Introduction

I.1. Pourquoi des analyses numériques en sciences environnementales ?

I.1.1. *Le chercheur face à ses données*

La recherche « environnementale » passe souvent par un échantillonnage sur le terrain ou bien l'acquisition automatisée de multiples paramètres (par exemple physico-chimiques, pédologiques, biologiques, etc.) à différentes stations et/ou différentes dates. Chaque station (ou date) constitue une *observation/objet/élément* qui sera donc caractérisée par plusieurs *descripteurs/variables/paramètres* (abiotiques ou biotiques).

Caractériser la variation d'un ou deux de ces descripteurs est facile à appréhender par une simple analyse graphique. Il est en effet possible d'observer, par exemple, (i) les fluctuations des débits fluviaux d'une rivière au cours des différents mois d'une année, (ii) de déterminer quelles dates présentent des débits similaires et donc se ressemblent d'un point de vue de ce paramètre (par exemple, périodes de crues et d'étiage) ou bien, (iii) d'analyser si une relation existe entre les débits fluviaux et la turbidité en un point de l'estuaire (voir la figure I.1A).

Toutefois, le chercheur ne se cantonne rarement qu'à la récolte d'un ou deux paramètres. Chaque paramètre supplémentaire représente une dimension de plus. Bien que la représentation graphique soit encore possible mais plus difficilement accessible pour trois paramètres (voir la figure I.1B), la représentation graphique *multidimensionnelle* n'est plus possible au-delà de ce nombre. L'analyse graphique devient alors laborieuse puisqu'elle passe par l'analyse de la projection en 2 dimensions de toutes les combinaisons de paramètres pris 2 à 2, afin de bien comprendre toutes les « relations » possibles existantes entre les paramètres et les « ressemblances » éventuelles entre dates/stations d'un point de vue de ces paramètres. En ce sens, les

analyses numériques vont permettre d'avoir une vision globale des relations entre toutes les variables d'intérêt et de déterminer comment les stations ou dates se ressemblent en termes d'évolution de ces paramètres.

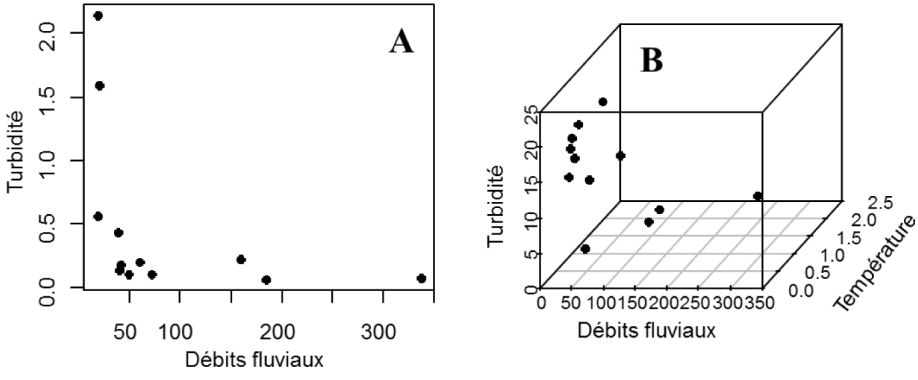

Figure I.1. *Représentation graphique de données issues de l'estuaire de la Charente (échantillonnage à point fixe pour une salinité de 5). A) Représentation en 2 dimensions de la turbidité en fonction des débits fluviaux, B) représentation en 3 dimensions de la turbidité en fonction des débits fluviaux et de la température.*

L'objectif du chercheur environnementaliste sera donc de *résumer l'information* portée par son jeu de données :

– en réduisant le nombre de dimensions (les variables corrélées entre elles peuvent apporter une redondance d'information) ;

– en mettant en évidence la ressemblance entre paramètres (par exemple, relations monotones linéaires positives ou négatives, relations non linéaires) ;

– en dégageant des grandes tendances/structures au sein des observations (par exemple, existe-t-il une structure particulière – gradient, groupes – des dates/ stations ?) ;

– de rechercher les causes des structures obtenues (c'est-à-dire quel(s) paramètre(s) explique(nt) la formation de ces groupes ou gradients ?).

I.1.2. *Pourquoi cet ouvrage ?*

L'objectif de ce livre est de présenter les analyses numériques les plus utilisées permettant de répondre à ces objectifs. Il ne sera bien évidemment pas exhaustif de par

le grand nombre d'analyses existantes à l'heure actuelle et le développement permanent dans ce domaine.

Deux écoles différentes existent pour l'utilisation de ces analyses :

– l'école francophone (Legendre et Legendre, 1998) adepte des approches paramétriques et incrémentant les analyses sous le logiciel R ;

– l'école anglophone (Clarke, 1993) préférant les approches non paramétriques et ayant développé le logiciel PRIMER (voir la section I.2.2 pour les notions d'approches paramétriques et non paramétriques).

Les chercheurs et enseignants sont la plupart du temps rattachés à l'une des deux écoles. Le constat est fait à travers la bibliographie que ceux rattachés à l'école anglophone optent la plupart du temps pour des analyses non paramétriques, même si leurs données leur permettraient d'accéder aux approches paramétriques plus puissantes dans certains cas et que ceux rattachés à l'école francophone utilisent des analyses paramétriques sans se poser la question sur les possibles biais engendrés pour un petit jeu de données.

Le principal souci de ce livre est donc de faire découvrir l'ensemble de ces analyses, les faire utiliser, permettre de les interpréter correctement et donner des pistes pour trouver les analyses les plus pertinentes pour les jeux de données utilisés en fonction des objectifs fixés en amont.

L'approche graphique deux à deux paramètres est une étape préalable obligatoire pour maîtriser le jeu de données. Elle est fortement conseillée avant et en parallèle des analyses multivariées afin de :

– vérifier la qualité des données, c'est-à-dire déceler des données aberrantes et les corriger si nécessaire (par exemple, faire la distinction entre des erreurs de mesures et des données extrêmes) ;

– vérifier que les grandes tendances observées sont plausibles, c'est-à-dire que l'approche numérique est correcte et que l'interprétation en est bonne ;

– vérifier que toutes les informations pertinentes sont mises en évidence à partir du jeu de données utilisé pour répondre aux objectifs scientifiques posés.

L'utilisation de ce genre d'analyses apporte une vision globale des résultats : cela permet notamment de résumer en 1 ou 2 figure(s) ce qui aurait nécessité une trentaine de figures. Tous les rapports de stage et toutes les publications scientifiques sont en effet limités en nombre de figures… Par ailleurs, ces analyses peuvent apporter des

informations qui n'auraient pas pu être visualisées par la simple observation de graphiques à 2 dimensions.

I.1.3. *Pourquoi le logiciel R ?*

Le logiciel R est à la fois un langage informatique de programmation et un environnement de travail permettant le traitement statistique des données. Il sert à manipuler des données, à tracer des graphiques et à faire des analyses statistiques sur ces données.

Parmi ses avantages, R est : 1) un logiciel « libre » téléchargeable et installable sur tous les ordinateurs ; 2) un logiciel « multi plates-formes » qui fonctionne sous Windows, Linux, Mac OS, etc. En conséquence, de nombreux scientifiques l'utilisent à l'échelle internationale et n'hésitent pas à partager leurs connaissances statistiques en développant de nouvelles fonctions/nouveaux packages et en communiquant à travers des forums. Cela en fait donc un logiciel en rapide et constante évolution. Ainsi, de nombreuses analyses statistiques y sont disponibles, des analyses aussi bien simples que complexes (c'est-à-dire statistiques descriptives et inférentielles, tests paramétriques ou non paramétriques, modèles linéaires ou non linéaires, écologie numérique – multivariées, traitement du signal, analyses spatiales, etc.). Très peu de logiciels ne sont en mesure de fournir non seulement un tel choix d'analyses mais encore de manière « libre ». Les approches paramétriques et non paramétriques présentées dans cet ouvrage sont à l'heure actuelle toutes incrémentées sous R. Enfin, R dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

L'outil est très efficace car, lorsque le langage R est dominé, il est alors possible de créer ses propres outils (ou « scripts »), permettant la réalisation d'analyses numériques plus sophistiquées sur les données (c'est-à-dire une succession de lignes de commandes) et de les reproduire très rapidement sur d'autres jeux de données.

I.2. Les grands types d'analyses multivariées

Les analyses multivariées font partie intégrante de l'écologie numérique, c'est-à-dire le domaine de l'écologie quantitative qui traite de l'analyse numérique des complexes de données (Legendre et Legendre, 1998). Elles permettent le traitement en bloc de tableaux de données où chaque élément d'échantillonnage (par exemple, stations/dates) est défini par plusieurs variables (descripteurs). Elles combinent à la fois des méthodes statistiques et non statistiques. On distinguera 1) *deux grandes familles d'analyses multivariées* qui permettront d'aller du descriptif à l'explicatif d'un jeu de données (c'est-à-dire analyses exploratoires, explicatives) et 2) *deux types*

d'approches selon la puissance et la robustesse de celles-ci (c'est-à-dire des approches paramétriques et non paramétriques).

I.2.1. *Les deux grandes familles d'analyses multivariées*

I.2.1.1. *Les analyses de type « exploratoire »*

Elles ont pour objectif commun de *décrire la structure des données.* Par exemple, de déterminer comment les stations se différencient entre elles sur la base de descripteurs biologiques (exemple : espèces phytoplanctoniques). Il en existe *3 grands types*.

I.2.1.1.1. Les analyses de groupement (*cluster analysis*)

Elles ont pour objectif de *classer les objets* (par exemple, date/stations) *ou les descripteurs* (par exemple, variables physico-chimiques, biologiques, etc.) *en groupes d'unités qui se ressemblent*. Elles visent donc à repérer les discontinuités dans un jeu de données. La formation de ces groupes tient compte de toute l'information contenue dans le jeu de données. Toutefois, cette seule analyse ne permettra pas de donner des informations sur les paramètres qui génèrent ces groupes. Si trois groupes de stations sont mis en évidence par une classification, cela veut dire que les stations contenues dans chacun de ces groupes se ressemblent sur la base de l'ensemble des paramètres considérés mais sans information sur l'importance de chaque paramètre quant à la distinction de ces groupes.

EXEMPLE. Classification ascendante hiérarchique (CAH).

I.2.1.1.2. Les projections multidimensionnelles ou analyse de proximité (*multidimensional scaling*)

Elles ont pour objectif de *représenter dans un espace à dimension réduite* (par exemple, un plan) *la ressemblance/dissemblance entre objets ou descripteurs*. Elles considèrent toute l'information contenue dans le jeu de données. Elles partent du principe que les données environnementales ou biologiques se structurent plutôt sous la forme de gradient et qu'il est difficile de les restreindre à de simples groupes. Cette capacité à mettre en évidence des groupes ou des gradients les rend plus « puissantes » que les analyses de groupement. Toutefois, la possibilité d'utiliser cette analyse dépendra de sa capacité à représenter toute l'information contenue dans le jeu de données (par exemple, 68 espèces) en quelques dimensions (2 ou 3). Enfin, elles présentent la même faiblesse que l'analyse de groupement : elles ne permettent pas à elle seule de donner des éléments sur les paramètres qui génèrent les gradients/groupes.

EXEMPLE. Projection multidimensionnelle non métrique (NMDS).

I.2.1.1.3. Les analyses factorielles (ou *unconstrained ordination*)

Elles ont pour objectif *de repérer et de hiérarchiser les grandes tendances dans un jeu de données pour les objets et les descripteurs*. Ces grandes tendances sont représentées sous la forme d'un nombre d'axes restreints qui résument l'information, mais toute l'information contenue dans le jeu de données initial ne sera pas traduite par les quelques axes gardés. En revanche, les gradients ou groupes de stations sont mis en relation avec les paramètres qui les expliquent. Si les stations se répartissent le long des deux premiers axes et que ces axes sont bien corrélés aux concentrations en nitrates et en phosphates pour le premier et la température pour le second, cela voudra dire que la structure des stations est expliquée de manière dominante par la disponibilité en sels nutritifs et secondairement et indépendamment par la température. Cette capacité à hiérarchiser l'information contenue dans le jeu de donnée *rend les analyses factorielles plus « puissantes » que les analyses de groupement ou de proximité*. Toutefois, ces analyses sont soumises au respect d'un certain nombre de conditions d'application qui les rendent *beaucoup moins robustes* notamment pour les petits jeux de données (c'est-à-dire peu d'objets par rapport aux nombres de descripteurs). Pour les termes de « puissance » et « robustesse » se référer à la section I.2.2.

EXEMPLE. Analyse en composantes principales (ACP).

I.2.1.2. *Les analyses de type « explicatives »*

Elles ont pour objectif commun de *confronter la structure des données obtenues par les méthodes exploratoires déclinées ci-dessus à d'autres jeux de données* en vue d'expliquer cette structure. Par exemple, la mise en évidence d'un gradient de stations sur la base des communautés phytoplanctoniques entraîne la nécessité d'expliquer ce gradient à l'aide de variables environnementales (par exemple, des paramètres physico-chimiques, hydrodynamisme, occupation des sols, etc.). Il existe 2 moyens pour arriver à ses fins.

I.2.1.2.1. Les analyses *a posteriori* (ou indirectes)

Elles ont pour objectif *de mettre en relation la structure obtenue (telle quelle) avec les paramètres potentiellement explicatifs*, soit par projection passive de variables supplémentaires, soit par corrélation de ces variables aux axes d'une analyse exploratoire ou à des groupes définis *a priori* (par exemple, des facteurs qualitatifs d'intérêt ou définis à partir d'autres analyses exploratoires). C'est donc une approche passive.

EXEMPLE. Variables supplémentaires.

I.2.1.2.2. Les analyses *a priori* (ou directes)

Elles ont pour principe de *contraindre la structure en amont de l'analyse* par les paramètres potentiellement explicatifs. C'est une approche active. Pour être plus efficace, ce type d'analyse nécessite de restreindre le nombre de variables potentiellement explicatives à celles apportant une information en tant que forçage explicatif réel.

EXEMPLE. Analyse canonique des correspondances (ACC).

I.2.2. *Notions de puissance et robustesse : les tests d'hypothèses pour analyses multivariées*

I.2.2.1. *Les tests d'hypothèses pour analyses multivariées*

Certains tests statistiques sont utilisés en analyse numérique pour vérifier de manière significative :

– si les structures mises en évidence (par exemple, les groupes) sont différentes ;

– si certains descripteurs peuvent expliquer ces structures ;

– si ces structures sont liées à certaines variables environnementales ;

– la correspondance entre des structures mises en évidences par deux analyses différentes.

C'est la première fois que le mot « significatif » est prononcé. En effet, la plupart des analyses numériques ne s'interprètent que graphiquement (avec quelques règles strictes bien sûr). Dans ces cas, décrire les résultats obtenus en parlant de corrélation ou de tendance « significative » est fortement déconseillé.

Pour faire simple, les méthodes statistiques sont celles qui, basées sur des tests, permettent de dire si les résultats sont « significatifs ». Dans les statistiques inférentielles les plus classiques, ces tests permettent de comparer les estimateurs (c'est-à-dire moyennes, variances, pourcentages, etc.) d'une variable (par exemple, la température) entre 2 ou plusieurs populations à partir d'échantillons aléatoires de celles-ci. Le principe de tests paramétriques est de partir de l'hypothèse que la variable considérée suit une loi de distribution théorique connue (par exemple, la loi normale) et donc, que l'hypothèse nulle associée au test (c'est-à-dire que les fluctuations d'échantillonnages expliquent les différences observées) suit cette distribution. Le problème concernant les tableaux de données, c'est qu'ils ne se restreignent pas qu'à

une seule variable mais présentent plusieurs descripteurs par élément. Les tests statistiques développés doivent donc intégrer ce côté multidimensionnel et le principe adopté sera donc différent : la distribution théorique sur laquelle le test sera basé sera calculée à partir de permutations aléatoires du tableau de données initial générant de l'aléatoire et permettant ainsi de construire ainsi une distribution correspondant à l'hypothèse nulle (c'est-à-dire tests par permutation).

EXEMPLE. Test SIMPROF.

I.2.2.2. *Approches dites paramétriques et non paramétriques en analyses multivariées*

Il est complètement *utopique d'obtenir l'ensemble des données d'une population* (par exemple, toutes les huîtres du Bassin d'Arcachon) de par des contraintes de temps, de personnels, de prélèvements nécessitant la mort des individus échantillonnés, etc. La recherche passe donc par des *échantillons représentatifs*, c'est-à-dire reflétant la complexité et la composition de la population. Ces échantillons représentatifs ne sont accessibles que par un échantillonnage aléatoire qui donne autant de chance à tous les individus de la population de se retrouver dans l'échantillon. Toutefois, les fluctuations d'échantillonnage font que deux échantillons d'une même population peuvent donner des résultats en moyenne différents pour un paramètre donné et que deux échantillons issus de populations différentes peuvent donner des valeurs en moyenne identiques pour ce paramètre. Seules les statistiques peuvent permettre de généraliser les valeurs obtenues à partir d'échantillon(s) à la (aux) population(s) entière(s). Dans les tests statistiques classiques, deux hypothèses de travail sont posées : l'hypothèse nulle et l'hypothèse alternative. Dans le cadre de la comparaison de deux populations (par exemple, des moyennes de températures), l'hypothèse nulle (H0) suppose que les fluctuations d'échantillonnages expliquent les différences entre les échantillons issus de populations différentes (c'est-à-dire, pas de différence visible) et l'hypothèse alternative (H1) admet que les fluctuations d'échantillonnage ne peuvent pas tout expliquer (c'est-à-dire différences conclues). La plupart du temps, l'objectif recherché par le chercheur environnementaliste est de rejeter H0, c'est-à-dire de mettre en évidence un effet significatif. Les tests statistiques permettent de conclure sur l'une ou l'autre hypothèse.

La *puissance* d'un test est son aptitude à rejeter H0 lorsque H0 est fausse. La *robustesse* est, quant à elle, sa sensibilité à des écarts par rapport aux hypothèses faites, soit son aptitude à donner des résultats fiables lorsque les conditions d'application ne sont pas respectées. Ainsi, pour chaque type d'analyse (par exemple, la comparaison de moyenne de 2 échantillons indépendants), il existera au moins 2 tests :

– *une alternative paramétrique* : plus puissante mais moins robuste. Elle aura plus de chance de rejeter H0 si H0 est fausse, c'est-à-dire de conclure qu'il y a une différence entre 2 échantillons s'il y en a réellement une, mais sera très sensible aux conditions d'application des tests. Si ces dernières ne sont pas respectées, les résultats seront erronés ;

– *une alternative non paramétrique* : moins puissante mais plus robuste. Elle aura moins de chance de rejeter H0 si H0 est fausse mais n'aura que très peu de contraintes du point de vue des conditions d'application des tests.

Pour les 2 grandes familles d'analyses multivariées, certaines analyses peuvent être qualifiées de paramétriques et d'autres de non paramétriques (voir la figure I.2). Les analyses paramétriques seront plus puissantes en termes de types de résultats apportés mais contraintes par un certain nombre de conditions d'application. C'est le cas par exemple des analyses factorielles qui permettent de hiérarchiser l'information du jeu de données, mais qui sont contraintes par une condition de multinormalité des données (c'est-à-dire que les données doivent suivre une loi normale pour chaque variable). D'autres analyses moins puissantes n'auront aucune condition d'application préalable. Elles sont généralement basées sur les rangs (de coefficient d'association par exemple) et non sur les valeurs brutes. C'est le cas par exemple de la projection multidimensionnelle non métrique (NMDS) : elle permet juste de représenter la ressemblance ou dissemblance entre stations sur un plan à 2 dimensions sans hiérarchisation possible de l'information. Cependant, elle ne présente aucune condition d'application si ce n'est que l'indicateur de qualité de représentation (c'est-à-dire le stress) soit bon pour que l'interprétation soit correcte.

I.2.2.3. *Conséquences : des analyses différentes pour les « petits » et les « grands » jeux de données*

Ainsi, les approches utilisées (paramétriques *versus* non paramétriques) dépendront de la taille de la base de données d'intérêt (voir la figure I.2). De manière générale, les « grands » jeux de données pour lesquels le nombre d'éléments (par exemple, stations/dates) sera important et notamment supérieur au nombre de descripteurs (par exemple, variables physico-chimiques), se prêteront aux analyses de type paramétrique (par exemple, analyses factorielles, analyses canoniques, etc.). Par opposition, les approches non paramétriques, pour lesquelles la robustesse est élevée, seront à préférer pour les « petits » jeux de données (c'est-à-dire un nombre d'éléments faibles et inférieurs aux nombres de descripteurs). La puissance de ces dernières approches sera certes plus faible (par exemple, moins d'information sur la hiérarchisation de l'importance des descripteurs dans la structuration des éléments) mais les analyses numériques utilisées ne seront pas biaisées par la faiblesse du jeu de données.

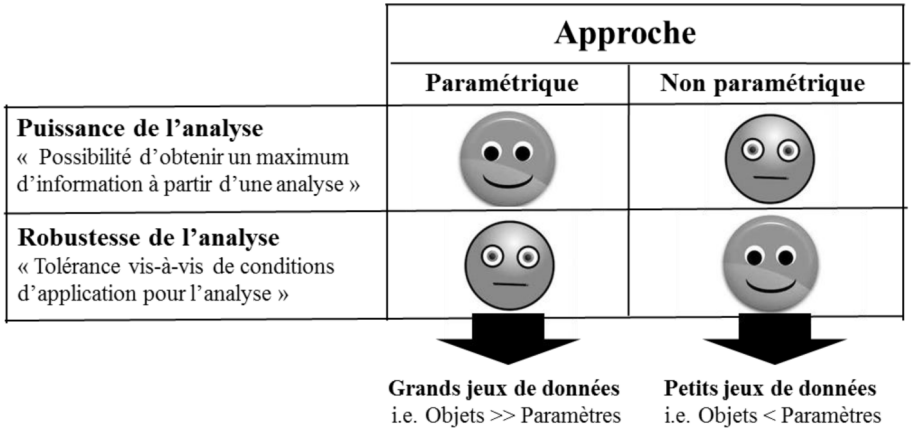

Figure I.2. *Approches paramétrique et non paramétrique appliquées aux analyses multivariées. Conditions en termes de puissance et de robustesse et conséquences en termes d'application sur des petits et grands jeux de données (nombre d'objets versus nombre de paramètres).*

Quoi qu'il en soit, qui peut le plus peut le moins : les approches non paramétriques pourront être appliquées à des « grands » jeux de données. Toutefois, si une approche paramétrique est possible pour l'objectif visé, celle-ci est conseillée afin de pouvoir tirer plus d'informations sur la base de données traitée. En effet, si une analyse factorielle est possible, elle apportera des informations sur la hiérarchisation de l'importance des descripteurs sur la structure des objets, alors que l'utilisation d'une analyse multidimensionnelle non métrique ne fournira que la structure en elle-même des objets sans information sur le rôle des descripteurs impliqués. Toutefois, certaines analyses permettront dans certains cas, d'avoir une approche intermédiaire avec une puissance plus forte qu'une approche non paramétrique mais moins contrainte en termes de conditions d'application (par exemple, analyses de variance par permutation). La possibilité de les utiliser pourra ainsi être testée pour de petits jeux de données avant de se restreindre à une approche non paramétrique.

I.3. De l'objectif de recherche aux choix des analyses

Dans le meilleur des cas, la stratégie d'échantillonnage ou le plan expérimental a été mis en place pour répondre à un ou des objectifs spécifiques. Il est d'ailleurs important d'avoir déjà en tête le type d'analyse numérique qui sera appliqué lors du traitement des données pendant l'élaboration de l'échantillonnage ou du plan expérimental (Scherrer, 1984).

Les objectifs ne doivent pas être perdus de vue pendant tout le traitement des données au risque de tomber dans le travers de faire des analyses pour faire des analyses et donc finalement, de ne pas répondre aux questions posées au départ.

Plusieurs analyses pourront permettre de répondre à une même question. L'important est de choisir la ou les analyses la (ou les) plus pertinente(s), apportant le maximum d'information dans la mesure où les conditions d'application sont respectées. Si plusieurs analyses sont utilisées, elles doivent être complémentaires, l'utilisation d'analyses redondantes est totalement inutile.

Il faut analyser, d'un regard critique, les analyses proposées dans les publications. Non pas que ces analyses soient critiquables, mais ce n'est pas parce qu'un chercheur utilise une analyse en particulier pour répondre au même objectif, qu'il est conseillé de pratiquer la même analyse. Sa stratégie d'échantillonnage peut être différente (par exemple, nombre de stations, types et nombres de paramètres, etc.) et par conséquent, les analyses à appliquer peuvent varier pour cause de conditions d'application par exemple.

De même, lorsqu'une base de données est récupérée d'une autre étude, il faut s'assurer que la stratégie d'échantillonnage ou le plan expérimental adopté en amont est bien adapté aux objectifs fixés.

I.3.1. *Est-ce possible de manipuler les résultats des analyses avec de tels outils ?*

Les analyses numériques restent avant tout des outils qu'il convient d'utiliser avec rigueur en toute connaissance des données et notamment, de la manière dont elles ont été acquises. Il ne viendrait pas à l'idée d'un chercheur d'utiliser une sonde à oxygène sans calibration préalable ou bien de déterminer des organismes vivants sans utiliser les critères pointus d'identification recensés dans des clés de détermination spécialisées. L'adage erroné mais répandu « *on peut faire dire n'importe quoi ou ce que l'on veut aux données avec les statistiques* » est un signe clair de méconnaissance de ces analyses.

I.3.2. *« Dire n'importe quoi ». Quelles erreurs peuvent conduire à une interprétation erronée ?*

La plupart des analyses sont en effet basées sur des hypothèses fortes qui conditionnent les propriétés des outils mathématiques développés dans celles-ci. Par exemple, les analyses factorielles partent du principe que les données suivent une loi normale pour tous les descripteurs considérés et l'interprétation qui va être faite n'est

rigoureusement valable que si cette condition est respectée. Le *respect des conditions d'application* est donc un gage d'interprétation fiable pour toutes ces analyses.

Il est également important d'*interpréter correctement les analyses en ayant conscience de leurs limites* : chacune d'entre elles présente ses propres règles d'interprétation qui découle notamment des propriétés des outils mathématiques utilisés et des hypothèses préalablement posées. Il faut donc avoir conscience de ce qu'il est possible de conclure ou non avec l'analyse choisie. Par exemple, les descripteurs mal représentés sur les axes d'une analyse factorielle ne pourront en aucun cas être interprétés sous peine d'en tirer des conclusions erronées.

Une autre limite particulièrement importante de ces analyses provient du fait que *ceux ne sont que des outils mathématiques* ! Les résultats qu'elles donnent doivent être abordés avec un *regard critique par le chercheur qui reste le « spécialiste » de son domaine*. Les connaissances empiriques du terrain ou de la bibliographie représentent une aide précieuse à toute interprétation. La mise en évidence d'une corrélation entre 2 ou plusieurs variables n'est pas un gage de relation de cause à effet. En effet, les relations entre variables sont particulièrement complexes.

– *La corrélation entre 2 variables peut être le fait d'une coïncidence*. Imaginons la mise en évidence d'une corrélation forte entre les pourcentages d'occupation du sol sur le bassin versant en surface agricole, en surface boisée et la quantité de nitrates dans les eaux du fleuve situés en aval. S'il est évident que les fortes concentrations en nitrates sont plus vraisemblablement expliquées par une surface agricole importante sur le bassin versant plutôt que par de fortes étendues boisées, seules des connaissances empiriques du terrain peuvent conduire à la conclusion que la relation entre surface agricole et boisée n'est que pure coïncidence sur la zone géographique considérée.

– *Deux variables peuvent être expliquées par une seule et même variable de manière indépendante*. Il est en effet possible qu'une corrélation observée entre les deux premières ne peut être due qu'à l'effet exercé par la tierce variable. On peut citer par exemple, deux espèces, l'une benthique et l'autre planctonique en milieu aquatique, dont les abondances seraient fortement expliquées par la température : les fortes températures favoriseraient l'espèce benthique et les faibles températures, l'espèce planctonique. Cette corrélation aux températures peut induire une anticorrélation forte entre les deux espèces sans pour autant qu'existe une réelle relation négative entre les espèces (c'est-à-dire prédation, compétition, etc.). Seules des connaissances approfondies sur ces deux espèces peuvent permettre au chercheur d'identifier si cette relation négative est réelle ou causée par le lien indépendant de chacune d'entre elles à la température.

– *Une relation entre deux variables peut être positive ou négative selon l'échelle considérée*. La température peut être un paramètre affectant positivement la physiologie d'une espèce à petite échelle (échelle annuelle) et négativement à long terme (échelle décennale). Par exemple, la température est souvent un facteur positif favorisant les fonctions physiologiques d'espèces poïkilothermes et contrôle ainsi le cycle saisonnier de celles-ci (en fonction, bien sûr, des *preferenda* thermiques de chacune). Toutefois, l'augmentation à long terme des températures peut amplifier les écarts thermiques entre maxima estivaux et minima hivernaux et être ainsi préjudiciable à ces espèces.

– *Plusieurs variables peuvent expliquer les fluctuations d'une autre variable (par exemple, une espèce) et ces variables explicatives peuvent également être reliées entre elles*. L'augmentation de la température peut causer une stratification de la colonne d'eau et donc un épuisement plus rapide des sels nutritifs défavorables aux diatomées (micro-algues planctoniques). Même si certaines espèces de diatomées peuvent être favorisées par des températures croissantes, le seuil de température déclenchant ou non une stratification favorisera ou non le développement de ce groupe.

I.3.3. *« Dire ce que l'on veut ». Cela sous-entend-il que parmi plusieurs analyses donnant des résultats différents, je choisis celle qui m'arrange ?*

Bien sûr que non ! Dans une approche rigoureuse, plusieurs analyses seront possibles mais les résultats seront cohérents, redondants ou complémentaires. Le choix des analyses gardées se portera sur la quantité et la qualité des informations apportées en fonction de l'objectif posé, en évitant de présenter des analyses totalement redondantes.

Une approche rigoureuse nécessite de *faire des choix* dans la préparation de la base de données et tout au long de l'analyse numérique :

– le tableau de données peut, par exemple, présenter des données manquantes dont le traitement dépend du type de données : une absence de données représentant une absence d'une espèce peut être remplacée par 0, alors qu'une absence représentant la panne d'un appareil empêchant la mesure doit être identifiée comme une donnée manquante (notée NA) ;

– les données peuvent nécessiter une transformation dont le choix dépend du type de données et des besoins pour l'analyse (par exemple, homogénéiser les unités de mesures, « normaliser » les données, réduire les gammes de variations, etc.) ;

– le choix du coefficient d'association qui permettra de mesurer les degrés de ressemblances entre objets ou descripteurs dépendra, quant à lui, des objectifs et de la

manière dont les données ont été acquises. La co-absence d'une espèce en 2 stations par exemple pourra être dans un cas considérée comme un critère de ressemblance entre station si cette co-absence traduit un phénomène particulier que je cherche à étudier (par exemple, pollution, etc.) ou bien, être ignorée si je ne veux pas donner trop de poids à celle-ci (par exemple, dans le cas où la méthode de prélèvement est biaisée par les espèces rares, etc.).

L'ensemble de ces choix va conditionner les relations ou ressemblances observées entre objets (c'est-à-dire station/dates) et descripteurs. Il est évident que pour répondre à un objectif, ces choix seront identiques quel que soit l'analyse qui sera effectuée. Par conséquent, les résultats seront donc cohérents entre toutes les approches utilisées même si certaines analyses pourront apporter quelques variantes selon leurs spécificités et propriétés.

I.4. Base de données traitée en exemple dans l'ouvrage

Les données présentées ici seront utilisées tout au long de l'ouvrage pour la mise en pratique des analyses utilisées. Pour chaque objectif énoncé ci-dessous, sont notés les chapitres dans lesquels les analyses appropriées seront appliquées.

Quatre marais desséchés de Charente-Maritime ont été échantillonnés. Ces types de marais ont été façonnés par l'Homme au cours des siècles pour le développement d'activités agricoles (c'est-à-dire agriculture, élevage, ostréiculture, pisciculture, saliculture, etc.) et ont également été soumis à une urbanisation progressive (implantation de stations d'épuration, d'habitations). Il s'agit de réseaux hydrographiques artificiels de chenaux et de fossés de plusieurs dizaines à centaines de kilomètres de long, pour lesquels l'action humaine vise à contrôler des portes à la mer pour éviter la pénétration d'eaux salées dans le réseau à chaque marée montante et éviter l'inondation des terrains adjacents durant les crues hivernales. Le réseau hydrographique se compose de trois types de chenaux : (i) les chenaux principaux larges et de profondeur supérieure à 1 mètre dans lesquels l'eau provenant des autres chenaux est drainée avant d'atteindre les eaux côtières, (ii) les fossés ou chenaux tertiaires, les plus étroits avec des profondeurs de moins de 50 cm, alimentant directement les terrains adjacents sur lesquels se développent les activités anthropiques, (iii) les chenaux intermédiaires ou secondaires assurant le drainage des eaux depuis les fossés jusqu'aux chenaux principaux.

Les quatre marais (A, B, C, D) ont été échantillonnés pendant la période estivale (juillet) : deux marais desséchés réalimentés par le fleuve adjacent (la Charente) pour assurer le réapprovisionnement en eau pendant les sècheresses estivales durant lesquelles les besoins des cultures en eaux sont particulièrement importantes (R) et deux marais

desséchés non réalimentés pour lesquels le niveau d'eau peut rapidement baisser en été jusqu'au dessèchement complet de certains fossés. Pour chacun de ces types de marais, un marais a été échantillonné dans l'intérieur des terres (I) et l'autre sur le littoral proche des portes à la mer (E). Pour chacun des 4 marais, 4 à 6 stations ont été échantillonnées (notées a à e) pour des types de chenaux différents et des occupations de sols sur des zones d'influence proche différentes. Les données sont disponibles sur le profil Researchgate de l'auteur (https://www.researchgate.net/profile/V_David). Chaque type de base de données présenté est fourni par un fichier .RData.

Pour chacune des 19 stations, différents paramètres ont été échantillonnés et notés (voir le tableau I.1) :

– *des paramètres environnementaux (base « fext »)* : station (variable « Station » de A à D), type de marais (variable « Type », D ou R), « Position » (I pour interne ou E pour externe), occupation du sol sur la zone d'influence proche (variable « Occupation », avec Prairie, Culture ou Urbain), types de chenaux (variable « Chenal » avec prim pour primaire, sec pour secondaire ou ter pour tertiaire), présence de macrophytes (variable « MP » avec oui si les macrophytes sont présentes), surface du bassin versant (variable « BV » exprimée en km^2). Les codes des stations apparaissent comme nom de lignes avec le type de marais (de A à D) et les stations échantillonnés dans ces marais (de a à f) :

```
fext<-as.data.frame(get(load("fext.Rdata")))
```

– *des paramètres physico-chimiques (base « PC »)* : la profondeur (« Prof »), la luminosité à la surface de l'eau (« Lumin »), la profondeur optique (« Prof_opt »), la température de l'eau (« Temp »), les concentrations en nitrites (« NO2 »), nitrates (« NO3 ») et phosphates (« PO4 »), le rapport (NO3+NO2)/PO4 (« N.P ») et la turbidité (« Turb ») :

```
PC<-as.data.frame(get(load("PC.Rdata")))
```

– *des paramètres biologiques globaux (base « bio»)* : la biomasse chlorophyllienne comme indicateur de la biomasse phytoplanctonique (« Chloa »), la production primaire sur la colonne d'eau (« P_phyto »), la productivité phytoplanctonique – rapport production/biomasse (P.B) :

```
bio<-as.data.frame(get(load("bio.Rdata")))
```

– *des abondances phytoplanctoniques (base « phyto »)* : les abondances de 68 genres ont été répertoriées mais seulement celles de 7 genres sont données dans le tableau I.1. L'ensemble des 68 genres et les sigles correspondant est noté dans le tableau I.1 :

```
phyto<-as.data.frame(get(load("phyto.Rdata")))
```

Stations	ENVIRONNEMENT							PHYSICO-CHIMIE									BIOLOGIE			ABONDANCES PHYTOPLANCTONIQUES							
	Marais	Type	Position	Occupation	Chenal	MP	BV	Prof	Lumin	Prof_opt	Temp	NO2	NO3	PO4	N.P	Turb	Chloa	P_phyto	P.B	Cycl	Melo	Licm	Frag	Nitz	Cosm	Eugl	...
A_a	A	D	I	Prairies	sec	1	4716	0,50	857	0,35	23	0,01	0,0	3,0	0,0	12	14	24	0,03	305	56	0	0	28	0	0	...
A_b	A	D	I	Cultures	prim	0	4716	0,70	1886	0,05	30	0,02	0,0	0,1	0,2	23	22	2641	1,91	0	411	0	0	17625	137	1646	...
A_c	A	D	I	Cultures	ter	1	4716	0,30	1941	0,05	26	0,03	0,0	2,0	0,0	38	87	13704	2,45	805	0	0	0	889	611	666	...
A_d	A	D	I	Cultures	prim	0	4716	0,80	1704	0,15	30	0,02	0,0	1,0	0,0	35	44	3819	1,36	154	0	0	0	463	695	116	...
A_e	A	D	I	Cultures	prim	0	4716	0,25	346	0,02	26	0,09	0,0	1,0	0,1	58	95	7395	1,21	0	0	0	0	4860	194	889	...
A_f	A	D	I	Cultures	ter	0	4716	0,08	818	0,00	26	0,07	0,0	2,0	0,0	60	189	18768	1,55	41098	0	0	0	9025	1944	2778	...
B_a	B	D	E	Cultures	prim	0	6971	0,48	1761	0,15	25	0,05	0,0	1,0	0,0	59	128	12740	1,55	15984	0	0	0	224	671	559	...
B_b	B	D	E	Prairies	sec	1	6971	0,30	1834	0,27	25	0,02	0,0	0,1	0,2	49	66	8524	2,02	139	0	0	0	333	0	1166	...
B_c	B	D	E	Urbain	prim	1	6971	0,75	133	0,30	24	0,05	1,0	3,0	0,4	60	330	698	0,03	11108	0	0	0	12774	6109	555	...
B_d	B	D	E	Prairies	sec	0	6971	0,33	1535	0,08	24	0,01	0,0	2,0	0,0	14	12	1947	2,43	139	139	69	0	208	0	417	...
C_a	C	R	E	Prairies	ter	1	1325	0,30	1306	0,15	22	0,00	0,0	0,1	0,0	10	24	4401	2,87	4	30	0	0	3	0	0	...
C_b	C	R	E	Urbain	sec	0	1325	1,00	1899	0,07	22	0,01	0,0	0,1	0,1	40	46	4287	1,46	28	0	0	250	14	167	458	...
C_c	C	R	E	Urbain	ter	0	1325	0,77	1427	0,47	22	0,00	0,0	2,0	0,0	13	21	1638	1,23	4041	0	0	0	458	14	625	...
C_d	C	R	E	Urbain	prim	0	1325	1,50	745	0,45	23	0,01	0,0	3,0	0,0	53	197	1230	0,1	0	0	0	0	19300	0	42488	...
D_a	D	R	I	Urbain	prim	0	982	1,33	1782	0,04	26	0,09	2,7	0,1	27,9	60	36	860	0,37	2	91	0	0	132	1	5	...
D_b	D	R	I	Prairies	ter	0	982	0,53	1854	0,10	26	0,07	2,5	0,1	25,7	60	346	27497	1,24	112	72	0	0	52	0	2462	...
D_c	D	R	I	Cultures	prim	1	982	0,93	2129	0,20	24	0,04	2,9	0,1	29,4	60	56	4501	1,26	695	0	0	0	144	0	1228	...
D_d	D	R	I	Prairies	ter	1	982	0,85	1310	0,33	23	0,04	2,9	1,0	2,9	55	61	3611	0,93	0	0	0	0	689150	0	7980	...
D_e	D	R	I	Prairies	sec	1	982	0,83	1888	0,35	24	0,03	2,9	0,1	29,3	44	19	1989	1,63	28	0	0	0	6	0	1028	...

Tableau I.1. *Paramètres environnementaux, physico-chimiques et biologiques des prélèvements réalisés dans les 19 stations des marais de Charente-Maritime. Seules les abondances de 7 genres microphytoplanctoniques sont données (sur les 68 au total)*

Genre	Code		Genre	Code
Cyclotella	Cycl		*Staurodesmus*	Staud
Melosira	Melo		*Actinastrum*	Acti
Licmophora	Licm		*Ankistrodesmus*	Anki
Fragillaria	Frag		*Crucigenia*	Cruc
Achnantinium	Achnm		*Haematococcus*	Haem
Meridion	Merid		*Monoraphidium*	Mono
Synedra	Syne		*Oocystis*	Oocy
Achnanthes	Achn		*Pandorina*	Pand
Rhopalodia	Rhop		*Pediastrum*	Pedi
Caloneis	Calo		*Scenedesmus*	Scen
Gomphonema	Gomp		*Desmodesmus*	Desm
Gyrosigma	Gyro		*Tetrastrum*	Tetrs
Haslea	Hasl		*Tetraedon*	Tetrd
Navicula	Navi		*Chroomonas*	Chroas
pleurosigma	pleu		*Cryptomonas*	Cryp
Amphora	Amph		*Chrysochromulina*	Chry
Cymbella	Cymb		*Chlamydomonas*	Chla
Cymatopleura	Cyma		*Tetraselmis*	Tetrl
Epithemia	Epit		Euglenes	Eugl
Bacillaria	Baci		*Lepocinclis*	Lepo
Cylindrotheca	Cyli		*Peranema*	Pera
Nitzschia	Nitz		*Phacus*	Phac
Pseudonitzschia	Pseu		*Strombomonas_*	Stro
Gymnodinium	Gymn		*Trachelomonas*	Trac
Peridinien	Peri		*Anabaena*	Anab
Gloecystis	Gloe		*Gomphospaeria*	Gompp
Cystodinium	Cyst		*Chroococcus*	Chro
Gonyostomum	Gony		*Merismopedia*	Meris
Synura	Synu		*Mycrocystis*	Mycr
Closterium	Clos		*Nostoc*	Nost
Dinobryon	Dino		*Plankto*	Plan
Netrium	Netr		*Oscillatoria*	Osci
Cosmarium	Cosm		*Spirulina*	Spir
Staurastrum	Staut		Divers cyano	cyan

Tableau I.2. *Genres phytoplanctoniques recensés au cours de l'étude et sigles utilisés pour la suite des analyses*

I.5. Structure de l'ouvrage

L'ouvrage est organisé en 5 chapitres.

Chapitre 1. Observer et préparer un jeu de données. Cette première partie vise à mieux connaître son jeu de données, c'est-à-dire établir ses limites et faire les choix nécessaires avant tout traitement en adéquation avec les objectifs de l'étude et la stratégie d'échantillonnage (voir les sections 1.1 et 1.2). Il donne également des moyens de simplifier les jeux de données afin de retirer les variables redondantes ou peu informatives (voir les sections 1.3 et 1.4).

Chapitre 2. Traitement préalable du jeu de données. Cette partie a pour objectif de montrer comment calculer des indices de diversité (voir la section 2.1), de transformer son jeu de données (voir la section 2.2) et choisir le coefficient d'association le plus adapté au jeu de données (voir la section 2.3).

Chapitre 3. Structure sous forme de groupes d'objets/variables. Cette partie permet d'exposer les méthodes de classification les plus utilisées, comment les appliquer en adéquation avec le jeu de données et les objectifs (voir la section 3.1) et comment caractériser les groupes obtenus (voir la section 3.2).

Chapitre 4. Structure sous forme de gradients d'objets/variables. Cette partie expose l'application des analyses factorielles pour une approche paramétrique (voir la section 4.1) et celle du positionnement multidimensionnel non métrique pour une approche non paramétrique (voir la section 4.2).

Chapitre 5. Comprendre une structure. Cette partie vise à présenter des méthodes permettant de comparer directement des structures sans hypothèses sur les variables explicatives et expliquées (voir la section 5.1), à trouver des facteurs quantitatifs et qualitatifs (voir la section 5.2) expliquant les structures obtenues dans les chapitres précédents.

La figure I.3 guide sur l'utilisation des analyses présentées dans ce livre en fonction des propriétés du jeu de données utilisé et des objectifs posés.

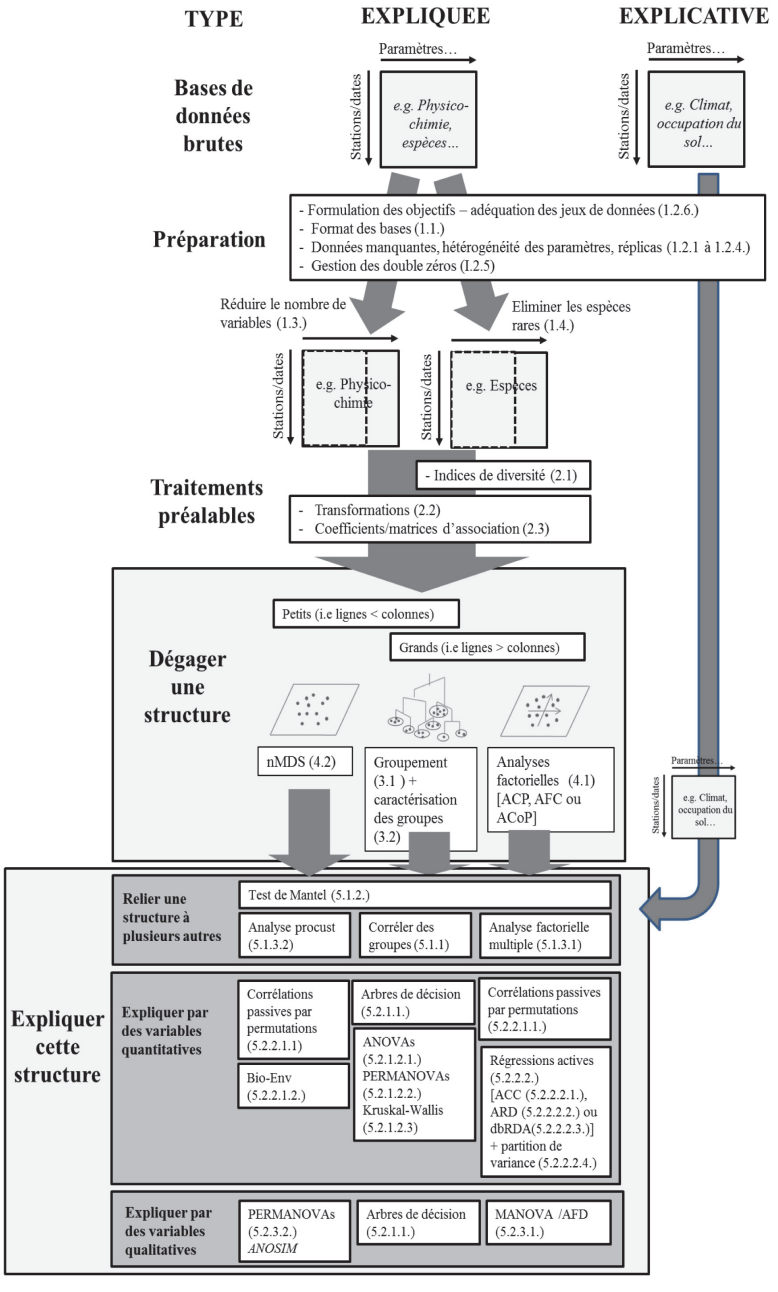

Figure I.3. *Schéma permettant le choix des analyses présentées dans ce livre en fonction du jeu de données et des objectifs visés*