualité du modèle et prendre des dispositions le cas échéant.
Pour un GLM, on vérifiera en particulier les points suivants :

- La relation entre réponse et variables explicatives quantitatives est-elle bien à peu près linéaire ? (graphiques résidus vs variables explicatives)
- Les hypothèses de variabilité des résidus sont-elles respectées (homogénéité de la variance résiduelle pour un modèle gaussien, pas de surdispersion pour les modèles de Poisson et binomiaux)
- La distribution des résidus suit-elle à peu près la distribution attendue (QQ plots) ?
- Y a-t-il des problèmes de multicolinéarité ? (corrélations multiples entre variables explicatives - VIFs) ?
- Y a-t-il des points extrêmes qui pourraient influencer fortement les résultats à eux seuls ?
- Est-ce qu'on peut s'attendre à des problèmes de surparamétrisation (overfitting : trop de variables explicatives)
- Les résidus sont-ils bien indépendants ? (pas de structure hiérarchique, mesures répétées, indépendance spatiale)
- Est-il nécessaire ou utile de centrer ou de standardiser les variables explicatives ?

Ici encore les représentations graphiques sont souvent indispensables si on ne veut pas travailler à l'aveugle.
Le premier modèle est rarement suffisant et nécessitera de multiples ajustements et nouvelles vérifications : changer les variables explicatives, transformer la réponse ou les variables explicatives, ajouter des interactions, changer de type de modèle (gaussien, poisson, ... modèle mixte), refaire l'analyse avec et sans les points extrêmes pour évaluer la robustesse des résultats, ...

Graphiques des résidus en fonction des 6 variables explicatives d'un modèle montrant clairement que la relation entre la réponse et la variable explicative x5 est non linéaire. Il semble que la relation avec x1, x2 et x3 soit également non linéaire dans une moindre mesure. (Exemple fictif sur base de données simulées)



b) Inférence statistique et sélection de variables

Lorsqu'on obtient un modèle satisfaisant on peut passer à l'inférence et regarder les résultats. Si on a peu de variables explicatives (par rapport au nombre de données), les p-valeurs et/ou les intervalles de confiance peuvent être suffisants pour évaluer la précision des coefficients et savoir quelles variables peuvent être interprétées.
Lorsqu'on a beaucoup de variables explicatives, il faut souvent passer par une étape de simplification du modèle ("sélection de modèle", "sélection de variables") en sélectionnant les variables explicatives (pour éviter la surparamétrisation ou overfitting). Plusieurs approches sont possibles : AIC, LASSO, Ridge régression,... Ces méthodes permettent souvent de classer les variables par ordre d'importance statistique.
Il peut être aussi utile à ce stade d'analyser les mêmes données avec une autre approche statistique pourvoir quels résultats sont robustes (même conclusion avec les deux approches) et quels résultats le sont moins. L'exploration des données avec d'autres outils peut également vous permettre d'améliorer vos GLMs (pex repérer des interactions potentielles entre variables explicatives). Les arbres de régression sont une approche populaire car très simple à mettre en œuvre et à interpréter même si cette méthode souffre de quelques limitations. Il existe aujourd'hui de très nombreuses autres possibilités...

Exemple de résultats de sélection de modèle pour un modèle visant à expliquer la présence d'un pesticide dans le pollen récolté par des abeilles domestiques en fonction de la période de récolte et de l'origine botanique du pollen.

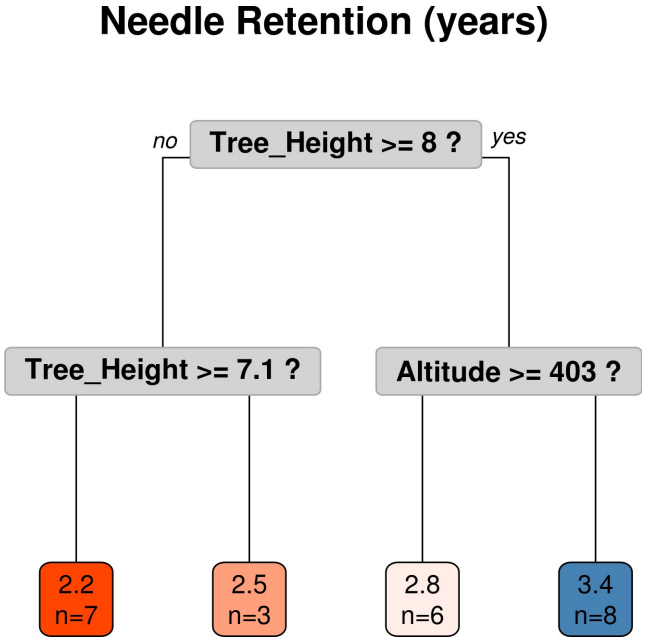
The first 10 best models (Models with delta.AICc<2 are equally supported by the data):

	model	AICc	AICc.delta	AICc.w	sum.w
4	Period+ bra	42.27	0	0.099	0.099
132	Period+ bra+ tar	44.36	2.088	0.035	0.134
20	Period+ bra+ pha	44.44	2.168	0.034	0.168
516	Period+ bra+ api	44.45	2.179	0.033	0.202
68	Period+ bra+ ast	44.48	2.208	0.033	0.235
8	Period+ bra+ ivy	44.51	2.231	0.033	0.267
12	Period+ bra+ tri	44.52	2.244	0.032	0.299
260	Period+ bra+ vic	44.55	2.274	0.032	0.331
36	Period+ bra+ ros	44.57	2.294	0.032	0.363
148	Period+ bra+ pha+ tar	46.58	4.303	0.012	0.375

Model averaging results (variables with w>0.6 are supported by the data)

	freq	w	av.coef	av.se
(Intercept)	1	1	-4.179	2.274
PeriodSepOct	0.5	0.959	-3.282	1.288
bra	0.5	0.952	0.747	0.341
ivy	0.5	0.25	0.002	0.064
tar	0.5	0.248	0.028	0.071
api	0.5	0.244	-0.021	0.067
pha	0.5	0.242	-0.014	0.045
ast	0.5	0.24	0.017	0.072
tri	0.5	0.24	-0.016	0.065
ros	0.5	0.237	0.004	0.048
vic	0.5	0.232	-0.011	0.068

Arbre de régression utilisé pour expliquer l'intensité d'une maladie sur le douglas en fonction de de l'altitude et de la hauteur de l'arbre (plusieurs autres variables explicatives n'ayant pas été retenues). Plus la rétention foliaire (needle retention) est faible plus l'arbre est malade.

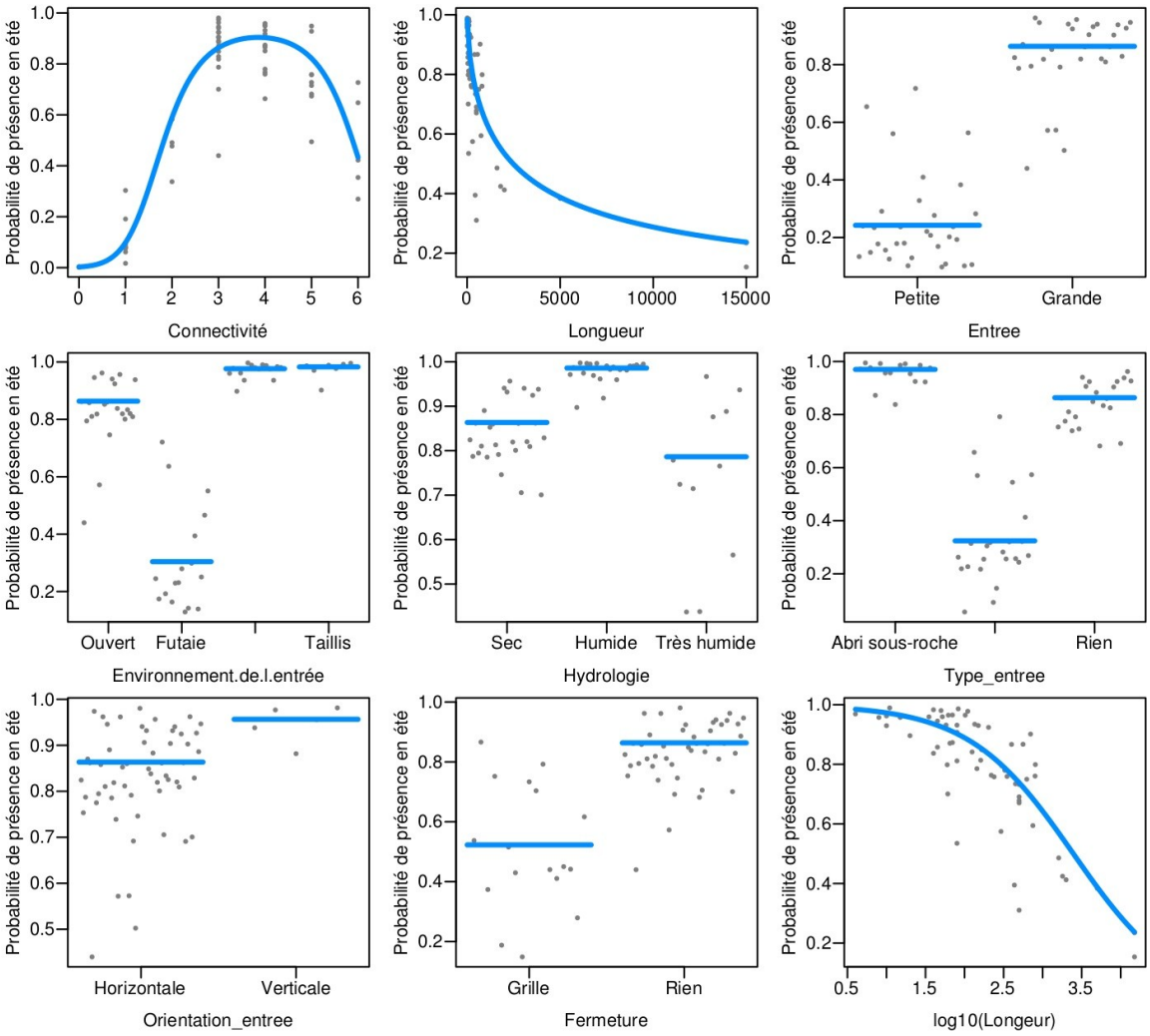


6) Interprétation biologique des résultats et communication

Lorsqu'on a obtenu un modèle final satisfaisant et identifié les variables explicatives supportées par les données, il reste à réaliser l'interprétation biologique des résultats. Pour rappel un résultat "statistiquement significatif" peut très bien être négligeable sur le plan biologique. Il faut à ce stade combiner les éléments de toutes les étapes précédentes (question scientifique, exploration des données, output des modèles,...) pour comprendre la signification biologique des résultats.

L'interprétation fine des coefficients des GLMs et des sorties de la plupart des méthodes statistiques est souvent délicate et la représentation graphique des variables importantes du modèle est souvent une étape indispensable pour comprendre les relations entre les variables. Ils permettent aussi une communication plus efficace vers l'extérieur et ces graphiques doivent être en général plus soignés que les graphiques exploratoires pour cette raison.
Selon l'objectif de l'étude et les besoins des utilisateurs finaux la présentation finale des résultats peut prendre d'autres formes comme par exemple des cartes prédictives.

Exemple de représentation graphique des résultats d'un modèle tentant de déterminer quelles sont les caractéristiques des grottes utilisées par une espèce de chauve-souris Natura 2000 en période estivale. Une telle représentation graphique permet une interprétation facile et précise des résultats ce que permettent plus difficilement une table avec les coefficient d'un modèle.



Prérequis et compétences que vous devrez consolider ou acquérir

Pour pouvoir réaliser l'analyse du jeu de données GBIF vous allez devoir acquérir de nouvelles connaissances ou consolider des connaissances existantes principalement dans vos capacités à manipuler des données et à vérifier la qualité d'un GLM.

Pour vous aider dans votre démarche d'auto-apprentissage on va pendant le cours :

1) examiner ensemble une analyse détaillée d'un jeu de données (abondance d'une libellule menacée : *Coenagrion mercuriale*) selon le même canevas qui vous est demandé pour l'analyse GBIF. On examinera à la fois les résultats de l'analyse, le processus et le code utilisé.
2) revoir ou apprendre les notions nécessaires pour manipuler des données (dans R principalement) et pour utiliser des GLMs en pratique (vérification du modèle, sélection de variables explicatives,...)

On considérera comme prérequis des connaissances de base en R et dans les GLM. Si certaines notions même de base ne vous semblent pas claires, il est important d'avoir une démarche active et de de poser des questions ou de demander des précisions.

Note : Pour des jeux de données très simples, il est possible d'utiliser R (ou tout autre logiciel adapté) juste pour l'analyse de données. On fait par exemple la mise en forme des données dans un tableur/base de données et on importe les données dans R juste pour faire l'analyse simple (anova, test de student, chi carré,...).

Dans un jeu de données complexe comme le jeu de données GBIF vous pourrez sans doute si vous le désirez effectuer certaines étapes préliminaires de nettoyage de données dans un tableur ou dans un environnement qui vous est familier. Mais cela deviendra vite impossible quand on commencera à analyser les données car il s'agit typiquement d'un processus cyclique/itératif : transformation des données --> analyse --> visualisation --> transformation des données.

L'utilisation de R n'est pas obligatoire tant que vous pouvez fournir un résultat final qui soit reproductible et clair.

Utilisation basique de R

Vous avez déjà eu au moins des notions de cette utilisation basique de R dans d'autres cours. On fera ici un rappel rapide. Ces notions de base sont bien entendu indispensables dès que vous voulez utiliser R.

- **Fonctions de base** (`c`, `paste` , `length`, `ncol`, fonctions arithmétiques,...)
- Savoir utiliser correctement **l'aide** (très très important !)
- Savoir **importer** un jeu de données (format csv ou txt) et vérifier le contenu (`read.table`, `summary`, `pairs`, `head`, `dim`). Savoir créer un fichier csv/txt depuis un tableur si nécessaire.
- Connaître très grossièrement les principaux **types d'objets** en R : vecteur, data.frame, liste, matrice, facteur et les principaux types de données : numeric, character, logical.
- Savoir changer les **niveaux d'un facteur** et en particulier leur ordre (fonction `factor` ou `levels`)
- Installer un **package** (fonction `intall.packages`) - charger un package (fonction `library`)
- Savoir changer les **noms de colonnes** (`colnames`)
- Notions de **base des graphiques**

Manipulation des données

Indispensable

Dans R, pendant votre analyse, les fonctions `[ ]`, `merge` et `apply` vous seront indispensables pour les manipulations de données de base. On les verra en détail.

- **Extraire des colonnes**. Pex pour ne garder que les colonnes de températures et de pluviométrie dans le jeu de données.
  R : fonction `[ ]`
- **Extraire des lignes** (typiquement selon un critère logique pex ne garder que les données où l'année est > 1980).
  R : fonction `[ ]` - Tableur : filtres
- **Fusionner des tables** de données de dimensions différentes selon les valeurs d'une colonne.
  Pex fusionner votre tableau carrés UTM x espèces et votre tableau UTM x variables environnementales selon la valeur de la colonne UTM.
  R : `merge` - tableur : possible mais compliqué et avec possibilités réduites !
- **Calculer des valeurs sur les lignes ou les colonnes**.
  Pex pour chaque ligne calculer la moyenne ou la médiane ou la variabilité, etc... des températures des 12 mois de l'année.
  R : fonction `apply`
- **if: exécution conditionnelle**
  Pex transformer des valeurs d'abondance en présence/absence (si la valeur est > 0, mettre 1 si non mettre 0):
  R : `d <- ifelse(d>0, 1, 0)` - tableur : fonction SI ou IF

Très utile

Les approches suivantes sont souvent très utiles mais vous seront moins indispensables parce que vos jeux de données on déjà été mis en forme par mes soins. On fera une petite démonstration des possibilité sans entrer dans trop de détails.

- Éliminer les **doublons** dans un jeu de données. Par exemple ne garder que les combinaisons uniques de carrés UTM et espèces
  R : `unique` - Tableur : collage spécial
- **Agréger** des données (pex : nombre d'observations pour chaque carré UTM)
  R : `aggregate` ou `reshape2:::dcast` - Tableur : Pilote de données/tableau croisé dynamique
- **Tableaux croisés** (pex passez d'un tableau avec des colonnes : carré UTM, espèce, date vers un tableau avec les carrés UTM en ligne, les espèces en colonnes et le nombre de dates différentes.
  R : `reshape2:::dcast` - Tableur : Pilote de données/tableau croisé dynamique
- **Boucles** : pour pouvoir répéter des opérations similaires.
  Pex créer un histogramme de chacune de vos variables explicatives, faire une carte pour chacune des 200 espèces de votre jeu de données,...
  R : `for (i in 1:ncol(d)) { hist(d[,i] }`
- **Fonctions** : créer vos propres fonctions.
  Par exemple une fonction pour calculer un indice de diversité du paysage (pex indice de simpson) : `simpson <- function(x) { 1 - sum((x/sum(x))^2)}`

Exploration des données

On combinera ici des approches manipulations de données et simples statistiques descriptives (en particulier matrices de corrélation) et des méthodes simples de visualisation de données (scatterplot matrix et heatmap).

Ce sera aussi l'endroit où vous pourrez mettre en pratique les méthodes d'analyse multivariées que vous avez apprises dans d'autres cours (PCA, CA, clustering,...)

Modèles linéaires généralisés

Les bases

Les bases sont sensées être connues via ce que vous avez appris dans d'autres cours.

- différents types de GLM (gaussien, Poisson, binomial) et quand on peut les utiliser
- interprétation des coefficients de ces modèles (pour des variables continues ou qualitatives)
- inférence (p-valeurs, comparaison de modèles emboîtés,...)
- représentation graphique des ces modèles (par exemple avec le package `visreg`)

Vérifier les conditions d'applications et les problèmes potentiels

Vous avez déjà quelques notions dans ce domaine, on va les approfondir ensemble. On va voir quels sont les principaux types de problèmes possibles, comment les diagnostiquer et quelques pistes de solutions en cas de problème.

- **Linéarité** - La relation entre réponse et variables explicatives quantitatives est-elle bien à peu près linéaire ? (graphiques résidus vs variables explicatives)
- **Surdispersion** et **homoscedasticité** : Les hypothèses de variabilité des résidus sont-elles respectées (homogénéité de la variance résiduelle pour un modèle gaussien, pas de surdispersion pour les modèles de Poisson et binomiaux)
- La **distribution des résidus** (normalité etc...) suit-elle à peu près la distribution attendue (QQ plots) ?
- Y a-t-il des problèmes de **multicolinéarité** ? (corrélations multiples entre variables explicatives - VIFs) ?
- Y a-t-il des **points extrêmes** qui pourraient influencer fortement les résultats à eux seuls ?
- Est-ce qu'on peut s'attendre à des problèmes de **surparamétrisation** (**overfitting** : trop de variables explicatives). Faut il faire de la sélection de modèle.
- Les résidus sont-ils bien **indépendants** ? (pas de structure hiérarchique, mesures répétées, indépendance spatiale)
- Est-il nécessaire ou utile de **centrer** ou de **standardiser** les variables explicatives ?

Sélection de variables (feature selection)

Lorsqu'on a beaucoup trop de variables explicatives on veut en général savoir lesquelles sont les plus "importantes". Typiquement aussi on a un risque de surparamétrisation : le modèle prédit très bien les données mais n'est pas extrapolable à d'autres jeux de données. On utilise alors des méthodes qui permettent de simplifier le modèle. Une série de méthodes relativement simples à comprendre et à mettre en œuvre utilisent l'AIC (Aikaike information criterion).
On comparera ces méthodes entre elles et avec des tests d'hypothèse nulle classique.
Ceux d'entre vous qui ont suivi le cours de Biométrie à l'UCL (LMAT1375 - N. Schtikzelle) ont déjà vu ces notions.

Exercice sur les données GBIF

Détails pratiques

Le jeu de données environnemental est décrit en détails dans le fichier "UTM5data_README.pdf" et les données sont disponibles dans le jeu de données "UTM5data.csv".

Les jeux de données espèces GBIF se trouvent dans GBIF/data/GBIF_data. Chaque dossier correspond à un groupe taxonomique différent et contient un fichier README.md décrivant le contenu des différents jeux de données. Ces dossiers contiennent également un fichier pdf montrant le processus de manipulation et de nettoyage de données depuis le jeu de données brut et décrivant quelques caractéristiques du jeu de données.
Je vous conseille de regarder le pdf pour le groupe des Coccinellidae qui montre l'utilisation d'une série de commandes simples pour manipuler des données dont vous aurez besoin (`merge`, `[ ]`, `table`, ...) ainsi que quelques exemples simples de graphiques (notamment des cartes).
Il est important également d'aller lire la description du jeu de données sur le site GBIF pour comprendre comment elles ont été récoltées.

Choix des données

Choisissez un groupe taxonomique parmi les jeux de données GBIF disponibles et une question sur laquelle vous voulez travailler. Le choix le plus évident est d'étudier la distribution d'une ou plusieurs espèces et d'essayer de comprendre quels facteurs environnementaux déterminent cette distribution. Mais vous pouvez aussi par exemple travailler sur la diversité (nombre d'espèces dans un carré pex) ou toute autre variable susceptible de vous intéresser.
Ce genre de base de données permet aussi d'étudier l'évolution temporelle des populations. Cependant ce genre d'approche demande l'utilisation de techniques statistiques plus complexes permettant de prendre en compte la nature répétée des données (non indépendance) et les nombreuses données manquantes (modèles mixtes, GEEs,...).

Commentaires
Vous devez choisir une espèce qui n'est ni trop commune ni trop rare. Il faut un minimum de variabilité dans les données. L'exercice sera d'autant plus intéressant que vous arriverez au final à trouver au moins une ou deux variables environnementales expliquant la distribution. Les espèces plus rares ont souvent des exigences écologiques plus spécifiques et sont de ce point de vue de meilleurs candidats. Ce sont aussi souvent les espèces d'intérêt en biologie de la conservation.

Si vous voulez travailler sur d'autres jeux de données n'hésitez pas à faire des propositions. Il faut que ce soit un vrai jeu de données avec un nombre de variables conséquent avec une question compatible avec les approches développées ici (les cas sont nombreux!).

Description du contexte et de la/les question(s)

Q01 - Décrivez brièvement le contexte (ea espèce et/ou variables choisies), les jeux de données utilisés (GBIF et environnement), la manière dont ces données ont été collectées et la ou les questions qu'on se pose (objectifs).

Commentaires
Cette partie doit permettre à une personne extérieure au cours/à votre projet de comprendre quels sont vos objectifs, d'où viennent les données et comment elles ont été récoltées et pré-traitées. Si vous utilisez les jeux de données GBIF déjà nettoyés, précisez les règles qui ont été utilisées pour les nettoyer (voir fichier README). Vous devez aussi comprendre comment les données originales ont été récoltées (voir sur le site GBIF).
Au fur et à mesure que vous progresserez dans votre analyse et l'exploration des données vous devrez très probablement éditer cette partie pour y incorporer les changements effectués.
Même lorsque vous travaillez uniquement pour vous même, c'est une bonne habitude d'organiser vos données et votre analyse comme vous le feriez pour une personne étrangère. Cette personne étrangère est souvent simplement votre "futur vous-même". Il est en effet souvent difficile de se rappeler exactement ce qu'on a fait et pourquoi quelques mois ou quelques semaines plus tard... N'oubliez pas non plus que les vols, incendies, pannes informatiques, accidents... n'arrivent pas qu'aux autres. Faites régulièrement des copies physiquement séparées (différents bâtiments, cloud) de vos données et de votre travail. N'oubliez pas non plus qu'il ne faut JAMAIS travailler sur vos données brutes mais TOUJOURS sur une copie pour éviter de corrompre, détruire, mélanger, perdre vos données brutes...

Q02 - En première approche, sous quelle forme présenterez-vous vos résultats finaux pour pour pouvoir répondre à votre question ?

Commentaires
Dans le cadre de cet exercice imposé, cette question est un peu artificielle... Dans toute autre circonstance l'idée est d'imaginer à l'avance quelle est la forme la plus efficace de communication des résultats en fonction du public visé. Un crayon et un papier suffisent à ce stade pour imaginer à quoi pourrait ressembler le résultat final. NB : cette vision du résultat final ne doit pas être figée, elle peut évoluer au cours de l'analyse. Le fait d'imaginer à quoi devra ressembler le résultat final aide souvent à préciser les questions lorsqu'elles sont un peu floues ou à les hiérarchiser.

Q03 - En première approche, vers quel type d'analyse statistique vous orienteriez-vous à ce stade pour répondre à votre question ?

Commentaires
Analyse supervisée/non supervisée ? Analyse multivariée/univariée ? Si vous envisagez d'utiliser des GLMs, vers quelle famille de GLMs vous orienteriez-vous en première approche ? Bas besoin de s'étendre ici... Le but est juste d'avoir une idée de vers où on va pendant la phase d'exploration des données.

Import et vérification des données

Q04 - Importez les données dans R et vérifiez leur contenu.

- Est-ce que toutes les données ont été importées (nombre de lignes, nombre de colonnes)
- Est-ce qu'il y a des variables qualitatives et si oui ont-elles bien été importées comme des facteurs dans R ?
- Est-ce que les variables quantitatives ont bien été importées comme des nombres ?
- Est-ce qu'il y a des valeurs manquantes (dans quelles variables?). Si le jeu de données original contenait des cases vides il faut s'assurer qu'elles ont été importées correctement (valeurs manquantes, simple champ texte vide)
- Examinez les valeurs pour voir si il n'y a pas manifestement des valeurs impossibles

Commentaire
Utilisez par exemple la fonction R `read.table` pour lire un fichier texte/csv. Il faut bien noter dans le fichier texte original les caractères séparateurs de champs et de décimales (pex : `sep = "\t"` pour des tabulations et `dec = ","`), indiquer que la première ligne contient les noms des colonnes (`header = TRUE`). Si les caractères accentués ne sont pas lus correctement il faut probablement spécifier l'encodage des caractères avec `encoding = "utf8"` ou `encoding = "latin1"`. Si toutes les lignes ne sont pas lues (warning message parlant de caractère "eof" - end of file pex) il peut être utile de spécifier `quote = ""`.
La fonction `dim` vous donne les dimensions du jeu de données et `head` les premières lignes.
La fonction `summary` devrait toujours être exécutée après l'import de données. Elle vous montre où sont les valeurs manquantes, quel est le type de chaque variables en fonction du type d'output (numérique : moyenne et quantiles, facteur : fréquence des différents niveaux, caractères) et vous permet par exemple de voir si les valeurs minimales et maximales des données sont plausibles. Certaines personnes préfèrent la fonction `str`. Il est fréquent que des variables qualitatives soient encodées sous forme de nombres (par exemple un numéro de site, la moyenne de ces numéro n'a pas de sens) et doivent être convertis en facteurs (fonction `factor`). Il peut être aussi utile de changer l'ordre des niveau d'un facteur par exemple :
`mydata$size <- factor(mydata$size, levels = c("small", "medium", "tall"))`
N'oubliez pas de spécifier votre répertoire de travail (`setwd`) et n'oubliez pas que le séparateur de chemins de Windows (le backslash \) doit soit être doublé : \\ soit être transformé en slashes : / .

Nettoyage des données

Q05 - En fonction de ce que vous savez sur la manière dont les jeux de données GBIF ont été récoltés décrivez les problèmes et/ou les biais potentiels dans votre jeu de données et les conséquences possibles pour la question que vous étudiez. Est-il possible de mitiger certains de ces problèmes en modifiant le jeu de données ou en modifiant votre manière d'analyser les données ?

Commentaires
Un des problèmes principaux avec ce genre de jeu de données (observations naturalistes occasionnelles) est que l'effort d'échantillonnage n'est en général pas homogène et pas aléatoire du tout ! En conséquence, le fait qu'il n'y ait pas d'observations d'une espèce dans un carré ne signifie pas automatiquement que l'espèce était réellement absente ("fausses absences"). Il se peut simplement que ce carré n'ait jamais été visité ou pas suffisamment inventorié. De même, le fait qu'une espèce ait été observée une seule fois dans un carré ne signifie pas automatiquement que les conditions environnementales de ce carré soient favorable à cette espèce ("fausses présences"). Il peut s'agir simplement d'un individu en transit ("espèces touristes"). Une manière classique de diminuer le problème des fausses absences est de ne garder comme "absences" que les carrés où l'effort d'échantillonnage a été "suffisamment" important. Ce seuil d'échantillonnage est malheureusement assez arbitraire. Plus on est exigeant sur l'effort d'échantillonnage nécessaire plus on doit éliminer des observations et moins on a de données. Il faut donc trouver un juste milieu...

Le problème des fausses présences est souvent considéré comme moins problématique (en particulier à une échelle de 5 km²). Mais vous pouvez aussi déterminer des seuils de présence si vous pensez que c'est important.
Les choix à poser ici sont difficiles et très subjectifs et peuvent avoir une grande influence sur le résultat final. Il est parfois utile de tester l'effet de ces choix arbitraires sur le résultat final en effectuant la même analyse avec plusieurs seuils. Il est important de se rendre compte de l'importance de ce genre de choix pour une lecture critique de la littérature scientifique. Ceci montre également l'importance pour toute publication scientifique de fournir les données et le détail des analyses de façon à pouvoir reproduire les résultats (très facile aujourd'hui avec les "supplementary materials" en ligne).
Il existe des approches alternatives. Il est par exemple fréquent de n'utiliser que les données de présence et de sélectionner au hasard des carrés de référence pour caractériser un "background environnemental". Si on a des données réelles d'absence il est toutefois toujours préférable de les utiliser.

Q06 - Posez-vous les mêmes questions sur le jeu de données environnemental.

Commentaires
Certains carrés ne couvrent qu'une toute petite partie de la Belgique. Or certaines variables environnementales ne sont disponibles que pour la Belgique...

Exploration des données et choix des variables explicatives

Il n'est ni raisonnable ni utile d'utiliser les ~150 variables environnementales telles quelles comme variables explicatives. Le choix des variables explicatives doit se faire selon une combinaison de deux catégories de critères : 1) biologiques et 2) statistiques.
On commence en général par faire une première sélection de variables explicatives sur base de leur importance biologique (quelles variables sont potentiellement importantes pour déterminer la distribution de votre espèce ?). On vérifie ensuite les problèmes statistiques de ces variables et on agit en conséquence (combinaison de variables, transformations, éliminations,...). Le processus est en général itératif, l'exploration des données et les contraintes statistiques nourrissant la réflexion sur le choix des variables sur base de critères biologiques.

Q07 - Explorez les données espèces (GBIF) et commentez vos observations.

Commentaires
L'exploration est relativement limitée ici une fois que vous avez choisi votre espèce ou votre variable d'intérêt. Il peut-être cependant utile de visualiser la distribution de votre espèce/variable réponse. Un simple plot des coordonnées xy peut suffire comme visualisation grossière. Il est cependant relativement facile de positionner ces points sur un fond de carte. Voir les fichiers GBIFdata_CreateData.pdf pour quelques exemples. Des couches cartographiques de la Belgique sont mises à votre disposition sur la plate-forme moodle.
Il peut être aussi utile de visualiser la répartition des carrés utilisés comme "absences". Est-ce que leur distribution est relativement uniforme ?

Q08 - Explorez les données environnementales et commentez vos observations.

Vous devez au minimum vous faire une idée des corrélations entre variables explicatives et une idée de leur distribution (histogrammes, density plots) et prendre les dispositions nécessaires en cas de problèmes potentiels pour votre futur modèle.

Commentaires
NB1 : vous n'êtes obligés d'explorer toutes les variables ! Vous pouvez choisir de laisser tomber certaines variables et vous concentrer sur les variables que vous pensez être importantes sur le plan biologique.

NB2 : vous n'êtes pas limités aux variables telles qu'elles sont dans le jeu de données. Il est même recommandé de combiner ces variables explicatives pour créer de nouvelles variables qui ont plus de sens sur le plan biologique (pex faire la somme de certains types de land cover). La série de variable bio01 à bio19 peuvent vous donner des idées pour les données climatiques.

NB3 : c'est une des étapes les plus importantes et les plus consommatrices de temps... Ceci explique aussi la longueur de ces commentaires...

Si des variables sont trop fortement corrélées il peut être nécessaire de les combiner (pex calculer les températures moyennes de plusieurs mois ou sommer les % de recouvrement d'occupation du sol). Si vous laissez tomber certaines variables corrélées il faut s'en rappeler pour l'interprétation des résultats.
NB : l'examen des corrélations deux à deux (colinéarité) n'est qu'une première étape. En effet c'est bien la "corrélation multiple" (multicolinéarité) entre une variable explicative et l'ensemble des autres variables explicatives qui posera des problèmes dans le futur modèle (augmentation des erreurs standard, instabilité du modèle, ...). Vous pouvez déjà à cette étape créer un modèle provisoire et calculer les "Variance Inflation Factors" (VIFs) pour évaluer à quel point la multicolinéarité est problématique (fonction `vif` du package `car`).

La distribution des variables explicatives ne doit pas spécialement suivre une distribution gaussienne. L'idéal est une distribution uniforme (chaque valeur de x a la même probabilité d'être observée) mais c'est rarement le cas pour des données non expérimentales. Il est généralement préférable éviter les distribution très asymétriques avec une ou deux valeurs extrêmes qui peuvent fortement influencer le résultat final. Dans ce cas on peut transformer la variable par exemple avec une transformation minimisant l'écart des grandes valeurs (log, sqrt). Il est par exemple fréquent d'appliquer une transformation racine carrée à des données de surface et des transformations log à des mesures physico-chimiques. On peut aussi la transformer en variable qualitative en créant 2 ou plusieurs classes. Lorsqu'on a une variable continue avec une large dominance de 0 et peu de données non nulles ou des données non nulles très disparates il est parfois utile de transformer cette variable en simples données de présence/absence. Il est parfois utile d'attendre la construction du premier modèle avant de faire des transformations pour évaluer leur effet sur la linéarité de la relation.

Si vous avez des variables qualitatives (y compris présence/absence) il faut vérifier que chaque classe est suffisamment représentée dans le jeu de données. Dans le cas contraire il faudra soit regrouper des classes soit laisser tomber une partie des données. Si vous comptez ajouter des interactions entre variables qualitatives il faut de plus s'assurer que vous avez des observations pour chaque combinaison des différentes classes.
La fonction `table` vous permet de facilement compter le nombre de données pour chaque niveau d'un facteur.

La fonction `pairs2` (fournie dans le script "`mytoolbox.R`") est particulièrement utile à cette étape car elle permet de visualiser les corrélations et la distribution des variables. `pairs2(mydata, reorder = TRUE)` permet de réordonner les variables explicatives de façon à regrouper les variables fortement corrélées. Vous pouvez bien sûr personnaliser cette fonction selon vos préférences. Le résultat devient cependant vite illisible (et les calculs très lents) quand vous avez beaucoup de variables. Une `heatmap` sur une matrice de corrélation (fonction `cor`) est souvent alors une bonne option pour visualiser les corrélations. La fonction `hist` permet de créer des histogrammes pour examiner la distribution des variables. La fonction `corheatmap` fournie dans le script '`mytoolbox.R`' vous permet de facilement faire ce genre de représentation. Pensez à utiliser une boucle `for` si vous devez faire des graphiques en série et `par(mfrow = c(6, 5))` pour diviser la fenêtre graphique par exemple en 6 lignes et 5 colonnes. Les paramètres graphiques `mar` et `mgp` vous permettent d'adapter la taille des marges et la position des étiquettes et légendes d'axes et `las` l'orientation des étiquettes : `par(mar = c(3,3,1,1), mgp = c(1.8,0.6,0), las = 1)`. Le pdf sur les données coccinelles contient un exemple de carte dans une boucle et d'utilisation des paramètres graphiques. L'analyse des données *C.mercuriale* montre également plusieurs exemples.

Au delà de l'exploration des colonnes de votre jeu de données, il peut être utile surtout pour l'interprétation finale des résultats d'explorer les relations entre les lignes de votre jeu de données (Q mode analysis - similarités entre les carrés UTM). Clustering, heatmaps et ordinations sont particulièrement utiles à ce stade. Ceci dit cette étape est moins cruciale pour la construction d'un GLM que l'exploration des colonnes (R mode analysis).

Q09 - Suite à la phase d'exploration des données, choisissez un nombre raisonnable (maximum 20) de variables environnementales que vous pensez pouvoir expliquer la distribution de votre espèce et qui ne posent pas à priori de problèmes statistiques.

Construction et examen critique d'un modèle

Q10 - Construisez un premier modèle et évaluez sa qualité et les conditions d'applications. Adaptez progressivement votre modèle en conséquence.

NB il n'est pas possible de donner ici une séquence logique et unique des éléments à vérifier. Il faut s'adapter au cas par cas...

Vérifiez <u>obligatoirement</u> les points suivants :

Q10a - Variabilité des résidus

Vérifiez que les conditions liées à la variabilité des résidus sont respectées : homogénéité des variances pour un modèle gaussien (sur base d'un graphique résidus vs valeurs prédites par exemple) et surdispersion pour les modèles de Poisson ou binomiaux (calcul du coefficient de surdispersion, graphique des résidus de Pearson).

Comment peut-on interpréter un coefficient de surdispersion et quelles sont les conséquences si ce coefficient est trop élevé ?

Si vous rencontrez des problèmes essayez d'appliquer les solutions vues au cours : changer de type de modèle ou adapter les méthodes d'inférence, transformer la variable réponse, ajouter des variables explicatives ou des interactions,...
NB : avec un modèle binomial sur données binaires vous ne devriez jamais avoir de problèmes de surdispersion... Avec des modèles gaussiens la transformation de la variable réponse est une pratique courante. Elle doit alors idéalement être faite avant d'explorer la linéarité des relations (point suivant)
La fonction `overdisp` (dans `mytoolbox.R`) calcule le coefficient de surdispersion de deux manières différentes. La fonction `diagplot` (dans `mytoolbox.R`) vous donne un graphique des résidus en fonction des valeurs prédites (pour évaluer la dispersion) et le coefficient de surdispersion pour les modèles de Poisson et binomiaux.

Q10b - Linéarité et points extrêmes

Vérifiez que les relations entre votre réponse et vos variables explicatives quantitatives sont bien approximativement linéaires et qu'il n'y a pas de points extrêmes qui pourraient influencer trop fortement la relation.

Vous pouvez pour ce faire utiliser des graphiques résidus ~ variable explicative. Si les résidus ne se distribuent pas de manière uniforme de part et d'autres de leur moyenne (0 ou valeur prédite) on peut suspecter un problème de non linéarité (la relation ne prend pas la forme d'une ligne droite) ou de non additivité (il manque des interactions). La fonction `diagplot2` (dans `mytoolbox.R`) vous donne automatiquement ces graphiques pour toutes les variables explicatives du modèle. On peut visualiser les résidus `diagplot2(monmodèle, partial = FALSE)` ou les résidus partiels `diagplot2(monmodèle)` (comportement par défaut) qui permettent de visualiser le sens de la relation.
En cas de problème, on transforme en général les variables explicatives (log, sqrt, ^2, exp,...), on utilise une régression polynomiale, on convertit une variable quantitative en variable qualitative, on ajoute des interactions, on se tourne vers d'autres méthodes statistiques pour les cas complexes (régression non linéaire, GAMs, ensemble modelling, ...), ... Pour les points extrêmes il peut être utile de refaire l'analyse avec ou sans ces points extrêmes pour évaluer leur effet sur le résultat final.
NB : la relation y~x ne doit pas être parfaitement droite. Une relation à peu près linéaire est souvent une approximation suffisante. Le but est d'éviter les relations les plus manifestement non linéaires. Comparez les R^2, RMSE, AUC,... ou les AIC quand vous faites des transformations de variables explicatives (pas du y!) pour évaluer le gain en qualité. Les graphiques de résidus de modèles binomiaux sont souvent plus difficile à interpréter.

Q10c - Multicolinéarité

Calculez les VIFs pour vérifier que la multicolinéarité n'est pas trop importante.

Comment interprète-t-on ces valeurs et quel est le risque si certaines valeurs de VIFs sont trop élevées ?

NB si vous transformez les variables explicatives il faut calculer les VIFs sur ces variables transformées ...
Vous avez normalement déjà minimisé ce genre de problèmes lors de l'analyse exploratoire des données et du choix des variables explicatives.
On voit régulièrement le chiffre 10 comme seuil acceptable. Personnellement je trouve que c'est déjà très élevé... La fonction `vif` du package `car` vous permet de cacluler des VIFs et des VIFs généralisés (GVIFs) lorsqu'il y a des variables qualitatives.

Q10d - Distribution des résidus

Vérifiez que la distribution des résidus correspond approximativement à la distribution attendue par votre modèle (QQ-plot, histogramme des résidus).

La fonction `diagplot` (dans `mytoolbox.R`) vous donne un QQ-plot pour les distributions gaussienne, binomiale et de Poisson (nécessite le package `mgcv`)

De manière générale le package `car` offre de nombreuses fonctions intéressantes pour diagnostiquer les problèmes dans les GLMs.

Quelques commentaires supplémentaires :

Indépendance
Vos répétitions (les lignes de votre tableau) doivent être indépendantes. Un cas typique de non indépendance sont les mesures répétées (pex : plusieurs mesures au cours du temps ou non sur un même site ou un même individu) et les designs hiérarchiques (avec des groupes d'observations : enfants groupés dans des classes groupées dans des écoles, individus groupés dans des familles (mêmes parents) groupées dans des populations). Il n'y a pas vraiment de diagnostique, c'est la structure de récolte des données qui implique la non indépendance. Il faut alors soit utiliser des méthodes adaptées (modèles mixtes pex) ou regrouper les données de façon à ce que chaque ligne représente une observation indépendante (pex sommer toutes les captures des pièges posés sur un même site). Dans certains cas extrêmes il n'y a pratiquement aucune vraie répétition. Il n'y a pas vraiment de solution dans ces cas là et il faut résister à la tentation d'analyser malgré tout ces données comme si elles étaient indépendantes ou alors être extrêmement prudent dans l'interprétation (les résultats ne sont pas extrapolables à d'autres cas).
Une autre source de non indépendance est la corrélation spatiale. Deux échantillons proches dans l'espace sont susceptibles d'avoir des valeurs similaires. Pour vérifier ce point on doit mesurer la corrélation entre les résidus en fonction de la distance entre eux, typiquement au moyen d'un corrélogramme. On abordera pas ce point dans le cadre de ce cours.

Erreurs standard énormes - problèmes de convergence
Regardez les erreurs standard des coefficients de votre modèle avec un regard critique. Il arrive qu'elles soient énormes par rapport au coefficient correspondant ce qui indique qu'il y a vraisemblablement un problème quelque part mais l'origine peut être très variée... Il s'agit souvent de problèmes de multicolinéarité ou d'un modèle trop complexe par rapport à une jeu de données trop petit (par exemple des interactions complexes avec peu de données pour chaque cas de figure), d'un cas de discrimination parfaite de données binaires (cfr infra),....

Discrimination parfaite pour des données binaires
Assez paradoxalement les GLMs binomiaux sur des données binaires peuvent rencontrer des problèmes lorsque une ou plusieurs variables explicatives prédisent parfaitement ou presque la présence ou l'absence (discrimination parfaite). Typiquement le modèle ne converge pas (R vous donne un avertissement) et/ou présente des erreurs standard énormes.

Il faut toujours un minimum de variabilité
Pour des données binaires (pex présence/absence), la distribution binomiale est presque assurée mais vous devez vous assurer de ne pas avoir ni trop ni trop peu de chaque catégorie. De même dans une variable réponse quantitative il faut vous assurer qu'il y a un minimum de variation sans quoi il n'y a rien à étudier... Si vous mesurez 3 mâles qui font exactement 5cm et 3 femelles qui font exactement 6cm, impossible de faire un test statistique pour comparer leur taille.

Excès de 0
Les données de comptage (pex : nombre d'individus - distribution de Poisson à priori) contiennent souvent trop de valeurs nulles. Il existe des méthodes adaptées pour ce genre de données (Zero infalted poisson models). Si vous avez vraiment beaucoup de 0 il peut être judicieux et plus facile de travailler en présence/absence (modèle binomial). Parfois la cause est simplement que vos unités d'échantillonnage sont trop petites et il faut dans ce cas, si possible, les regrouper. Par exemple au lieu de compter le nombre de pucerons par feuille, il faudrait compter le nombre de pucerons par 100 feuilles... Il existe un continuum entre des données binaires et des données de comptage.

Centrage/standardisation
Si vous avez des interactions ou des termes polynomiaux et que les VIFs concernant ces variables sont élevés, il peut être utile de centrer les variables correspondantes pour réduire les corrélations. En dehors de ces cas et si vous interprétez les résultats de votre modèle sur base de graphiques, le centrage ou la standardisation des variables explicatives est en général inutile. Certaines personnes préfèrent cependant standardiser systématiquement les variables explicatives de façon à pouvoir comparer directement les coefficients du modèle.

Inférence et sélection de modèle

Une fois que vous avez un modèle que vous estimez globalement satisfaisant vous pouvez passez à l'étape suivante qui est de rechercher quelles sont les variables qui semblent effectivement pouvoir expliquer au moins partiellement la distribution de votre espèce (sans que la relation observées puisse être due au hasard).

Q11 - Réalisez un test d'hypothèse nulle pour chaque variable de votre modèle. Quelle est l'hypothèse précise que vous testez ? Quel problème peut-on rencontrer lorsqu'on réalise de tels tests avec un grand nombre de variables explicatives ?

Vous pouvez utiliser simplement la fonction `summary`. Alternativement vous pouvez faire une comparaison de modèles avec et sans chaque variable explicative (test de rapport de vraisemblance) avec les fonctions suivantes : `drop1(m, test = "Chisq")` ou `car:::Anova(m)` où m est le nom de l'objet dans lequel vous avez sauvé votre modèle. Attention si vous avez des interactions soyez extrêmement prudents dans l'interprétation des variables impliquées.

Q12 - Réalisez une sélection du meilleur modèle par une sélection par étape sur base de l'AIC (stepwise AIC selection). Quelles sont les variables sélectionnées par cette méthode ? Quel(s) problème(s) peut-on rencontrer avec cette approche.

Vous pouvez simplement utiliser `step(m)` avec les options par défaut qui utilise l'AIC (et pas l'AICc)

Q13 - Calculez tous les modèles possibles, leur AICc et le poids du modèle (model AICc weight). Calculez l'importance relative de chaque variable (variable AICc weight) et les model averaged coefficients. Comparez les résultats avec les deux approches précédentes.

Est-ce que le meilleur modèle sélectionné ici est le même que celui de la question précédente ? Qu'est-ce qui pourrait amener à des résultats différents ? Combien avez-vous de modèles dont la différence d'AICc par rapport au meilleur modèle est <2 (ceci vous donne une idée de l'incertitude sur le choix du meilleur modèle) ? Quel est le poids du meilleur modèle et comment interprétez-vous cette valeur ?

Examinez maintenant le poids des variables. On considérera qu'une variable avec un poids de plus de 0.6 est "supportée par les données". Si vous avez des interactions dans votre modèle, comparez plutôt le poids du modèle à sa fréquence dans la série de modèles. Est-ce que ces variables sont les mêmes que les variables des deux questions précédentes (variables du meilleur modèle sélectionné par stepwise selection et variables "significatives" dans les tests d'hypothèse nulle) ? Discutez les différences.

Vous pouvez utiliser la fonction `model.select` fournie dans le cadre de ce cours ou bien une combinaison de fonctions du package `MuMin`.
Attention si vous avez beaucoup de variables explicatives les calculs peuvent prendre beaucoup de temps... Faites un premier essai sur un sous groupe de variables pour évaluer le temps nécessaire (le temps double à peu près pour chaque variable explicative ajoutée...). Quand on a trop de variables seule l'approche stepwise est possible...

Interprétation des résultats et communication

Une fois que vous avez une idée des variables qui peuvent potentiellement expliquer votre variable réponse, il faut procéder à l'interprétation biologique des résultats. Si il s'avère à l'étape précédente que vos données ne vous permettent de relier aucune des variables explicatives à votre variable réponse choisissez une ou deux variables pour continuer l'exercice (en gardant bien sûr toutes les réserves nécessaires sur les résultats).

Q14 - Examinez les coefficients (model averaged coefficients) des variables qui semblent les plus importantes. Comment pouvez-vous interpréter ces coefficients (sans faire de calculs il s'agit juste d'une interprétation très grossière à ce stade) ?

Q15 - Faites une représentation graphique de vos données et de votre modèle et interprétez les résultats en termes biologiques. Concentrez-vous sur les variables les plus importantes. Il peut être aussi intéressant de discuter les variables qui ne sont pas "significatives" sur le plan statistique mais qui montrent malgré tout une relation avec la variable réponse. Il faut alors s'interroger sur la raison qui fait que ces variables ne sont pas supportées par les données.

La fonction `visreg` du package du même nom est très pratique pour faire une représentation rapide de toutes les variables du modèle avec des résidus partiels. La fonction fixe les autres variables continues à leur valeur médiane et les variables qualitatives à la classe la plus fréquente. Ce dernier choix, un peu arbitraire, peut avoir des effets dramatiques sur la représentation graphique des autres variables. Vous pouvez cependant changer ce choix par défaut. Il peut-être intéressant d'explorer un peu. Si vous avez des interactions entre certaines variables vous devez également explorer plus en détails la gamme de valeurs possibles et ne pas vous contenter de la première représentation graphique.
Malheureusement `visreg` ne permet de représenter le modèle sur base des "model averaged coefficients". Vous pouvez construire "à la main" une représentation graphique pour les variables les plus importantes avec les valeurs réellement observées plutôt que les résidus partiels fournis par `visreg`.