



N° d'ordre :

École Doctorale Mathématiques, Sciences de l'Information et de l'Ingénieur

UdS - INSA - ENGEES

THÈSE

présentée pour obtenir le grade de

Docteur de l'Université de Strasbourg
Discipline : informatique
Spécialité : Fouille de données

par

Aurélie Bertaux

Treillis de Galois pour les contextes multi-valués flous **Application à l'étude des traits de vie en hydrobiologie**

soutenue publiquement le 1er octobre 2010

Membres du jury

Directrice de thèse : Florence Le Ber, Ingénieure en chef des Ponts, des Eaux
et des Forêts, HDR en informatique, LHyGeS et LORIA

Rapporteuse externe : Marianne Huchard, Professeure, LIRMM

Rapporteuse externe : Maguelonne Teisseire, Directrice de recherche Cemagref,
LIRMM et TETIS

Examinatrice : Michèle Trémolières, Professeure, LHyGeS

Examineur : Philippe Usseglio-Polatera, Professeur, LIEBE

Examinatrice : Agnès Braud, Maître de conférence, LSIIT

Laboratoire d'Hydrologie et de Géochimie de Strasbourg — UMR 7517
Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection — UMR 7005



Mis en page avec la classe thloria.

Remerciements

Avant toute chose, je tiens à remercier Marianne Huchard et Maguelonne Teisseire pour avoir accepté de rapporter ma thèse.

Le doctorat étant une formation l'encadrement est donc fondamental, c'est pour cela que je souhaite remercier Agnès Braud et Florence Le Ber pour leur encadrement et direction. Elles sont, l'une comme l'autre très occupées mais ont toujours eu le temps d'orienter, suivre et baliser mon travail (avec beaucoup de patience), et plus encore ont toujours relu nos articles et cette thèse autant de fois (incalculable) que nécessaire.

Cette thèse a également été suivie par un comité. C'est une initiative rare en informatique, mais qui est pourtant très enrichissante et rassurante car on peut se référer aux experts qui suivent la thèse. C'est une aide inestimable. Mon comité de thèse comporte des informaticiens, outre Florence Le Ber et Agnès Braud, il y a Amedeo Napoli, figure incontournable des treillis de Galois. C'est lui qui m'a mis le pied à l'étrier de l'ACF par les cours du CNAM, m'a souvent accueillie au LORIA pour des réunions de travail très riches et conviviales et m'a conseillée pour mon poste d'ATER.

Du côté biologie, mon comité accueille Philippe Usseglio-Polatera qui est un éminent spécialiste des invertébrés, d'une réactivité impressionnante aux mails que j'ai pu lui écrire. Il nous a également accueillies à Metz pour discuter de comment traiter le jeu de données des invertébrés (qui vient également de lui) et nous a fourni des pistes pour des années de travail. Je le remercie également de faire partie du jury. Corinne Grac travaille également sur les invertébrés et s'est faite le relai de compréhension entre P. Usseglio-Polatera et moi. Dans son emploi du temps de ministre, elle a trouvé moyen de m'expliquer tout ce qui m'était obscur. Elle a également passé beaucoup de temps sur l'interprétation de l'application de mon travail aux invertébrés.

Ma thèse est conseillée par Michel Lafont sur le milieu sédimentaire. Quel pianiste ! Ceci mis à part, il fourmille d'idées et ses critiques et suggestions m'ont amené beaucoup de pistes à développer. Enfin, et surtout je dirais, Michèle Trémolières, l'experte macrophyte de cette thèse. Elle fait partie de l'UdS et donc a une proximité géographique qui a fait que je l'ai particulièrement sollicitée. Elle a suivi toutes les propositions et tentatives avortées ou non. Je la remercie aussi pour ses explications pédagogiques, et la simplicité avec laquelle elle a fourni ses expertises, et également d'avoir accepté d'être examinatrice de ma thèse.

Je voudrais aussi remercier les membres de l'équipe Fouille de Données et Bioinformatique Théorique dans laquelle je commence à faire partie des meubles. J'y suis entrée sous la direction de Jerzy Korczak. Il est professeur émérite actuellement et donc n'est plus présent au laboratoire, mais je veux le remercier parce que c'est par lui que

m'est venue l'idée et l'intérêt pour la recherche, et qu'il a tout fait pour m'y encourager et pour soutenir cette décision. De manière générale, je remercie toute l'équipe que j'é-tends à l'IUT informatique d'Illkirch pour la bonne ambiance et la bonne humeur de travail. Ce fut un vrai plaisir.

Chapeau bas particulier pour Sébastien Derivaux, Jonathan Weber et Gwenaël Im-feld, mes colloc' de bureau qui ont réussi à me supporter toutes ces années.

Par ailleurs, une thèse est un travail permanent qu'on ne laisse pas derrière soi en fermant la porte du bureau. C'est partie prenante de la vie quotidienne. Pour arriver au bout il vaut mieux être bien entouré. Je tiens donc à remercier ma famille pour leur soutien. Par ailleurs, j'ai trouvé en Alsace une seconde famille et des amis, dont la liste est trop importante pour être citée mais qui se reconnaîtront si jamais ils ouvrent cette thèse. Je ne veux surtout pas les oublier ici.

Table des matières

1	Hydrobiologie	9
1.1	Caractérisation de l'état des masses d'eau	10
1.1.1	La directive cadre européenne	10
1.1.2	La physico-chimie	10
1.1.3	Les indices biologiques	11
1.2	Traits biologiques et traits écologiques	15
1.2.1	Etat de l'art de travaux connexes	16
1.2.2	Traits fonctionnels	16
1.2.3	Analyse fonctionnelle	18
1.3	Méthodes de fouille de données en hydrobiologie	19
1.3.1	Réseaux de neurones	19
1.3.2	Le système CN2	20
1.3.3	Arbre de régression	20
1.4	Le projet Indices	20
1.4.1	Qualification des milieux (alsaciens) à l'aide des traits de vie . .	21
1.4.2	Données considérées pour nos travaux	22
1.4.3	Contexte des invertébrés	24
2	Treillis de Galois et Analyse de Concepts Formels	31
2.1	Ensembles ordonnés	32
2.2	Treillis et propriétés des treillis	34
2.2.1	Demi-treillis, treillis et treillis complet	34
2.2.2	Approche algébrique	34
2.2.3	Treillis de Galois	35
2.2.4	Analyse de Concepts Formels	36
2.3	Treillis sur des contextes non triviaux	38
2.3.1	Contexte multi-valué	38
2.3.2	Structures de patrons	40
2.3.3	Contexte flou	42
2.3.4	Objets symboliques	43
2.3.5	Connexion de Galois par similarité	44

2.3.6	Règles d'association et implications	45
2.4	Conclusion	46
3	Traitement binaire	47
3.1	Les données et leurs pré-traitements	49
3.1.1	Règles d'implication	49
3.1.2	Ajustement du jeu de données	52
3.2	Méthodes de conversion en tableau binaire	53
3.2.1	Echelle nominale	56
3.2.2	Echelle histogramme	57
3.3	Comparaison et combinaison des treillis et des méthodes statistiques . .	62
3.3.1	Analyse factorielle multiple	62
3.3.2	Apport des treillis	71
3.4	Démarche complète d'identification de traits écologiques	77
3.4.1	Analyse des concepts formels	77
3.4.2	Validation	80
3.5	Conclusion	84
4	Traitement flou par histogrammes	85
4.1	Le format histogramme	86
4.1.1	Echelonnage histogramme	86
4.1.2	Fermeture de Galois sur les histogrammes	86
4.1.3	Connexion min-max seuillée	88
4.1.4	Illustration d'un contexte multi-valué flou	88
4.1.5	Structure de patron d'histogramme	91
4.2	Algorithmes de construction de treillis	92
4.2.1	Algorithme NextClosure de Ganter	93
4.2.2	Algorithme Charm	96
4.3	Algorithmes adaptés	97
4.3.1	MinMaxNC	97
4.3.2	MinMaxC	100
4.4	Test de validation des algorithmes	103
4.4.1	Données	104
4.4.2	Concepts	104
4.4.3	Délai d'exécution	105
4.5	Seuillage des histogrammes	106
4.5.1	Seuil brut par trait	107
4.5.2	Seuil fondé sur les écarts réels entre les affinités	108
4.5.3	Seuil par la variance et l'écart-type	108
4.5.4	Seuil par les quantiles	111
4.5.5	Comparatif des deux algorithmes seuillés par quantiles	112

4.6	Conclusion	115
5	Applications à la biologie	117
5.1	Macrophytes	117
5.1.1	Processus de mise en évidence des traits de vies caractéristiques de groupes taxonomiques	117
5.1.2	Application	118
5.2	Invertébrés	122
5.2.1	Groupes biotypologiques d'invertébrés	122
5.2.2	Groupes d'invertébrés par guidage latticiel	123
5.2.3	Groupes d'invertébrés selon la classification de P. Usseglio-Polatera et al.	127
5.3	Conclusion	133
6	Conclusion et perspectives	135
6.1	Traitement binaire	135
6.2	Traitement flou	137
6.3	Application	138
6.4	Discussion et perspectives	139
	Bibliographie	141
	Glossaire biologique	147
A		155
A.1	Stations du projet Indices	156
A.2	Indice IBGN	158
A.3	Groupes biotypologiques de Verneaux	161
A.4	Implications entre les propriétés Lxy	164
A.5	Algorithme NextClosure	181
A.6	Concepts issus du contexte histogramme pour la validation des algorithmes	182

Table des figures

1.1	Macroinvertébré <i>Hydropsyche</i>	12
1.2	Diatomée <i>Cocconeis krammerii</i>	12
1.3	Macrophyte <i>Nymphaea alba</i>	13

1.4	Espèces des indices normalisés dans un schéma de rivière.	30
2.1	Treillis du contexte binaire filmographique de la table 2.4.	41
2.2	Individus, descriptions et objets symboliques	44
3.1	Traitements binaires	48
3.2	Répartition des implications entre les différents supports.	50
3.3	Jeu de données actuel	54
3.4	Treillis pour le trait de "taille potentielle" pour un sous-ensemble de 15 espèces.	57
3.5	Treillis complet pour le contexte échelonné nominalement (limité à 21 objets et 10 traits).	58
3.6	Exemples d'histogrammes pour le trait "taille potentielle".	60
3.7	Treillis pour le trait histogramme de "taille potentielle" pour un sous ensemble de 15 espèces.	62
3.8	Treillis de Galois pour le contexte histogramme complet (15 traits). . . .	63
3.9	Comparaison et combinaison des méthodes statistique et latticielle . . .	64
3.10	Résultats de l'AFM. Plan principal : traits biologiques	65
3.11	Cercle de corrélation avec les propriétés du contexte binarisé	66
3.12	Résultats de l'AFM. Plan principal : les espèces de macrophytes	69
3.13	Treillis du contexte de l'axe 1. A gauche le sous-treillis des plantes à coordonnées négatives, à droite celui des positives	72
3.14	Concepts du treillis du contexte de l'axe 1 avec les propriétés à coordonnées positives	73
3.15	Sous-treillis du treillis du contexte de l'axe 2 avec les plantes à coordonnées positives	75
3.16	Sous-treillis du treillis du contexte de l'axe 2 avec les plantes à coordonnées négatives	76
3.17	Méthode complète d'identification des traits écologiques	78
3.18	Seconde étape de la méthode : analyse de concepts formels.	78
3.19	Dernière étape de la méthode : la validation.	81
3.20	Treillis de Galois pour un sous-ensemble de 25 macrophytes, 10 traits biologiques et 5 écologiques.	82
3.21	Mécanisme d'abstraction des espèces	83
4.1	Treillis de Galois de l'extrait de données sur les traits écologiques en correspondance de Galois classique.	90
4.2	Treillis de Galois par la correspondance min-max.	91
4.3	Nombre de concepts en fonction d'un seuil selon la variance des valeurs des modalités.	108

4.4	Délai de calcul des concepts en fonction d'un seuil selon la variance des valeurs des modalités.	109
4.5	Nombre de concepts en fonction d'un seuil selon les quantiles.	112
5.1	Apports relatifs de 21 unités taxonomiques à 6 groupes de taxons possédant des attributs biologiques et écologiques "similaires".	129
A.1	Carte de localisation des stations dans la plaine d'Alsace [64].	156
A.2	Tableau d'échantillonnage vierge pour l'indice IBGN	159

Introduction

Analyse de Concepts Formels

Cette thèse a pour objectif d'étudier des méthodes pour la fouille de données complexes, en particulier l'Analyse de Concepts Formels (ACF). De nombreuses méthodes existent dans le cadre de la fouille de données et permettent l'extraction de connaissances. Les méthodes de classification se divisent en deux branches : la classification supervisée et non supervisée. L'approche supervisée permet de guider l'apprentissage par des connaissances préalables grâce à des jeux tests. Les méthodes non supervisées sont dites *guidées par les données* afin de les regrouper selon leur similarité intrinsèque et extraire des classes d'objets pouvant être interprétées.

L'Analyse de Concepts Formels (ACF) est une méthode de classification non supervisée qui offre une représentation visuelle naturelle et aisément lisible pour une interaction plus facile avec des experts et a également l'avantage d'être un processus d'extraction de connaissance [55] sans effet « boîte noire ». Cette méthode permet d'extraire des groupes d'objets et de propriétés nommés *concepts formels* (ou plus simplement appelés *concepts*). Un concept est formé d'un ensemble maximal d'*objets* (appelé *extension*) et d'un ensemble maximal d'*attributs* (appelé *intension*) tels que tous les objets de l'extension partagent tous les attributs de l'intension et vice-versa [9, 24, 37]. Les concepts sont ordonnés dans un graphe nommé *treillis de Galois* selon une double hiérarchie, telle qu'un concept hérite des propriétés de ses concepts pères et des objets de ses concepts fils. L'ACF se base sur des *connexions* et *fermetures de Galois* qui sont des fonctions mathématiques, pour mettre en correspondance ces couples d'ensembles. Ces connexions manipulent des données qui se présentent initialement sous forme d'une table binaire appelée *contexte* indiquant quels objets et quels attributs sont reliés. Cependant de récents travaux montrent que l'ACF peut être appliquée à des données plus complexes que des contextes binaires, telles que des données floues, des intervalles, ou des données numériques, notamment par l'utilisation de mesures de similarité ajustées aux attributs et des connexions de Galois adaptées à ces mesures [17, 47, 52, 56].

Objectif de la thèse

Nos travaux s'intéressent à la manipulation de contextes spécifiques, que nous appelons multi-valués flous par les treillis de Galois et proposent également des solutions au problème de la génération massive de concepts. Nous étudions des contextes complexes, dont la complexité repose sur deux axes : d'une part, les *contextes multi-valués* dont les attributs se divisent en plusieurs modalités ; d'autre part, les *contextes flous* dont les relations entre les objets et les attributs ne sont pas binaires. Nous définissons les *contextes multi-valués flous* qui peuvent être vus comme une combinaison de contextes flous, à savoir que chaque trait forme à lui seul un contexte flou entre ses différentes modalités. Nous présentons alors deux conversions des données multi-valuées floues. La première conversion est une binarisation par une disjonction totale des attributs permettant d'une part l'exploitation d'outils comme des *règles d'implication* [28], et d'autre part de comparer et combiner les treillis avec des méthodes statistiques telles que l'*analyse factorielle* [44]. La seconde conversion est issue de l'*échelonnage histogramme*, que nous définissons, qui convertit les attributs en *histogrammes* et permet ainsi de conserver la notation multi-valuée de la structure des données. Afin de générer les concepts à partir des histogrammes, nous proposons une fermeture de Galois basée sur une mesure de similarité entre ces histogrammes. Cette fermeture permet d'obtenir des concepts pour lesquels les objets possèdent des attributs non plus égaux mais similaires, compris entre un minimum et un maximum communs. Nous proposons également des mesures de seuillage pour limiter le nombre de concepts générés et diminuer les temps de calculs. Enfin, cette fermeture de Galois et ces seuillages ont été implémentés dans de nouvelles versions de deux algorithmes issus de l'état de l'art, NextClosure et Charm, que nous testons et dont nous comparons les performances.

Applications à la biologie

Cette thèse trouve son application notamment dans le domaine hydrobiologique dont une problématique est la définition d'assemblages de traits de vie permettant de caractériser l'état écologique des cours d'eau. La qualité de l'eau est un problème majeur dont l'importance a été soulignée par la Directive Cadre Européenne (DCE) [10] adoptée en octobre 2000. Un des principaux objectifs de cette directive est d'atteindre d'ici 2015 un bon état écologique pour les milieux aquatiques naturels. Pour cela, il convient d'établir la typologie des cours d'eau, de définir les sites et les valeurs de référence permettant d'évaluer l'état écologique de tout l'écosystème par rapport aux pressions qu'il endure. Les pressions sont le résultat de l'exercice des activités humaines pouvant avoir une incidence sur les milieux aquatiques. Il peut s'agir, par exemple, de prélèvements ou de rejets dans les milieux aquatiques ou encore de leur artificialisation. En France, les eaux courantes sont qualifiées par des paramètres physico-chimiques

et des paramètres biologiques. La physico-chimie permet d'obtenir des résultats très précis et fiables, mais ces mesures peuvent être onéreuses et ne permettent de relever que les pollutions présentes au moment du prélèvement. Les paramètres biologiques sont les taxons (espèces, familles...) de différents compartiments d'organismes vivants (macrophytes, invertébrés, diatomées, oligochètes et poissons). Ces cinq catégories d'organismes sont utilisées pour définir des indices normalisés (AFNOR) pour l'évaluation des états écologiques. Au contraire de la physico-chimie, la biologie fournit des informations sur l'écosystème et garde des traces de pressions momentanées. Cependant les résultats biologiques sont difficiles à comparer puisqu'ils sont établis sur des expertises compartimentées (différents indices biologiques) et régionales : les espèces (plantes, poissons, ...) vivent dans différentes régions. Ainsi la biologie et la physico-chimie ne sont pas suffisantes et de nouveaux outils sont nécessaires pour l'évaluation de l'écosystème. Les biologistes étudient ainsi les *traits de vie* des espèces : leurs caractéristiques biologiques et écologiques. Les traits biologiques regroupent des caractéristiques morphologiques et physiologiques. Les traits écologiques caractérisent le fonctionnement des espèces au sein de leur milieu. Dans la littérature, de nombreux traits écologiques ont été décrits [14, 2, 69] et le principal problème est de sélectionner les traits les plus adaptés à la caractérisation de la qualité de l'eau. Pour sélectionner les *traits écologiques*, nous recherchons des groupes d'espèces et de *traits biologiques* qui soient caractéristiques d'un milieu, autrement dit des groupes d'espèces partageant des traits biologiques communs que nous pouvons associer à des traits écologiques. De tels groupes peuvent être considérés comme des concepts, et l'ACF se révèle être un outil puissant à la fois pour la construction de concepts et l'interaction avec l'expert dans un processus d'extraction de connaissances.

Nous avons ainsi développé une démarche complète qui a été appliquée avec deux experts différents sur deux groupes de taxons : macrophytes et invertébrés.

Organisation du mémoire

Plan

Le document se découpe comme suit.

Le premier chapitre présente le contexte hydrobiologique de la thèse. Après avoir développé l'objectif de la directive cadre européenne sur l'eau, nous décrivons les moyens de caractériser les masses d'eau par les outils physico-chimiques et biologiques. Nous introduisons les traits biologiques et écologiques auxquels nous nous intéressons puis les outils biologiques et les méthodes de fouille de données que les hydrobiologistes utilisent. Enfin nous présentons le projet Indices dans lequel cette thèse s'inscrit, ainsi que les données sur lesquelles nous travaillons.

Dans le second chapitre nous présentons l'ACF et les treillis de Galois. Nous rap-

pelons les définitions mathématiques sur lesquelles s'appuient les treillis de Galois. Nous présentons les travaux menés sur des données complexes sur lesquels nous nous appuyons, et enfin les outils connexes aux treillis.

Le troisième chapitre présente le traitement binaire des données. Avant toute chose nous présentons les pré-traitements effectués sur les données ainsi que leur ajustement. Puis nous présentons deux méthodes de binarisation de nos données afin de leur appliquer les treillis d'une part et les méthodes statistiques d'autre part. Nous étudions donc ces deux approches par comparaison mais également par combinaison. Enfin nous présentons une démarche complète d'identification des traits écologiques.

Le chapitre 4 propose une conversion des données hydrobiologiques sous un format respectueux de l'information qu'elles représentent. Pour cela, nous introduisons la notion de *contexte multi-valué flou*, puis l'*échelonnage histogramme*; avant de proposer une *fermeture de Galois* permettant de les manipuler. Ensuite nous présentons les algorithmes NextClosure et Charm et leurs adaptations pour intégrer ces connexions de Galois nommées MinMaxNC et MinMaxC. Enfin, nous testons ces algorithmes et présentons les résultats obtenus. Dans un premier temps nous effectuons les tests nécessaires pour valider les modifications apportées aux algorithmes. Puis nous proposons d'introduire une notion de seuil afin de traiter le problème du volume des treillis qui peut devenir rapidement faramineux.

Enfin, le dernier chapitre présente les conclusions biologiques tirées sur nos données par l'utilisation des techniques non binaires. Dans un premier temps nous traitons des macrophytes, et dans un second temps nous abordons les invertébrés.

Notations

Dans ce mémoire nous utilisons diverses notations que nous précisons ici. La première d'entre elles est classique puisqu'il s'agit des références à la bibliographie. Par exemple pour indiquer un de nos articles intitulé "Mining Complex Hydrobiological Data with Galois Lattices", nous utilisons ce référencement [11] qui renvoie à la onzième référence de la bibliographie.

La seconde notation se marque en position exponentielle avec un chiffre incrémental et indique les notations de pied de page, elle se note ainsi ¹.

Enfin la troisième notation est un chiffre entre parenthèses en position exponentielle également et se rapporte au glossaire biologique se trouvant immédiatement après la bibliographie, elle se note ainsi ⁽⁰⁾.

Par ailleurs, le lecteur pourra être surpris de trouver le verbe *classifier* alors qu'il lui paraîtrait plus français d'utiliser le verbe *classer*. En fait, en fouille de données, ces deux termes ont un sens bien défini qu'il convient de préciser. *Classifier* est l'action

1. Ceci est une référence à une note de bas de page

de créer des classes et d'y ranger divers éléments. *Classer* est l'action de ranger divers éléments, dans des classes qui existent au préalable.

Données

Au cours de ce mémoire nous présentons des données qui peuvent paraître être issues de différents jeux de données. Nous clarifions donc ici l'origine des quelques différences. Nous traitons essentiellement de plantes aquatiques dont 50 espèces sont présentées page 23. Nous renseignons des caractéristiques biologiques et écologiques pour ces espèces de plantes. Cependant si tous les traits biologiques sont renseignés pour ces 50 espèces (page 54), seulement 47 le sont également pour les traits écologiques. Ainsi lorsque nous avons besoin de comparer tous les traits de vie (biologiques et écologiques) nous nous basons sur le jeu des 47 espèces. Pour d'autres applications, nous sommes contraints de limiter le nombre de données pour simplifier des calculs, ainsi nous ne prenons qu'une portion des données, mais il s'agit toujours des mêmes 50 espèces d'origine.

Résumé

Cette thèse en informatique se place dans le cadre de l'*Analyse de Concepts Formels* (ACF) ou les *treillis de Galois*. Ce sont des outils basés sur des opérateurs mathématiques appelés *fermetures de Galois* permettant de calculer des *concepts*. Un concept est formé d'un ensemble d'*objets* partageant tous un ensemble d'*attributs* communs. Ces concepts sont extraits à partir d'un *contexte* qui est une table de relation binaire entre ces objets et ces attributs.

Nous nous intéressons à des contextes complexes dont la complexité repose sur deux axes. D'une part, les contextes multi-valués dont les attributs se divisent en plusieurs modalités. D'autre part, les contextes flous dont la relation entre objets et attributs n'est pas binaire. Nous définissons les *contextes multi-valués flous* qui héritent de ces deux complexités et présentons deux conversions des données multi-valuées floues. La première conversion est une binarisation par une disjonction totale des attributs permettant d'une part l'exploitation d'outils comme des implications et d'autre part de comparer et combiner les treillis avec des méthodes statistiques telles que l'analyse factorielle. La seconde conversion est issue de l'*échelonnage histogramme* que nous définissons et qui permet de convertir les attributs en *histogrammes*. Afin de générer les concepts à partir des histogrammes, nous proposons une nouvelle *fermeture de Galois* basée sur une mesure de *similarité* entre ces histogrammes. Cette fermeture permet d'obtenir des concepts pour lesquels les objets possèdent des attributs non plus égaux mais similaires compris entre un minimum et un maximum communs. Nous proposons également des mesures de seuillage pour limiter le nombre de concepts générés et diminuer les temps de calculs. Enfin, deux algorithmes ont été testés pour implémenter cette fermeture : MinMaxNC et MinMaxC, dont nous comparons les performances.

Cette thèse trouve son application notamment dans le domaine hydrobiologique dont une problématique est la sélection de *traits écologiques* de taxons permettant de caractériser l'état écologique des cours d'eau par le comportement des espèces au sein de leur environnement. La sélection de ces traits s'appuie sur la recherche de groupes de taxons possédant des caractéristiques morphologiques et physiologiques (appelés *traits biologiques*) communes. Ces groupes correspondent à des concepts au sens de l'ACF et les données biologiques se présentent sous la forme d'un contexte multi-valué flou pour lequel nous montrons l'efficacité de notre approche.

Mots-clés: Hydrologie, Indices biologiques, Traits écologiques, Treillis de Galois, Analyse Formelle de Concepts, Analyse Factorielle Multiple, algorithme.

Abstract

This computer information science PhD takes place in the framework of *Formal Concept Analysis* (FCA) or *Galois lattices*, which are tools based on mathematical operators called *Galois connections* allowing to generate *concepts*. A concept is composed with a set of *objects* sharing a set of *attributes*. These concepts are generated from a *context* which is a table of binary relations between these objects and these attributes.

We are interested in complex contexts for which the complexity is based on two elements. On one hand, on many-valued context for which the attributes are divided into several modalities. On the other hand, it is based on fuzzy contexts for which the relation between objects and attributes is not binary.

We define *fuzzy many-valued contexts* which inherit of both complexities and introduce two conversions for fuzzy many-valued data. The first conversion is a binarisation by a complete disjunctive operation allowing to use tools such as implications and to compare and combine lattices with statistical methods such as factorial analysis. The second conversion is issued from *histogram scaling* which we define and which converts attributes into *histograms*. To generate concepts from histograms, we propose new Galois connections based on a similarity measure between these histograms. These connections allow to obtain concepts where objects share attributes which are not equal but similar between the same minimum and maximum. We also propose to use thresholds to limit the number of generated concepts and decrease calculating time. We have tested and compared the performance of two algorithms : MinMaxNC and MinMaxC implementing this connection.

This PhD is applied to the hydrobiological domain for which it is needed to select *ecological traits* allowing to characterize ecological quality of water surfaces due to the behaviour of species in their environment. The selection of these traits is based on the search of groups of taxons sharing morphological and physiological (called *biological traits*) characteristics. These groups correspond to concepts in FCA and biological data can be considered as fuzzy many-valued context for which we show the efficiency of our approach.

Keywords: Hydrobiology, biological indices, ecological features, Galois lattice, Formal Concept Analysis, Multiple Factor Analysis, algorithm.

Chapitre 1

Hydrobiologie

Sommaire

1.1	Caractérisation de l'état des masses d'eau	10
1.1.1	La directive cadre européenne	10
1.1.2	La physico-chimie	10
1.1.3	Les indices biologiques	11
1.2	Traits biologiques et traits écologiques	15
1.2.1	Etat de l'art de travaux connexes	16
1.2.2	Traits fonctionnels	16
1.2.3	Analyse fonctionnelle	18
1.3	Méthodes de fouille de données en hydrobiologie	19
1.3.1	Réseaux de neurones	19
1.3.2	Le système CN2	20
1.3.3	Arbre de régression	20
1.4	Le projet Indices	20
1.4.1	Qualification des milieux (alsaciens) à l'aide des traits de vie	21
1.4.2	Données considérées pour nos travaux	22
1.4.3	Contexte des invertébrés	24

Un des objectifs de cette thèse est de répondre à des problématiques hydrobiologiques par le biais de méthodes informatiques de fouille de données. Nous commençons, dans ce chapitre, par donner des précisions sur le contexte hydrobiologique. Nous évoquons la caractérisation de l'état des masses d'eau qui est un problème soulevé par les instances européennes ainsi que les mesures actuelles d'évaluation qui passent par des analyses physico-chimiques ou biologiques. Certaines analyses biologiques s'appuient sur des caractéristiques biologiques d'animaux ou de végétaux vivant dans ces eaux. Nous détaillons donc ces caractéristiques et les approches scientifiques qui les utilisent.

Par ailleurs des solutions informatiques de fouille de données ont déjà été utilisées pour aborder ce problème, nous décrivons lesquelles ainsi que leur pertinence. Enfin nous présentons le projet *Indices*, dans lequel cette thèse s'inscrit, qui a pour objectif de proposer un système d'harmonisation fondé sur plusieurs indices biologiques pour l'évaluation de l'état des masses d'eau.

1.1 Caractérisation de l'état des masses d'eau

1.1.1 La directive cadre européenne

La mise en place de mesures visant à limiter les problèmes environnementaux est à l'heure actuelle un enjeu majeur, notamment en ce qui concerne la pollution de l'eau. L'Europe a ainsi mis en place la Directive Cadre Européenne sur l'eau (DCE) du 23 octobre 2000 (Directive 2000/60/CE) qui a été rédigée pour homogénéiser et améliorer la surveillance et l'évaluation de la qualité des eaux au niveau européen [10], et a pour objectif le *bon état écologique* de toutes les masses d'eau⁽³⁰⁾ en 2015. La DCE a pour objet "d'établir un cadre pour la protection des eaux intérieures⁽¹⁰⁾ de surface⁽¹²⁾, des eaux de transition⁽¹³⁾, des eaux côtières⁽⁹⁾ et des eaux souterraines⁽¹¹⁾" [10]. Cette thèse s'intéresse à l'eau des rivières.

Un "bon état écologique" est défini ainsi : "un état ne présentant que de légères distorsions d'origine humaine par rapport à une situation non perturbée. Ce bon état écologique doit tenir compte des caractéristiques naturelles des masses d'eau de surface". Il existe cinq classes d'état écologique qui s'établissent par rapport à des états de référence à définir localement : très bonne, bonne, moyenne, médiocre et mauvaise. Les classes sont les mêmes quels que soient les régions et les pays. Par contre les outils sont différents (caractéristiques physico-chimiques différentes, peuplements différents) et les seuils diffèrent également pour un même pays dans chaque hydro-éco-région, pour chaque type de masse d'eau. Différents programmes de recherche nationaux et européens sont mis en œuvre pour proposer un système d'évaluation et des outils DCE compatibles, comparables entre les divers pays, pour les différents types de masse d'eau.

En France, à l'heure actuelle, on utilise des indicateurs physico-chimiques et biologiques que nous détaillons.

1.1.2 La physico-chimie

La circulaire DCE 2006/16 du 13 juillet 2006 précise les éléments physico-chimiques à mesurer ainsi que les conditions de prélèvements pour les masses⁽³⁰⁾ d'eau. Que ce soit pour les cours d'eau ou les plans d'eau, il existe trois catégories de paramètres : ceux mesurés in situ, ceux mesurés en laboratoire sur eau brute et enfin ceux mesurés en laboratoire sur la phase solide du sédiment. Différents paramètres sont mesurés comme par

exemple, la température, l'oxygène dissous, le pH... ou les teneurs en nitrates, chlorures ou phosphates...

1.1.3 Les indices biologiques

Les organismes vivants caractérisent l'état écologique des milieux qu'ils peuplent car ils sont en contact direct avec ces milieux. Toute pression sur un milieu entraîne des modifications plus ou moins importantes des peuplements⁽³⁸⁾ qui y vivent. La présence des taxons⁽⁶¹⁾ les plus polluo-sensibles, la prolifération des taxons les plus polluo-résistants ainsi que leur état témoignent des pressions subies par le milieu sur toute une période de temps et en sont donc des mémoires révélant des perturbations actuelles comme passées. La définition des communautés floristiques et faunistiques permet donc d'évaluer l'état de ces milieux.

Pour cela elle se base sur des *indices biologiques* qui s'établissent sur des taxons végétaux ou animaux vivant dans les milieux aquatiques. Il existe cinq *indices biologiques* (normalisés en France) que nous présentons par ordre chronologique de normalisation.

Indice Biologique Général Normalisé (IBGN)

Cet indice a été normalisé par l'AFNOR en 1992 et révisé en mars 2004 par la norme NF T90-350 [1] et se base sur les macroinvertébrés benthiques⁽¹⁾ (insectes, mollusques, crustacés...) d'une taille supérieure à 0,5 mm.

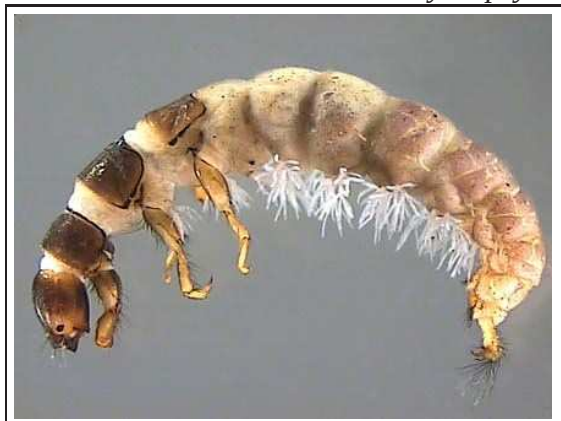
Il s'agit de petits animaux qui vivent au fond des rivières et plus particulièrement sur et dans les sédiments. La figure 1.1 présente l'un d'eux : un *Hydropsyche*². Cet indice repose sur l'absence ou la présence de groupes indicateurs (des taxons⁽⁶¹⁾) ainsi que sur la diversité du peuplement⁽³⁸⁾ évalué par le nombre de groupes représentés. Parmi l'ensemble des macroinvertébrés⁽²⁸⁾, 138 taxons⁽⁶¹⁾ sont retenus pour calculer l'IBGN.

Cet indice qualifie la capacité d'un milieu à accueillir une biocénose⁽²⁾ en fonction de la qualité physico-chimique de l'eau et des caractéristiques morphologiques de la masse d'eau. Il permet de mesurer les effets d'une perturbation (et non pas de déterminer sa nature) que ce soit dû à une mauvaise qualité de l'eau ou une dégradation des habitats⁽¹⁹⁾.

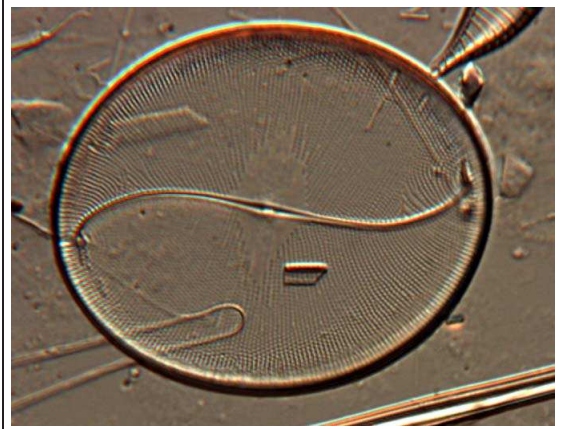
Indice Biologique Diatomées (IBD)

Cet indice a été normalisé par l'AFNOR par la norme NFT 90-354 [3] en juin 2000 et révisée en décembre 2007. Il vise à étudier la qualité des eaux à partir des diatomées benthiques⁽¹⁾. Il s'agit d'algues microscopiques unicellulaires mesurant de quelques

2. Photographie de Guillaume DOUCET.

FIGURE 1.1 – Macroinvertébré *Hydropsyche*

millièmes de millimètres à un demi millimètre. Leur grande diversité (plusieurs milliers d'espèces en eau douce) et leur grande sensibilité à la pollution en font des indicateurs très intéressants et répandus dans de nombreux pays. L'IBD repose sur la *polluo-sensibilité* des diatomées qui ne dépendent pas du support où elles vivent, mais de la qualité physico-chimique de l'eau. Cet indice permet donc d'obtenir le niveau de pollution générale physico-chimique de l'eau (et pas des sédiments). La photographie³ 1.2 présente l'espèce de diatomée : *Cocconeis krammerii*.

FIGURE 1.2 – Diatomée *Cocconeis krammerii*

3. prise par Ania Wachnicka, sous licence GNU.

Indice Oligochètes de Bioindication des Sédiments (IOBS)

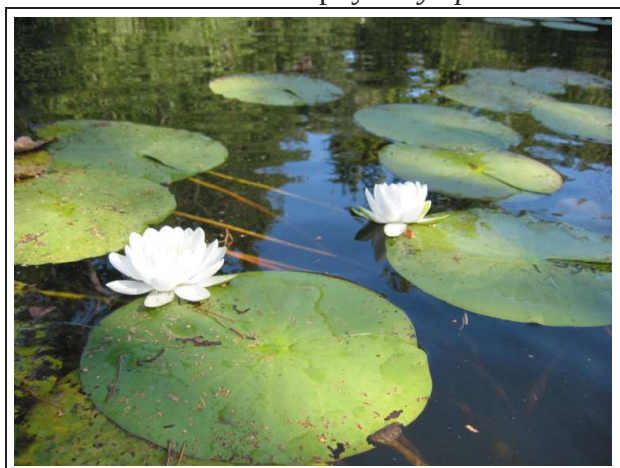
L'IOBS a été normalisé par l'AFNOR par la norme NFT 90-390 [4] en avril 2002. Il se base sur la richesse taxonomique⁽⁶²⁾ et le pourcentage de Tubificidae (famille d'oligochètes) qui sont les plus résistants à la pollution. Il permet l'évaluation des sédiments fins ou sableux, permanents ou stables des cours d'eau ou des canaux et indique des tendances fortes sur l'incidence écologique des rejets polluants (charge organique, micro-polluants organiques et métalliques).

Les oligochètes, sont des vers annélides (avec des anneaux) cylindriques et segmentés. Ils mesurent d'1 mm à plusieurs mètres et vivent dans les substrats⁽⁵⁹⁾ de tous les milieux aquatiques et certains peuvent nager à la surface. Ils ont un cycle de vie d'une ou deux années. La plupart survivent à de faibles quantités d'oxygène. Ils se nourrissent de débris organiques et d'algues qu'ils tirent de la boue des fonds qu'ils ingèrent. Ainsi on les trouve de manière abondante dans les zones polluées par les matières organiques.

Indice Biologique Macrophytique en Rivière (IBMR)

Cet indice a été normalisé par l'AFNOR en octobre 2003 par la norme NFT 90-395 [5]. Il se fonde sur les macrophytes⁽²⁹⁾ qui désignent les grands végétaux aquatiques (bryophytes⁽⁵⁾, ptéridophytes⁽⁴⁵⁾ et spermatophytes⁽⁵⁶⁾) et les algues filamenteuses telles que les Characées⁽⁶⁾, visibles et le plus souvent identifiables à l'œil nu. Les macrophytes, (voir exemple par la figure 1.3), servent de base alimentaire, d'habitat⁽¹⁹⁾, de lieux de refuge et de reproduction et ont une incidence très forte sur leur milieu, mais sont aussi très influencés par lui. Ils sont très sensibles à la qualité de l'eau et leurs présence, absence, multiplication, disparition ou abondance caractérisent l'état de leur milieu.

FIGURE 1.3 – Macrophyte *Nymphaea alba*



Cet indice se détermine par l'inventaire des espèces de macrophytes⁽²⁹⁾ et leur taux de recouvrement à chaque station⁴ échantillonnée, ce qui permet de définir l'état d'eutrophisation⁽¹⁸⁾ de ces stations.

Dans cette thèse, nous nous intéressons particulièrement à cet indice.

Indice Poisson Rivière (IPR)

L'Indice Poisson Rivière a été normalisé en mai 2004 par l'AFNOR par la norme NFT 90-344 [6]. Il consiste à mesurer l'écart entre la composition du peuplement sur une station donnée et la composition du peuplement attendue en situation de référence.

L'IPR se base sur les 34 espèces ou groupes d'espèces les plus représentatives en France et pour lesquelles il a été possible de modéliser la répartition en situation de référence.

Les classes d'indices

Chaque indice est une note calculée sur la base des taxons, leur abondance, leur sensibilité à une pollution ou une altération du milieu. Cependant ces notes ne sont pas établies sur les mêmes échelles d'une part, et d'autre part ces échelles n'évoluent pas dans le même sens. En effet, l'IPR évolue en sens inverse des autres, c'est à dire que plus sa note est faible, plus l'eau est de bonne qualité. L'IBD, l'IBGN et l'IBMR sont notés sur 20, alors que l'IOBS est sur 10 et l'IPR qui n'est en théorie pas limitée est plafonnée en pratique à 160.

Le Ministère chargé de l'environnement et les Agences de l'eau ont reconsidéré les grilles de qualité utilisées pour les remplacer par un nouveau Système d'Évaluation de la Qualité de l'Eau (SEQ-Eau). Le principe général de cet outil est d'évaluer une eau selon sa qualité physico-chimique ou selon l'aptitude de l'eau aux usages (i.e. la production d'eau potable, l'abreuvement des animaux, l'irrigation etc) ainsi que selon la biologie. Les concentrations mesurées sont confrontées à des limites de classes notamment établies sur la base de recommandations de l'Organisation Mondiale de la Santé (OMS) et converties en indices de qualité. Ces indices permettent de juger de la qualité de l'eau pour un paramètre, une altération (en retenant l'indice le plus faible obtenu pour l'ensemble des paramètres de l'altération) ou un ensemble d'altérations (en retenant l'indice le plus faible obtenu pour l'ensemble des altérations considérées).

Le SEQ (Système d'Evaluation de la Qualité) permet d'harmoniser les mesures au niveau national et d'avoir une évaluation de la qualité conforme à la Directive Cadre

4. Le Secrétariat d'Administration National des Données Relatives sur l'Eau (SANDRE⁽⁵¹⁾) [23] définit une *station de mesure*⁽⁵⁴⁾ comme "un lieu situé sur un cours d'eau ou plan d'eau, sur lequel sont effectués des mesures ou des prélèvements en vue d'analyses biologiques, afin de déterminer la qualité des milieux aquatiques à cet endroit. Il s'agit d'un volume dans lequel il est possible de faire des mesures en différents points réputés cohérents et représentatifs de la station."

Eau. Il se décompose en :

- SEQ Eau, qui identifie la nature des perturbations et évalue les incidences sur l'environnement grâce à la physicochimie.
- SEQ Physique, qui s'intéresse à l'ensemble des paramètres intervenant dans l'architecture, la forme du cours d'eau et dans son fonctionnement hydrodynamique.
- Le SEQ Bio, qui s'appuie sur des méthodes fondées sur l'examen de différents groupes d'organismes, végétaux et animaux.

Pour échelonner ces indices, cinq classes de qualité de "très mauvaise" à "très bonne" ont été créées [54]. Chaque indice se situe ensuite par rapport à ces classes en fonction de sa note comme l'indique le tableau 1.1.

Note IBD (sur 20)	Note IBGN (sur 20)	Note IOBS (sur 10)	Note IPR (sur 160)	Note IBMR (sur 20)	Couleur	Qualité
$IBD < 5$	$IBGN \leq 4$	$IOBS < 1$	$36 < IPR$	$IBMR \leq 8$	Rouge	Mauvaise
$5 \leq IBD < 9$	$5 \leq IBGN \leq 8$	$1 \leq IOBS < 2$	$25 < IPR \leq 36$	$8 \leq IBMR < 10$	Orange	Médiocre
$9 \leq IBD < 13$	$9 \leq IBGN \leq 12$	$2 \leq IOBS < 3$	$17 < IPR \leq 25$	$10 \leq IBMR < 12$	Jaune	Moyenne
$13 \leq IBD < 17$	$13 \leq IBGN \leq 16$	$3 \leq IOBS < 6$	$7 < IPR \leq 16$	$12 \leq IBMR < 14$	Verte	Bonne
$17 \leq IBD$	$17 \leq IBGN$	$6 \leq IOBS$	$IPR \leq 7$	$14 \leq IBMR$	Bleue	Très bonne

TABLE 1.1 – Classes de qualité.

1.2 Traits biologiques et traits écologiques

Le problème de l'homogénéité des classes est soulevé par le fait que les taxons servant d'indicateur aux indices ne vivent pas tous dans toutes les régions où les analyses sont effectuées. En effet, les indices fonctionnent bien sur toute la France métropolitaine, mais on sait que ce n'est pas très opérationnel sur certains secteurs comme l'IBGN en façade méditerranéenne alors qu'il l'est dans l'est de la France ou l'IBD opérationnel surtout en Adour-Garonne, Artois-Picardie car c'est dans ces régions qu'il a été mis au point et testé. De plus, comme la DCE est européenne, il est important d'avoir des outils européens et les indices ne fonctionnent pas en dehors de l'hexagone (exceptés le Luxembourg et la Belgique). Ainsi il convient de chercher un palliatif à ce problème. Une des solutions retenues par les biologistes est de considérer les caractéristiques inhérentes à l'adaptation au milieu des espèces (physiologiques, morphologique) plutôt que les espèces elles mêmes. En se basant donc sur les traits de vie, un outil pourrait être appliqué de façon plus large. Par ailleurs, l'utilisation des traits de vie pourraient fournir des informations sur les types de pressions subies par les masses d'eau, contrairement aux indices qui sont très intégrateurs de presque toutes les pressions supportées.

1.2.1 Etat de l'art de travaux connexes

L'état écologique des masses d'eau est estimé à partir de la physico-chimie ou des cinq indices biologiques. La physico-chimie permet d'obtenir des résultats bornés par des seuils, autorisant ainsi des comparaisons qui restent cependant relatives dans le cadre des rivières car la chimie évolue en fonction de la géologie et de la distance à la source. De plus, les analyses physico-chimiques peuvent être onéreuses (notamment pour la recherche d'éléments toxiques comme des métaux ou des produits phytosanitaires⁵) à obtenir et ne peuvent témoigner de l'état de la masse d'eau qu'au moment de leur prélèvement. En outre, les indices biologiques gardent traces des pollutions ou des changements environnementaux puisqu'ils sont influencés par eux. Mais en échange, ils influencent aussi leur milieu en le détoxifiant par la capture de certaines substances par exemple, qui n'apparaîtront alors plus dans l'eau, ou par le relâchement dans l'eau de ces substances lors de la mort de l'individu (phénomène de résistance et de résilience⁽⁴⁷⁾).

De nombreux travaux abordent l'évaluation des masses d'eau par leurs caractéristiques biologiques. On peut notamment citer les travaux de Willby [69] qui s'appuient sur ceux de [14, 7, 22, 32, 39, 46, 57, 58, 60, 63, 65] avant lui. Cette approche implique d'une part la sélection d'un ensemble d'individus (ici des macrophytes⁽²⁹⁾) selon un degré taxonomique à choisir. D'autre part elle implique également de déterminer un ensemble de caractéristiques biologiques appelées *traits biologiques*. Willby a sélectionné 120 espèces en fonction des informations disponibles concernant leurs habitats et leurs caractéristiques. Il a également retenu 17 traits en fonction, là encore, de la disponibilité des informations. Et, dans un but d'analyses, ces auteurs, ont scindé certains traits en attributs (appelés *modalités*) nominaux (comme une texture de feuille : souple ou rigide) ou ordinaux (comme un index morphologique : 2 ou 10). Les 58 modalités ainsi obtenues servent de référence et sont présentées avec leur traits dans le tableau 1.2 [69] (dans leur version originale : en anglais).

Par ailleurs, des scores ont été alloués à chaque modalité pour prendre en compte la répartition des individus de chaque espèce entre les différentes modalités. Ces scores sont '0' pour indiquer l'absence d'individus possédant cette modalité, '2' pour indiquer leur abondance, et '1' pour traduire un caractère de présence occasionnelle. Des degrés de codage plus fins ont été envisagés mais finalement rejetés car ils étaient trop subjectifs. Nous appelons ces scores des *affinités*.

1.2.2 Traits fonctionnels

Utiliser les traits écologiques pour l'évaluation des cours d'eau a aussi été abordée dans [48, 49, 50]. Michel Lafont s'intéresse aux oligochètes et détermine à partir d'eux des traits fonctionnels. Il explique que les ressources alimentaires assimilables (biodisponibilité trophique⁽⁶⁴⁾) ajoutées à l'accessibilité constituent un facteur

5. Produit utilisé pour soigner ou prévenir les maladies des organismes végétaux.

Traits	Modalités	Traits	Modalités
Growth form	Free-floating, surface	Number of reproductive organs per year per individual	Low (< 10)
	Free-floating, submerged		Medium (10-100)
	Anchored, floating leaves		High (100-1000)
	Anchored, submerged leaves		Very high (> 1000)
	Anchored, emergent leaves	Perennation	Annual
	Anchored, heterophylly		Biennial/short lived perennial
Vertical shoot architecture	Single apical growth point		Perennial
architecture	Single basal growth point	Evergreen leaf	-
	Multiple apical growth point	Amphibious	-
Leaf type	Tubular	Gamete vector	Wind
	Capillary		Water
	Entire		Air bubble
Leaf area	footnotesize (< 1 cm2)		Insect
	Medium (1-20 cm2)		Self
	Large (20-100 cm2)	Body flexibility	Low (< 45°)
	Extra large (> 100 cm2)		Intermediate (> 45 – 300°)
	High (> 300°)		
Morphology index (score)	(1) 2	Leaf texture	Soft
	(2) 3-5		Rigid
	(3) 6-7		Waxy
	(4) 8-9		Non-waxy
	(5) 10		
Rooting at nodes	-	Period of production of reproductive organ	Early (March-May)
High below-ground : above-ground biomass	-		Mid (June-July)
			Late (August-September)
Mode of reproduction	Rhizome		Very late (post-September)
	Fragmentation	Fruit size	< 1 mm
	Budding		1-3 mm
	Turions		> 3 mm
	Stolons		
	Tubers		
	Seeds		

TABLE 1.2 – Traits et modalités sélectionnés par Willby

clé de développement des organismes vivants. Le nombre d'espèces et les effectifs d'oligochètes interstitiels hyporhéiques⁽²³⁾ et superficiels augmentent significativement dans les zones alimentées par des apports d'eaux souterraines.

Il précise que de considérer que le milieu récepteur est apte à dégrader les rejets est une erreur possible et s'intéresse aux sédiments fins qui sont une photographie globale de toutes les substances dangereuses ou non. Pour des recherches sur le fonctionnement des écosystèmes⁽¹⁴⁾, les sédiments grossiers superficiels et le milieu hyporhéique⁽²³⁾ doivent être considérés en premier selon [45].

Ces travaux menés ont conduit à une caractérisation de l'état écologique par 4 traits fonctionnels qui sont :

1. Le pourcentage d'espèces d'oligochètes indiquant les échanges hydrobiologiques actifs entre les eaux de surface et souterraines. Il s'agit d'espèces perméables à la pollution de l'eau.
2. Le pourcentage d'espèces d'oligochètes tolérants à une forte pollution de l'eau, incluant les substances toxiques.
3. Le pourcentage d'espèces d'oligochètes étant tolérants à la pollution de l'eau en sédiments grossiers.
4. Le pourcentage de taxons indiquant la présence de boues polluées dans les interstices des sédiments.

Ces traits s'appliquent particulièrement à refléter les états de pollution des milieux.

1.2.3 Analyse fonctionnelle

Dans [44], les auteurs expliquent que classifier les habitats en types de végétation selon leur composition en espèces similaires est classique en écologie et a largement inspiré les instances de conservation européenne [21]. En outre, le besoin de prédire les effets de perturbation sur la biodiversité et de comprendre le rôle de la biodiversité dans le fonctionnement d'un écosystème a récemment gagné en intérêt en étudiant les assemblages de plantes d'un point de vue fonctionnel. Classifier les espèces par leurs caractéristiques morphologique, physiologique et phénologique communes plutôt que par leur seule phylogénie⁽⁴⁰⁾ est un thème central de cette approche. Deux arguments préconisent des approches centrées sur des espèces complémentaires avec une analyse fonctionnelle :

- les classifications (centrées sur les espèces) d'habitats d'une région biogéographique ne peuvent être facilement utilisées pour classifier les habitats d'une autre région ayant des espèces différentes, contrairement à une classification basée sur des traits comportementaux.
- un but commun pour les gestionnaires environnementaux est de fournir des directives de conservation pour beaucoup d'espèces simultanément, et les *Groupes Fonctionnels* englobent les espèces ayant des attributs biologiques différents.

Bien que l'utilisation des traits comportementaux pour la classification des espèces est devenue de plus en plus fréquente en littérature, la classification des habitats a reçu moins d'attention. En fait, les classifications fonctionnelles proposées sont souvent basées sur une matrice [*habitat* $\times$ *traits comportementaux*] obtenue par la multiplication des matrices [*habitat* $\times$ *espèces*] et [*espèces* $\times$ *traits comportementaux*]. Cependant, pour les auteurs, classer les habitats par l'abondance des traits comportementaux semble plutôt irréaliste puisque de telles procédures considèrent les habitats comme des assemblages libres de traits comportementaux. De plus, au cours de l'histoire de l'évolution, les traits de vie n'ont pas été regroupés aléatoirement et par conséquent, ne sont pas distribués aléatoirement entre les espèces [30, 40] .

1.3 Méthodes de fouille de données en hydrobiologie

Les biologistes ont habituellement recours à des méthodes statistiques telles que l'Analyse Factorielle Multiple (AFM) ou la Classification Ascendante Hiérarchique (CAH). Cependant des rapprochements avec d'autres domaines comme l'informatique ont suscité l'emploi de techniques différentes telles que des méthodes de fouille de données. Dans l'article [38], les auteurs référencent les outils de fouille de données auxquels les hydrobiologistes ont parfois recours. On peut notamment citer les réseaux neuronaux, le système CN2 et les arbres de régression.

1.3.1 Réseaux de neurones

Les réseaux de neurones sont des outils de calcul s'inspirant des modèles neurologiques du cerveau et sont régulés par des méthodes d'apprentissage statistiques. Les données de départ se présentent sous la forme de *neurones d'entrée* et les résultats sous celle de *neurone(s) de sortie*. Entre les deux, on peut placer une ou plusieurs couches de neurones qui sont interconnectés et l'information va circuler en va et vient entre les couches en modifiant des poids attribués à ces neurones jusqu'à fournir un réseau fournissant la bonne solution. Pour le paramétrage du réseau, on travaille sur un jeu de données dont la solution est connue, appelé *jeu d'apprentissage*. Une fois le réseau étalonné, il est utilisable pour prédire les résultats dépendant d'un nouveau jeu de données. Les systèmes écologiques étant non linéaires, les réseaux neuronaux sont souvent utilisés pour les modéliser. Ils fournissent de bonnes prédictions, cependant elles sont compensées par l'effet "boîte noire" qui les rend difficile à comprendre. Ceci est pénalisant pour la validation de l'expert ainsi que pour les ajustements qu'il pourrait suggérer car ils ne peuvent être pris en compte sans risquer de perturber l'étalonnage du réseau.

1.3.2 Le système CN2

Il s'agit de règles d'induction [19] qui s'appliquent à des variables discrètes et permet d'obtenir des règles de la forme "SI prémisse ALORS conclusion" indiquant la ou les conditions (prémisse) pour lesquelles une ou des propriétés (conclusion) sont remplies (les prémisses sont de la forme : attribut-comparateur-valeur, par exemple : âge > 20 ans). Comme les réseaux de neurones, cette méthode s'appuie sur un ensemble d'apprentissage.

Cette méthode a déjà été utilisée pour prédire la qualité de l'eau à partir de propriétés biologiques ou physico-chimiques, et pour analyser l'influence de propriétés physico-chimiques sur la présence de taxons⁽⁶¹⁾.

La lisibilité des règles favorise l'interaction avec l'expert lui permettant d'ajouter des attributs pertinents pour ses évaluations.

1.3.3 Arbre de régression

Il s'agit d'un arbre de décision appliqué à des données à valeurs continues. Le principe général, aisément compréhensible, est que plus on est près de la racine de l'arbre, plus les paramètres influent sur la prédiction. Cette méthode a été appliquée pour les paramètres physico-chimiques en se basant sur les paramètres biologiques.

Diverses études ont été menées :

- Une étude double en considérant les taxons⁽⁶¹⁾ au niveau des espèces et de la famille. Les résultats au niveau des espèces sont plus précis. Cependant en considérant les familles, une famille particulière et importante dans l'évaluation de la qualité de l'eau se distingue.
- Une étude pour trouver des relations entre des propriétés physico-chimiques de l'eau et la diversité des peuplements⁽³⁸⁾ en nombre de taxons. Cela permet de mettre en évidence l'interdépendance des paramètres.
- Une étude pour déterminer des paramètres physico-chimiques passés en fonction de paramètres biologiques pour savoir comment les premiers influencent les seconds.

L'utilisation de ces méthodes souligne la tendance à rechercher de nouvelles approches permettant d'apporter d'autres types de réponses que celles habituellement fournies par les méthodes statistiques.

1.4 Le projet Indices

Comme nous l'avons déjà indiqué, en France, nous utilisons classiquement deux types d'indicateurs pour caractériser les états des masses d'eau. Le premier type consiste en des prélèvements physico-chimiques qui ont l'intérêt de fournir des résultats

précis aux erreurs de mesure près, et d'être une valeur de référence indépendante de tout facteur géographique ou temporel. Cependant la physico-chimie présente deux inconvénients majeurs, d'une part son coût peut être important et, d'autre part, les résultats n'informent de l'état du cours d'eau qu'au moment du prélèvement. La seconde méthode d'évaluation s'appuie sur cinq indices biologiques normalisés qui se basent sur des taxons⁽⁶¹⁾ végétaux ou animaux vivant dans les milieux aquatiques. Ces taxons étant en contact direct avec le milieu, leur présence, leur absence, leur prolifération, leur disparition ainsi que leur état témoignent des pressions subies par le milieu sur toute une période de temps.

Par ailleurs, comparer les indices entre eux serait très intéressant car ils ne témoignent pas des mêmes éléments sur un même site.

L'objectif du projet *Indices* est de comparer les réponses des 5 indices normalisés français (IBGN, IOBS, IPR, IBD et IMBR) et leurs données associées avec les pressions anthropiques pesant sur les masses d'eau de la plaine d'Alsace. L'objectif final est de donner des pistes pour un système d'évaluation de la composante biologique (de l'état écologique) des cours d'eau et de permettre d'identifier les éventuelles pressions anthropiques⁶ subies par ces milieux ou, à l'inverse, de prédire les peuplements⁽³⁸⁾ floristiques et faunistiques attendus à partir des pressions connues sur les masses d'eau.

Ce projet s'intéresse à deux axes, d'une part à la comparaison des indices biologiques et d'autre part, par cette thèse, à la qualification des milieux par les traits de vie : biologiques et écologiques.

Ce projet, démarré en 2005 et soutenu par l'Agence de l'Eau Rhin-Meuse (AERM), est mené en collaboration avec différents organismes apportant une expertise au niveau biologique ou informatique. Les organismes impliqués sur les aspects biologiques, sont l'ONEMA (thème poissons), le Cemagref de Lyon (thème oligochètes) et le Centre de Recherche Gabriel Lippmann du Luxembourg (thème diatomées). Le LORIA de Nancy et le LSIIT d'Illkirch apportent leur contribution sur les aspects informatiques.

1.4.1 Qualification des milieux (alsaciens) à l'aide des traits de vie

A l'heure actuelle, pour comparer les résultats des indices, on utilise les 5 classes. Cependant, ces indices caractérisent des aspects différents de l'environnement, comme la qualité de l'eau ou des sédiments par exemple, qui peuvent avoir des couleurs (qualités) différentes pour le même site. Ainsi, à travers le projet *Indices*, nous cherchons le moyen de pouvoir comparer et combiner ces indices.

Le point commun entre eux est qu'ils appartiennent tous à un environnement. Cette thèse propose donc d'une part de considérer ces indices et plus particulièrement les propriétés morphologiques des taxons qui permettent de les déterminer. En effet, les taxons sont déterminés suivant leurs caractéristiques écologiques, biologiques et phys-

6. Relatif à l'activité humaine.

iologiques qui peuvent être partagées par d'autres taxons même s'ils ne vivent pas au même endroit. Ainsi, nous nous intéressons à rechercher les propriétés morphologiques (appelés *traits biologiques*) et les propriétés de l'environnement (appelées *traits écologiques*) que des taxons peuvent partager afin de comparer les différents indices.

1.4.2 Données considérées pour nos travaux

Le possible coût mais surtout le manque de projection temporelle de la physico-chimie font que le projet *Indices* s'intéresse à la partie biologique. L'approche est de considérer les taxons (servant de base aux indices) par leur caractéristiques morphologiques appelées *traits biologiques*. Dans le cadre de ce projet, cette thèse cherche à généraliser ces indices en identifiant des *traits écologiques* caractéristiques du comportement des espèces dans leur habitat. La mise en relation des traits biologiques et écologiques permettrait de s'abstraire des espèces qui diffèrent d'une région à une autre et empêchent d'uniformiser les évaluations des états écologiques des masses d'eau.

Sélection des espèces

La démarche permettant d'arriver à cet objectif a été mise en œuvre en s'intéressant à un premier indice : l'IBMR. Pour cela nous nous sommes particulièrement intéressés aux travaux de [14] d'une part qui s'attache à la variation des traits biologiques selon la distribution des espèces dans les différents habitats. Et d'autre part à ceux de Willby [69] qui réunit les données bibliographiques de 120 espèces et 14 traits concernant leur dispersion, leur localisation et leur persistance. Ces 14 caractéristiques correspondent à des valeurs potentielles que ces espèces peuvent atteindre.

A partir de cette littérature, Michèle Trémolières, professeure d'écologie à l'Université de Strasbourg, et une spécialiste des macrophytes⁽²⁹⁾, a constitué un jeu de données de 50 plantes. Elle a effectué cette sélection pour ne conserver que les espèces présentes dans la plaine d'Alsace et pour lesquelles nous avons toutes les informations nécessaires. Le tableau 1.3 présente ces espèces par leur nom latin et le code à 4 caractères que nous utilisons [61].

Sélection des traits biologiques

Outre les espèces, nous devons également sélectionner les propriétés de ces plantes. En nous appuyant toujours sur les travaux de [14, 69], nous avons sélectionnés 10 *traits biologiques*. Il s'agit de données bibliographiques sur leur potentialité communément admises par la communauté biologique, et non pas sur des prélèvements effectués sur nos stations alsaciennes.

Ces traits ont été sélectionnés pour correspondre à des caractéristiques morphologiques et physiologiques uniquement, telles que leur taille potentielle ou leur mode de repro-

Nom latin	Code	Nom latin	Code
<i>Alisma plantago-aquatica</i>	ALIP	<i>Nasturtium officinale</i>	NASO
<i>Berula erecta</i>	BERE	<i>Nuphar lutea</i>	NUPL
<i>Callitriche obtusangula</i>	CALO	<i>Nymphaea alba</i>	NYMA
<i>Callitriche platycarpa</i>	CALP	<i>Oenanthe fluviatilis</i>	OENF
<i>Carex acutiformis</i>	CARA	<i>Phalaris arundinacea</i>	PHAA
<i>Ceratophyllum demersum</i>	CERD	<i>Phragmites australis</i>	PHRG
<i>Chara fragilis</i> (=C. globularis)	CHAR	<i>Polygonum hydropiper</i>	PLGH
<i>Chara hispida</i> var. major	CHAH	<i>Potamogeton coloratus</i>	PTCO
<i>Chara vulgaris</i>	CHAV	<i>Potamogeton crispus</i>	PTCR
<i>Eleocharis acicularis</i>	ELEA	<i>Potamogeton lucens</i>	PTLU
<i>Elodea canadensis</i>	ELOC	<i>Potamogeton natans</i>	PTNA
<i>Elodea ernstae</i>	ELOE	<i>Potamogeton nodosus</i>	PTNO
<i>Elodea nuttallii</i>	ELON	<i>Potamogeton pectinatus</i>	PTPE
<i>Groenlandia densa</i>	GROD	<i>Potamogeton pusillus</i>	PTPU
<i>Hippuris vulgaris</i>	HIPV	<i>Ranunculus circinatus</i>	RANC
<i>Hottonia palustris</i>	HOTP	<i>Ranunculus fluitans</i>	RANU
<i>Hydrocharis morsus-ranae</i>	HYDM	<i>Ranunculus trichophyllus</i>	RANT
<i>Iris pseudacorus</i>	IRIP	<i>Rorripa amphibia</i>	RORA
<i>Juncus articulatus</i>	JUNA	<i>Sagittaria sagittifolia</i>	SAGS
<i>Lemna minor</i>	LEMM	<i>Sparganium emersum</i>	SEFC
<i>Lemna trisulca</i>	LEMT	<i>Typha latifolia</i>	TYPL
<i>Mentha aquatica</i>	MENA	<i>Utricularia vulgaris</i>	UTRV
<i>Myosotis palustris</i> (= scorpioides)	MYOP	<i>Veronica anagallis-aquatica</i>	VERA
<i>Myriophyllum spicatum</i>	MYRS	<i>Veronica beccabunga</i>	VERB
<i>Myriophyllum verticillatum</i>	MYRV	<i>Zannichellia palustris</i>	ZANP

TABLE 1.3 – 50 espèces macrophytes retenues.

duction, afin de ne pas mélanger les informations que nous voulons prendre en considération (caractéristiques écologiques). Chacun de ces traits recouvrant un panel trop vaste, ils sont segmentés en plusieurs *modalités* tel qu'indiqué dans le tableau 1.4. Par exemple, la taille potentielle a été scindée en 4 modalités par les experts : < 0.08 mètre, 0.08-0.3 mètre, 0.3 -1 mètre et 1-5 mètres. Au total il existe 35 modalités pour ces 10 traits.

Ajustement des affinités

Chaque plante possède une *affinité* plus ou moins importante pour chacune des modalités des différents traits. Elle est exprimée par une valeur quantitative pouvant valoir :

- 0 pour signifier qu'aucune plante ne possède cette modalité
- 1 pour indiquer qu'il y en a quelques-unes
- 2 pour dire qu'il y en a beaucoup.

Ces données sont issues majoritairement des données bibliographiques. Cependant pour certaines d'entre elles, un peu délicates, nous avons dû faire appel à des connaissances d'experts pour les compléter. A ce jour, 3 plantes : *Chara fragilis*, *Chara hispida* et *Chara vulgaris* viennent juste d'être renseignées pour certains traits écologiques que nous présentons dans la table 3.9 page 80. Ainsi malgré notre choix de 50 plantes, nos travaux ne prennent en compte que 47 d'entre elles.

Les choix effectués sur les plantes, les traits biologiques, leurs modalités et affinités résultent d'ajustements opérés suite à nos travaux validant ou invalidant certains de ces choix, comme nous l'expliquons au fil des chapitres abordés ultérieurement.

1.4.3 Contexte des invertébrés

Nos travaux ont été appliqués initialement sur le jeu de données des macrophytes. Cependant nous les appliquons maintenant sur un autre jeu de données : celui des invertébrés qui sert à la détermination de l'indice IBGN. Ce jeu de données est beaucoup plus important que celui des macrophytes et est issu du livre [2].

Ce contexte contient initialement 472 taxons, 22 traits et 135 modalités. Les valeurs des affinités vont de 0 à 5. Pour chaque taxon sont renseignés le nom du groupe taxonomique, la famille, la sous-famille, le genre et l'espèce. Notons que nous ne parlons plus d'espèces comme avec les macrophytes, justement car l'espèce est un degré taxonomique et que nous ne considérons pas les invertébrés à cet unique degré.

Les taxons

La circulaire de la directive cadre européenne [25] établit le protocole de prélèvement et de traitement des invertébrés pour la mise en œuvre du programme de surveil-

Traits biologiques	Modalités
Taille potentielle	- < 0.08 mètre - 0.08-0.3 mètre - 0.3 -1 mètre - 1-5 mètres
Forme de croissance	- non ancrée - ancrée, feuilles flottantes - ancrée, tissus de soutien - ancrée, submerse
Index morphologique ⁷	valeurs de 0 à 4
Flexibilité	- aucune (< 10°) - faible (10 - 45°) - intermédiaire (45 - 300°) - forte (> 300°)
Pérennité	- organes pérennes (souterrains ou aériens) - bisannuelle - annuelle
Phénologie ⁸	- toute l'année - seulement durant la saison de végétation
Floraison	- oui (annuelle ou pas) - non (rarement ou jamais)
Reproduction végétative	- bulbes ou tubercules - rhizomes ou stolons - bulbilles, turions ou apex dormants - fragments non spécialisés
Période de reproduction	- mars-mai - juin-juillet - août-septembre - octobre
Dispersion potentielle	- élevée (petites graines volantes) - intermédiaire (propagules flottantes, véhiculées par les flux hydriques) - faible (potentialités de transport faibles : graines ou propagules non flottantes, ou de faible longévité, ou absentes)

TABLE 1.4 – Traits biologiques des macrophytes et leurs modalités

lance des cours d'eau. Il s'agit notamment de définir le niveau systématique⁹ à considérer en fonction des différents taxons. Le tableau 1.5 présente ces corrélations. Prenons par exemple la première ligne de ce tableau : "Plecoptera" avec le niveau systématique "Genre". Plecoptera est en fait un *ordre* de la *classe* Insecta. La mention "Genre" stipule que pour tous les individus de l'ordre "Plecoptera" rencontrés nous ne retiendrons que le niveau systématique "genre" auquel il appartient. Nous avons suivi cette circulaire et regroupé les 472 taxons selon le niveau systématique adéquat.

Les affinités profilées

Afin de suivre les recommandations de la directive cadre, nous devons résoudre certaines difficultés concernant la fusion de certains taxons. En effet prenons l'exemple du groupe Hirudinea qui doit être regroupé en famille. Le tableau 1.6 présente les différents niveaux taxonomiques de ce groupe. Il indique également les affinités de chaque modalité du trait "taille maximale potentielle".

Pour le groupe *Hirudinea*, la directive demande que nous ne conservions que les familles, qui sont au nombre de 5 : *Piscicolidae*, *Glossiphoniidae*, *Hirudinidae*, *Haemopidae* et *Erpobdellidae*. Pour les familles ne comprenant qu'une espèce les choses sont simples. Pour les autres familles, nous devons savoir quelles affinités conserver. La première suggestion serait de garder la moyenne pour chaque modalité. Cependant les différentes espèces ne sont pas représentées dans les mêmes proportions. Certaines espèces sont en effet plus communes et fréquentes que d'autres. C'est pourquoi Philippe Usseglio-Polatera, expert des invertébrés, a pondéré les valeurs des affinités en fonction de leur fréquence et abondance.

La mise en forme des données selon la directive cadre entraîne un nombre final de 316 taxons. Notons que l'on parle ici de *taxons* et non plus d'espèces comme avec les macrophytes, parce que l'espèce est un degré taxonomique particulier et que les invertébrés s'appuient sur d'autres degrés taxonomiques : famille, genre, espèce, etc.

Les affinités qui étaient des entiers entre 0 et 5, répartis selon la distribution des individus de chaque taxon entre les modalités de chaque trait, deviennent des pourcentages dont la somme par trait égale 1, comme pour le contexte des macrophytes.

Les taxons pleinement renseignés de la plaine d'Alsace

Parmi les 316 taxons conformes à la circulaire DCE, nous avons rejeté les taxons qui ne sont pas relevés dans la plaine d'Alsace par le projet *Indices*, afin de garder le même protocole que pour les macrophytes et ajuster notre expertise. Ce travail a été mené avec Corinne Grac, Ingénieure de l'Agriculture et de l'Environnement. Nous nous sommes appuyées sur la base de données des prélèvements effectués sur les stations alsaciennes.

9. Classification des êtres vivants.

Taxons	Niveau systématique
Plecoptera	Genre
Ephemeroptera	Genre
Trichoptera (sauf Limnephilidae)	Genre
Trichoptera Limnephilidae	Sous-Famille
Coleoptera (sauf Dytiscidae, Hydrophilidae et Curculionidae)	Genre
Coleoptera (Dytiscidae, Hydrophilidae)	Sous-Famille
Coleoptera Curculionidae	Famille
Megaloptera	Genre
Heteroptera (sauf Corixinae)	Famille
Heteroptera Corixinae	Sous-Famille
Planipennia	Genre
Odonata (sauf Coenagrionidae)	Genre
Odonata Coenagrionidae	Famille
Lepidoptera	Famille
Hymenoptera	Genre
Diptera	Famille
(Hydr)acarina	Présence
Crustacea (sauf Asellidae)	Genre
Crustacea Asellidae	Famille
Bivalvia	Genre
Gastropoda (sauf Planorbidae)	Genre
Gastropoda Planorbidae	Famille
Hirudinea et Branchiobdellida	Famille
Oligochaeta	Classe
Bryozoa	Présence
Nematoda	Présence
Gordiacea	Présence
Turbellaria	Famille
Hydrozoa	Présence
Porifera	Présence
Nemertea	Présence

TABLE 1.5 – Niveaux d'identification requis pour les différents groupes taxonomiques.

Famille	Genre	Espèce	≤.25 cm	>.25-.5 cm	>.5-1 cm	>1-2 cm	>2-4 cm	>4-8 cm	>8 cm
Piscicolidae	Piscicola	geometra	0	0	0	0	2	3	0
Glossiphoniidae	Hemicleipsis	marginata	0	0	0	3	0	0	0
Glossiphoniidae	Glossiphonia	sp.	0	0	0	3	3	0	0
Glossiphoniidae	Helobdella	stagnalis	0	0	3	1	0	0	0
Glossiphoniidae	Placobdella	costata	0	0	0	0	1	3	0
Glossiphoniidae	Glossiphonia	verrucata	0	0	0	0	3	0	0
Glossiphoniidae	Glossiphonia	paludosa	0	0	0	3	0	0	0
Glossiphoniidae	Theromyzon	tessulatum	0	0	0	0	3	0	0
Hirudinidae	Hirudo	medicinalis	0	0	0	0	0	3	1
Haemopidae	Haemopis	sanguisuga	0	0	0	0	0	3	0
Erpobdellidae	Dina	lineata	0	0	0	0	3	0	0
Erpobdellidae	Erpobdella	sp.	0	0	0	0	3	0	0
Erpobdellidae	Trocheta	sp.	0	0	0	0	0	0	3

TABLE 1.6 – Trait *taille potentielle* pour le groupe Hirudinea.

Enfin, nous avons choisi de ne conserver que les taxons pour lesquels nous avons toutes les informations nécessaires. Ces deux opérations ont restreint le contexte à 122 taxons.

Notre contexte est basé sur les informations contenues dans l'ouvrage [2] qui fait référence pour les invertébrés et qui distingue trois catégories de traits : biologiques, écologiques et physiologiques. Après discussion avec l'un des auteurs, Philippe Usseglio-Polatera, nous avons conclu qu'une répartition en deux catégories était plus adéquate : biologiques et écologiques.

Par ailleurs, nous avons révisé les traits et les modalités afin d'ajuster leur pertinence à nos travaux et également diminuer la taille du jeu de données. Nous présentons dans cette section les ajustements opérés.

Ajustements des traits et des modalités

Pour certains traits, certaines modalités ont été fusionnées ou enlevées car elles étaient trop peu souvent renseignées ou rarement rencontrées. Pour chaque trait concerné, listons ces modifications.

- Reproduction : la parthénogenèse (reproduction monoparentale à partir d'un individu femelle) a été enlevée car il y avait peu d'information et elle a été agrégée à la reproduction asexuée.
- Forme de résistance : "cocons" et "habitat contre le dessèchement" : les cocons étant des habitats contre le dessèchement, ils ont donc été fusionnés.

- Respiration : la modalité "vésicules hydrostatiques" est peu utilisée et concerne des organismes d'eau stagnante ce qui est peu intéressant dans l'étude des rivières.
- Alimentation : "micro-organismes" et "détritiques" ont été regroupés, car l'invertébré ne fait pas de distinction entre l'un et l'autre quand il absorbe sa nourriture.
- Méthode d'alimentation : "absorption à travers les téguments⁽⁶³⁾" concerne très peu de taxons et a donc été enlevé.
- Distribution transversale : "habitat souterrain" a été enlevé car il s'agissait d'une information difficilement codable car il est difficile de savoir si l'individu habite vraiment là.
- Régions bio-géographiques : le "Jura" a été déplacé des "Alpes" aux "Vosges et Massif Central". En effet il appartient à la zone bio-géographique numéro 8 définie dans Limnofauna Europea d'Illies.

Les traits ayant beaucoup de modalités impliquent que leur affinités vont avoir une variance plus importante. Pour contrer cet inconvénient, nous avons regroupé certaines modalités :

- Taille potentielle maximale : " ≤ 0.25 cm" a été regroupée avec "> 0.25 – 5 cm" et "> 4 – 8 cm" a été regroupée avec "> 8 cm".
- Reproduction : "ponte dans la végétation" et "ponte terrestre" ont été regroupées car ce sont deux façons de protéger les œufs de la prédation.
- Locomotion : "vol" est supprimé car il ne concerne que les insectes. "Nage de surface" et "nage en pleine eau" sont regroupées en "nage".
- Alimentation : "macro-invertébrés vivants" et "vertébrés" sont également regroupés.
- Saprobie⁽⁵²⁾ : "alpha-mésosaprobe" et "polysaprobe" sont aussi regroupés.

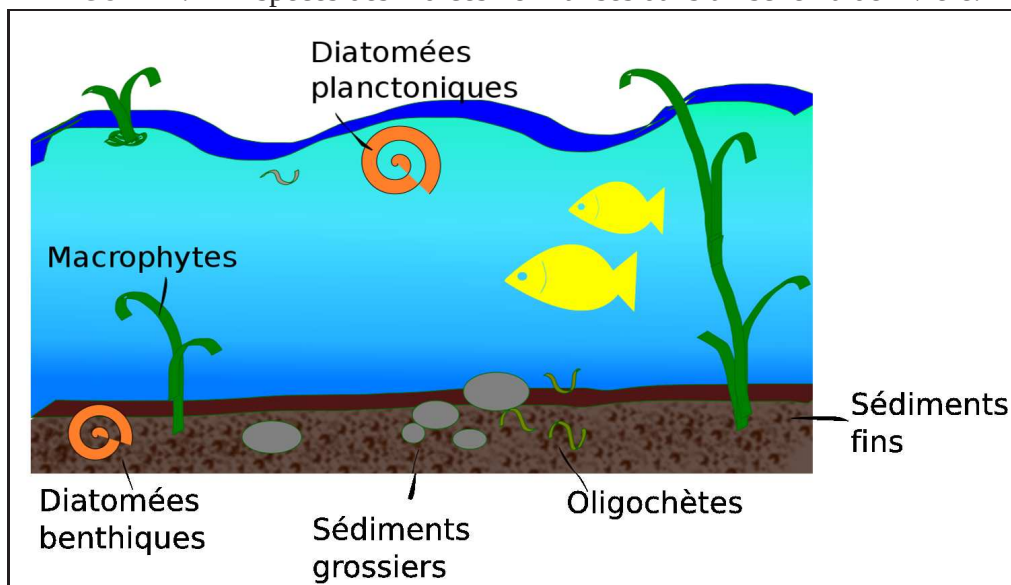
Conclusion

Il existe cinq indices biologiques basés sur des organismes vivants : poissons, diatomées, invertébrés⁽²⁸⁾, oligochètes et macrophytes⁽²⁹⁾. Leur localisation dans la rivière est présentée dans la figure 1.4. Ces indices permettent de qualifier des portions sélectionnées des masses d'eau en fonction du lieu géographique où ces organismes sont présents. Ceci implique que les différentes masses d'eau ne sont pas évaluées avec les mêmes organismes, et donc pas de manière homogène. Le projet *Indices* vise à comparer les indices biologiques qui sont utilisés pour qualifier l'état des masses d'eau.

Afin de comparer ces indices, diverses méthodes (certaines informatiques, mais principalement statistiques) sont utilisées par les biologistes. Cependant ils restent difficilement comparables car ils sont basés sur des expertises compartimentées et régionales. En effet, ces outils biologiques sont basés sur des espèces (plantes, poissons...) qui vivent dans des zones différentes, ce qui entrave la comparaison des outils existants.

L'hypothèse biologique pour parer à ce problème est de déterminer des traits écologiques. Ceux-ci caractériseraient les plantes, invertébrés... en fonction de leur milieu et perme-

FIGURE 1.4 – Espèces des indices normalisés dans un schéma de rivière.



traient donc d'établir des liens au sein d'un même indice mais également entre les indices.

Les traits écologiques sont liés aux espèces et à leurs traits biologiques c'est pourquoi nous avons choisi d'explorer ces données par des méthodes d'Analyse Formelle de Concepts (AFC) [9, 24, 37] définies dans le chapitre 2. Nous avons choisi d'utiliser les treillis de Galois parce qu'ils sont des outils pertinents pour l'extraction de connaissances [55] et permettent d'interagir avec les experts. Plus particulièrement, ils permettent de trouver des concepts, ici des ensembles de traits biologiques partagés par des espèces, qui peuvent être interprétés selon leur écosystème⁽¹⁴⁾ et ainsi déterminer des *traits écologiques* à partir des espèces et de leurs traits biologiques tel que présenté dans la section 3.4. Ces traits écologiques pourront ensuite être associés aux traits biologiques comme cela a été mis en application sur les macrophytes et les invertébrés dans le chapitre 5.

Chapitre 2

Treillis de Galois et Analyse de Concepts Formels

Sommaire

2.1	Ensembles ordonnés	32
2.2	Treillis et propriétés des treillis	34
2.2.1	Demi-treillis, treillis et treillis complet	34
2.2.2	Approche algébrique	34
2.2.3	Treillis de Galois	35
2.2.4	Analyse de Concepts Formels	36
2.3	Treillis sur des contextes non triviaux	38
2.3.1	Contexte multi-valué	38
2.3.2	Structures de patrons	40
2.3.3	Contexte flou	42
2.3.4	Objets symboliques	43
2.3.5	Connexion de Galois par similarité	44
2.3.6	Règles d’association et implications	45
2.4	Conclusion	46

Ce chapitre présente tout d’abord les définitions nécessaires à la compréhension des treillis de Galois et l’analyse formelle de concepts. Puis nous introduisons les notions de contexte et concept formels, et élargissons le spectre des définitions au traitement de contextes complexes.

Enfin, nous présentons des utilisations de treillis en fouille de données. La bibliographie est basée sur les ouvrages [9, 24, 37, 18, 20], mais dans la mesure du possible afin de conserver une cohérence de notation, nous nous appuyons sur le formalisme du livre [37] de Ganter et Wille.

2.1 Ensembles ordonnés

Nous rappelons ici quelques définitions générales concernant les ensembles ordonnés sur lesquelles s'appuie à la théorie des treillis.

Définition 1 Relation binaire. Une relation binaire R entre deux ensembles O et T est un ensemble de couples (o, t) avec $o \in O$ et $t \in T$, i.e. un sous-ensemble de l'ensemble $O \times T$. On écrit oRt pour signifier $(o, t) \in R$, et R^{-1} est la relation inverse de R entre T et O avec $tR^{-1}o \Leftrightarrow oRt$.

Définition 2 Relation d'ordre. Une relation binaire R sur $T \times T$ est appelée une relation d'ordre (ou simplement **ordre**) si elle satisfait les trois conditions suivantes pour tous les éléments $x, y, z \in T$:

- | | |
|-------------------------------------|----------------|
| 1. xRx | (réflexivité) |
| 2. xRy et $yRx \Rightarrow x = y$ | (antisymétrie) |
| 3. xRy et $yRz \Rightarrow xRz$ | (transitivité) |

Dans le cas d'une relation d'ordre, on utilise souvent le symbole $\leq$ pour R (et respectivement $\geq$ pour R^{-1}), et on note $x \leq y$ pour signifier que "x est inférieur ou égal à y". On note $x < y$ pour signifier que "x est strictement inférieur à y".

On dit qu'un **ordre** est **total** si les éléments d'un ensemble sont tous comparables deux à deux :

$$\forall x, y \quad x \neq y \rightarrow x < y \text{ ou } y < x$$

Un ordre qui n'est pas total est dit **partiel**.

Définition 3 Ensemble ordonné. Un ensemble ordonné est un couple $(O, \leq)$ avec O un ensemble et $\leq$ une relation d'ordre sur O .

Définition 4 Majorant, minorant. Soit O un ensemble ordonné par une relation $\leq$, $y \in O$ est un **majorant** (dualement un **minorant**) de $x \in O$ si et seulement si $x \leq y$. On dit alors que "x minore y" ou que "y majore x".

L'ensemble des majorants d'un élément a donné peut se noter **Maj.(a)** (dualement : **Min.(a)**).

Définition 5 Couverture. Soit O un ensemble ordonné et $x, y \in O$, on dit que x est couvert par y (ou que y couvre x) si $x < y$ et s'il n'existe pas de $z \in O$ tel que $x < z$ et $z < y$.

On note alors $x \prec y$ et x est appelé prédécesseur de y (et respectivement y est appelé successeur de x).

Diagramme de Hasse

Un diagramme de Hasse est une représentation visuelle d'un ordre fini. Cette représentation utilise des points pour représenter les éléments de O et des segments qui les interconnectent pour indiquer les relations de couverture. Les segments caractérisent des relations de spécialisation / généralisation entre les éléments de O .

Principe de dualité des ensembles ordonnés

Etant donné un ensemble ordonné $(O, \leq)$, on peut construire un nouvel ensemble ordonné $(O, \geq)$ (son **dual**) en définissant la relation inverse " $\geq$ " de " $\leq$ " comme étant aussi une relation d'ordre sur O qui sera alors appelée **duale** [9].

Le diagramme de Hasse correspondant à $(O, \geq)$ peut être aisément obtenu en renversant tête en bas le diagramme de $(O, \leq)$.

Définition 6 Infimum. Soient $(O, \leq)$ un ordre partiel, A un sous-ensemble de O , et m un élément quelconque de O . m est un infimum de A (noté $\bigwedge A$) si et seulement si :

$$\begin{aligned} &\forall a \in A, (m \leq a) \\ &\text{et } \forall c \in O, (\forall a \in A, c \leq a) \rightarrow (c \leq m). \end{aligned}$$

Synonymes d'infimum : plus grand commun diviseur, borne inférieure, greatest lower bound en anglais (glb).

Définition 7 . Supremum. Soient $(O, \leq)$ un ordre partiel, A un sous-ensemble de O , et M un élément quelconque de O . M est un supremum de A (noté $\bigvee A$) si et seulement si :

$$\begin{aligned} &\forall a \in A, (a \leq M) \\ &\text{et } \forall c \in O, (\forall a \in A, a \leq c) \rightarrow (M \leq c). \end{aligned}$$

Synonymes de supremum : plus petit commun multiple, borne supérieure, lowest upper bound en anglais (lub).

Définition 8 Minorant et majorant universels. Soit O un ensemble ordonné fini, il peut avoir un élément appelé **minorant universel** noté $\perp$, ayant pour propriété que $\perp \leq x$ pour tout $x \in O$. De manière duale il peut aussi avoir un élément appelé **majorant universel** et noté $\top$ qui vérifie $\top \geq x$ pour tout $x \in O$. Ces minorant et majorant universels sont chacun unique quand ils existent.

Définition 9 Maximum, minimum. Un élément x est **maximum** (dualement **minimum**) d'une partie X d'un ensemble ordonné si et seulement si :

- $x \in X$
- $\forall y \in X, y \leq x$.

En particulier si un ensemble ordonné admet un majorant universel, celui-ci est maximum dans cet ensemble¹⁰ [9].

2.2 Treillis et propriétés des treillis

Les treillis ont été étudiés en parallèle aux Etats Unis, en France et en Allemagne. Il existe deux écoles, qui utilisent des vocabulaires différents. Nous définissons dans cette section les **treillis de concepts** selon l'approche allemande de Ganter et Wille [37] et les **treillis de fermés** autrement appelés **treillis de Galois** selon l'approche française de Barbut et Monjardet [9]. Dans ce document, nous utiliserons de préférence la notation allemande.

2.2.1 Demi-treillis, treillis et treillis complet

Avant d'aborder la définition des treillis nous précisons les notations sur lesquelles elle s'appuie : celles des demi-treillis.

Définition 10 Inf-demi-treillis. Un inf-demi-treillis est un ensemble muni d'une relation d'ordre tel que chaque couple (x, y) d'éléments admet un infimum de x et de y , noté $x \frown y$ ¹¹.

Définition 11 Sup-demi-treillis. Un sup-demi-treillis est un ensemble muni d'une relation d'ordre tel que chaque couple (x, y) d'éléments admet un supremum de x et de y , noté $x \smile y$.

Définition 12 Treillis, treillis complet. Soit O un ensemble ordonné non vide. O est un **treillis** si chaque couple (x, y) d'éléments de O possède un supremum et un infimum. Il est à la fois un inf-demi-treillis et un sup-demi-treillis. O est un **treillis complet** si $\bigwedge S$ et $\bigvee S$ existent pour tout $S \subseteq O$.

2.2.2 Approche algébrique

Propriétés

Un treillis se définit également comme une structure algébrique $(O, \wedge, \vee, \leq)$. Il s'agit alors d'un ensemble O muni de deux lois de composition interne $\wedge$ et $\vee$ vérifiant :

10. Il convient de distinguer maximum et maximal (dualement minimum et minimal) : dans un ordre total, un maximum majore tous les éléments de X , alors qu'un élément maximal n'est majoré par aucun autre élément de X mais ne majore pas forcément tous les éléments de X .

11. On distingue l'infimum $\bigwedge$ d'un ensemble et l'infimum $\frown$ d'un couple d'éléments, et respectivement pour le supremum.

- l’associativité : $\forall x, y, z \in O, (x \vee y) \vee z = x \vee (y \vee z)$ (resp. pour $\wedge$)
- la commutativité : $\forall x, y \in O, x \vee y = y \vee x$ (resp. pour $\wedge$)
- l’idempotence : $\forall x \in O, x \vee x = x$ et $x \wedge x = x$
- l’absorption : $\forall x, y \in O, x \wedge (x \vee y) = x$ et $x \vee (x \wedge y) = x$

Le principe de dualité des ensembles ordonnés s’applique aux treillis, et on obtient le dual d’un ordre en remplaçant chacun des symboles : $\leq, \wedge, \vee, \dots$ par $\geq, \vee, \wedge, \dots$

Eléments particuliers

Dans un treillis, il existe un ensemble d’éléments par lesquels on peut représenter l’intégralité des éléments de ce treillis : les éléments inf et sup-irréductibles.

Définition 13 Élément sup-irréductible. Considérons le treillis T et $x \in T$ un élément de ce treillis. On dit que x est un élément **sup-irréductible** (aussi noté $\vee$ -irréductible) si et seulement s’il a un et un seul prédécesseur x_1 . Autrement dit le supremum de lui-même et de ce prédécesseur $x = x \vee x_1$. On note l’ensemble de ces sup-irréductibles $J(T)$ ¹².

Définition 14 Élément inf-irréductible. On définit de manière duale les éléments **inf-irréductibles** ($\wedge$ -irréductibles) comme étant ceux n’ayant qu’un et un seul successeur. Un élément x inf-irréductible est l’infimum de lui-même et de ce successeur : $x = x \wedge x_1$. L’ensemble de ces éléments se note $M(T)$ ¹³.

Définition 15 Atomes et coatomes. Dans un treillis avec minorant universel, les successeurs de $\perp$ sont toujours $\vee$ -irréductibles et sont appelés **atomes**. Par dualité, les éléments prédécesseurs de $\top$ sont toujours $\wedge$ -irréductibles quand $\top$ existe et sont appelés **coatomes**.

2.2.3 Treillis de Galois

Définition 16 Fermeture [9] Une fermeture dans un ensemble ordonné $(O, \geq)$ est une isotonie $h : O \rightarrow O$ extensive et idempotente, c’est-à-dire, une application ayant les trois propriétés :

1. $\forall x, \forall y, x \geq y \Rightarrow h(x) \geq h(y)$ (monotone croissante)
2. $\forall x, h(x) \geq x$ (extensive)
3. $\forall x, h \circ h(x) = h(x)$ (idempotente)

12. Pour Join : terme anglais de l’union.

13. Pour Meet : terme anglais de l’intersection.

Un élément x de O est dit **fermé** s'il est égal à sa fermeture : $x = h(x)$

Soient O et T deux ensembles ordonnés quelconques, et $\uparrow: O \rightarrow T$ et $\downarrow: T \rightarrow O$ deux applications monotones décroissantes :

$$\begin{aligned} \forall x, \forall y \in O, x \leq y &\Rightarrow x^\uparrow \geq y^\uparrow \\ \forall x', \forall y' \in T, x' \leq y' &\Rightarrow x'^\downarrow \geq y'^\downarrow \end{aligned}$$

Considérons les applications composées $h = \downarrow\uparrow$ ($\uparrow$ suivie de $\downarrow$) de O dans O , et $h' = \uparrow\downarrow$ de T dans T . Ces deux applications sont monotones croissantes. Si en outre elles sont extensives alors elles sont idempotentes et sont donc des fermetures dans O et T respectivement.

Définition 17 Connexion et fermeture de Galois. Le couple $(\uparrow, \downarrow)$ d'applications vérifiant les propriétés précédentes (monotone croissante, extensive et idempotente) est une **correspondance de Galois** entre les ensembles ordonnés O et T . Les fermetures $h = \downarrow\uparrow$ dans O et $h' = \uparrow\downarrow$ dans T associées à une correspondance de Galois $(\uparrow, \downarrow)$ sont appelées **fermetures de Galois** [9].

2.2.4 Analyse de Concepts Formels

L'Analyse de Concepts Formels (ACF) dérive des mathématiques appliquées et se base sur les travaux de Birkhoff [13] en 1940 (révisés en 1948 puis 1967). Sur ces travaux s'appuient ensuite ceux de Ganter et Wille en 1999 [37].

Nous présentons ici les définitions de contextes et concepts formels. Nous nous appuyons sur des exemples pour illustrer ces définitions.

Contexte formel

Un contexte formel traduit une relation binaire entre un ensemble de propriétés et un ensemble d'objets.

Définition 18 Un contexte formel $K = (O, T, I)$ est composé de deux ensembles O et T et d'une relation binaire I entre O et T . Les éléments de O sont appelés des **objets** et les éléments de T sont appelés des **attributs**¹⁴. Pour exprimer qu'un objet o est en relation I avec un attribut t , on écrit oIt ou $(o, t) \in I$, ce qui se lit "l'objet $o \in O$ possède l'attribut $t \in T$ ".

Définition 19 Concept formel. Soit la connexion de Galois suivante définie sur le contexte (O, T, I) :

$$\uparrow: 2^O \rightarrow 2^T$$

14. Dans le cadre de l'application biologique, les attributs portent le nom de **traits**.

$$A \subseteq O \rightarrow A^\uparrow = \{t \in T \mid \forall o \in A, oIt\}$$

$$\downarrow: 2^T \rightarrow 2^O$$

$$B \subseteq T \rightarrow B^\downarrow = \{o \in O \mid \forall t \in B, oIt\}$$

Avec $A \subseteq O$ et $B \subseteq T$. Les couples $(A, B) \in 2^O \times 2^T$ d'ensembles vérifiant $A = B^\downarrow$ et $B = A^\uparrow$ sont appelés des **concepts formels** (ou simplement **concepts**). A est alors appelé l'**extension** du concept et B son **intension**. L'**extension** est l'ensemble de tous les objets partageant les attributs de l'intension. L'**intension** est l'ensemble de tous les attributs partagés par tous les objets de l'extension.

Le sous-ensemble A de O est l'extension du concept unique $(A, A^\uparrow)$ si et seulement si $A^{\uparrow\downarrow} = A$. On définit dualement les sous-ensembles B de T en tant qu'intensions. L'ensemble de tous les concepts du contexte (O, T, I) est noté par $\mathfrak{B}(O, T, I)$.

Considérons le contexte de la table 2.1. L'ensemble des objets est $O = \{\text{Alien, Men in black, Stargate, Une nuit en enfer, Scary movie}\}$, l'ensemble des propriétés est $T = \{\text{US, FR, GB}\}$ et la relation I associe à chaque film sa nationalité par une croix indiquant que l'objet concerné possède la propriété concernée. Par exemple, le concept $(\{\text{Alien, 28 jours plus tard}\}, \{\text{GB}\})$ issu de ce contexte indique que ces deux films (et aucun autre) ont comme seule caractéristique commune d'être britanniques.

Film	US	FR	GB
La cité de la peur		×	
Alien			×
Men in black	×		
Stargate	×		
Une nuit en enfer	×		
Scary movie	×		
28 jours plus tard			×

TABLE 2.1 – Contexte filmographique.

Ordre sur les concepts

Définition 20 Soit (O, T, I) un contexte. Soient $(A_1, B_1) \in \mathfrak{B}(O, T, I)$ et $(A_2, B_2) \in \mathfrak{B}(O, T, I)$, on écrit $(A_1, B_1) \leq (A_2, B_2)$, si et seulement si $A_1 \subseteq A_2$.

$A_1 \subseteq A_2$ implique $A_1^\uparrow \supseteq A_2^\uparrow$, et inversement parce que $A_1^{\uparrow\downarrow} = A_1$ et $A_2^{\uparrow\downarrow} = A_2$. Ainsi nous avons $(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_1 \supseteq B_2$.

(A_1, B_1) est alors appelé **sous-concept** de (A_2, B_2) qui, lui, est son **super-concept**.

Définition 21 Treillis de concepts. L'ensemble $\mathfrak{B}(O, T, I)$ muni de l'ordre $\leq$ définit un treillis dit treillis de concepts et se note $\langle \mathfrak{B}(O, T, I); \leq \rangle$.

Représentation des treillis et héritage

Un treillis est représenté par un diagramme de Hasse où chaque nœud est un concept, tel que le montre la figure 2.1. Il peut être représenté de manière simplifiée, c'est-à-dire que les individus et propriétés indiqués sont ceux qui sont propres au concept et non pas ceux hérités. En effet, l'ordre permet également de définir un héritage entre les concepts. Ainsi chaque concept hérite des objets des concepts qui sont entre lui et $\perp$ (ses sous-concepts) et des attributs qui sont entre lui et $\top$ (ses super-concepts). C'est pour cela qu'un treillis est dit être une **double hiérarchie** de concepts. Nous appellerons donc cet ordre un **ordre hiérarchique**.

2.3 Treillis sur des contextes non triviaux

2.3.1 Contexte multi-valué

La notion de contexte formel a été étendue par Ganter et Wille [36] pour introduire le fait qu'un attribut puisse avoir plusieurs valeurs comme l'illustre l'exemple présenté dans la table 2.2.

Définition 22 Un contexte multi-valué K est défini par un quadruplet (O, T, M, I) , où O est un ensemble d'**objets**, T est un ensemble d'attributs multi-valués appelés **traits**, M est un ensemble de valeurs pour ces attributs, valeurs appelées **modalités**. I est une relation ternaire $I \subseteq O \times T \times M$ telle que : $(o, t, m) \in I$ et $(o, t, n) \in I$ implique toujours $m = n$. La notation $(o, t, m) \in I$ (ou $t(o) = m$) signifie que l'attribut t possède la valeur m pour l'objet o .

Film	Catégorie	Nationalité	Durée
La cité de la peur	Comédie	FR	1h30
Alien	Fant./SF	GB	1h30
Men in Black	Fantastique/SF	US	1h45
Stargate	Fantastique/SF	US	2h
Une nuit en enfer	Comédie	US	1h45
Scary movie	Comédie	US	1h30
28 jours plus tard	Horreur	GB	2h

TABLE 2.2 – Contexte filmographique.

Prenons un exemple nous permettant d'illustrer ces définitions : un contexte filmographique présenté par la table 2.2. L'ensemble des objets est composé des différents films représentés par leur titre. L'ensemble des attributs est $T = \{\text{Catégorie, Nationalité, Durée}\}$ et l'ensemble des modalités de ces attributs est $M = \{\text{Horreur, Comédie, Fantastique/SF, US, FR, GB}^{15}, 1\text{h30}, 1\text{h45}, 2\text{h}\}$. Ainsi $(\text{Men in Black, Nationalité, US}) \in I$ signifie que le film "Men in Black" est américain.

Ces contextes ne peuvent être traités directement dans le cadre de l'ACF. Des pré-traitements sont nécessaires.

Echelonnage conceptuel

Il s'agit d'un processus où chaque attribut multi-valué est interprété au moyen d'une **échelle conceptuelle** permettant de binariser un contexte multi-valué.

Définition 23 Une **échelle conceptuelle** (ou **échelle**) pour un attribut t d'un contexte multi-valué est un contexte (mono-valué) $S_t := (R_t, T_t, I_t)$.

Illustrons ce processus par l'exemple filmographique. Considérons l'attribut multi-valué "Durée". Il existe plusieurs échelles selon l'interprétation que l'on souhaite considérer. Par exemple nous pouvons établir pour une échelle de taille que tout ce qui mesure plus de 10 cm mesure aussi plus de 5 cm. Dans notre cas nous nous intéressons aux échelles conceptuelles nominales, c'est-à-dire que les différentes modalités s'excluent mutuellement comme nous pouvons le voir dans la table 2.3 pour le trait multi-valué "Durée". Ce contexte présente les ensembles $M_{\text{Durée}} = \{1\text{h30}, 1\text{h45}, 2\text{h}\}$ et $R_{\text{Durée}} = \{\text{Durée-1h30}, \text{Durée-1h45}, \text{Durée-2h}\}$

Film	1h30	1h45	2h
Durée-1h30	×		
Durée-1h45		×	
Durée-2h			×

TABLE 2.3 – Echelle conceptuelle nominale (contexte) du trait "Durée".

Après avoir défini une échelle pour chaque attribut multi-valué (dont les modalités peuvent être multiples comme par exemple : FR, GB et US), la seconde étape du processus pour obtenir un contexte mono-valué (dont les modalités sont binaires : présence / absence) consiste à regrouper dans un même contexte les échelles de chaque attribut.

En regroupant les différentes échelles (et les éventuels attributs mono-valués) nous obtenons un contexte complet mono-valué tel que présenté dans la table 2.4. Plus formellement on pose la définition suivante [37] :

15. SF : Science Fiction, US : américain, FR : français, GB : britannique.

Définition 24 Contexte mono-valué. Si (O, T, M, I) est un contexte multi-valué et $S_t = (R_t, M_t, I_t), t \in T$ un ensemble d'échelles, alors le contexte qui en découle compte tenu de l'échelonnage conceptuel est le contexte (O, R, J) avec :

$$R := \bigcup_{t \in T} T_t \times M_t$$

et

$$oJr_t \Leftrightarrow t(o) = m \text{ et } mI_tr_t$$

Film	Catégorie			Nationalité			Durée		
	Horreur	Comédie	Fantastique/SF	US	FR	GB	1h30	1h45	2h
La cité de la peur		×			×		×		
Alien			×			×	×		
Men in Black			×	×				×	
Stargate			×	×					×
Une nuit en enfer		×		×				×	
Scary movie		×		×			×		
28 jours plus tard	×					×			×

TABLE 2.4 – Contexte filmographique multi-valué.

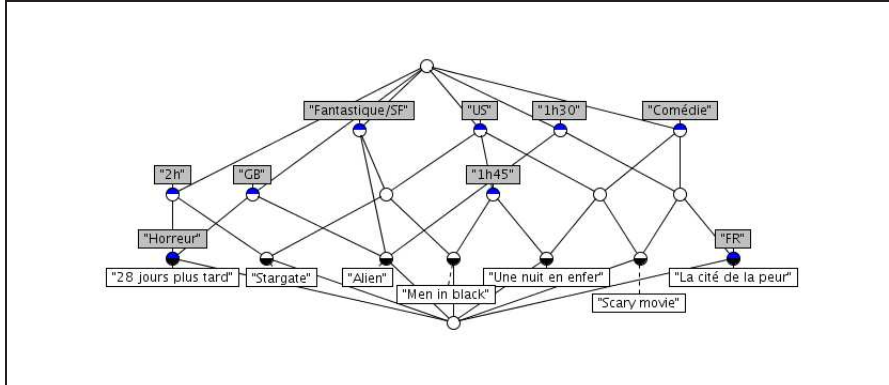
La figure 2.1 présente le treillis de Galois correspondant au contexte filmographique binarisé de la table 2.4 obtenu par le logiciel Conexp [70]. Les étiquettes grises concernent les attributs et les blanches concernent les objets.

Par exemple si on considère le concept avec l'étiquette "1h45" dans la figure 2.1, ce concept se lit ainsi : $(\{\text{Men in black, Une nuit en enfer}\}, \{1h45, \text{US}\})$ puisqu'il hérite de l'attribut "US". Comme indiqué par la définition 20, ce concept est bien le *sous-concept* de celui juste au dessus dans le treillis de la figure 2.2 et portant l'étiquette "US" : $(\{\text{Stargate, Men in black, Une nuit en enfer, Scary movie}\}, \{\text{US}\})$. En effet, ce concept hérite des objets de ses sous-concepts : "Stargate", "Men in black", "Une nuit en enfer" et "Scary movie".

2.3.2 Structures de patrons

Nous abordons ici une autre représentation des données, les structures de patrons, appelées en anglais : "**pattern structures**" et présentée par [35]. Il s'agit d'une représentation générale pour les échelonnages conceptuels. Les objets possèdent alors des descriptions ou "patrons" ("*patterns*" en anglais) qui sont structurées dans un demi-treillis et permettent des opérations latticielles.

FIGURE 2.1 – Treillis du contexte binaire filmographique de la table 2.4.



Définition 25 Structure de patrons. Soit G un ensemble, $(D, \sqcap)$ un inf-demi-treillis et $\delta : G \rightarrow D$ une application. Alors $(G, \underline{D}, \delta)$ avec $\underline{D} = (D, \sqcap)$ est appelé une **structure de patrons** impliquant que l'ensemble

$$\delta(G) := \{\delta(g) \mid g \in G\}$$

génère un sous-demi-treillis $(D_\delta, \sqcap)$ de $(D, \sqcap)$ ¹⁶.

Les éléments de D sont appelés des **patrons**. L'ordre sur D , appelé **ordre de subsomption** est donné par :

$$c \sqsubseteq d \Leftrightarrow c \sqcap d = c$$

Une structure de patrons s'interprète comme étant un ensemble d'objets avec des "descriptions" (patrons) acceptant entre elles une "opération de similarité" $\sqcap$, i.e. une opération telle que pour un ensemble quelconque d'objets, elle produise une "description" qui représente les similitudes entre les objets du sous-ensemble. La similarité est indépendante de l'ordre dans lequel les objets apparaissent, c'est une opération idempotente, commutative et associative.

Définition 26 Opérateurs de fermeture. Si $(G, \underline{D}, \delta)$ est une structure de patrons, alors on définit les opérations de fermeture :

$$A^\square := \sqcap_{g \in A} \delta(g) \quad \text{pour } A \subseteq G$$

$$d^\square := \{g \in G \mid d \sqsubseteq \delta(g)\} \quad \text{pour } d \in D$$

16. En considérant que chaque sous-ensemble X de $\delta(G)$ possède un infimum $\sqcap X$ de $(D, \sqcap)$ et que D_δ est l'ensemble de ces infima.

Les opérateurs $(.)^\square$ forment une connexion de Galois entre l'ensemble des parties de G et $(D, \sqsubseteq)$. Les couples (A, d) satisfaisant

$$A \subseteq G, \quad d \in D, \quad A^\square = d \quad \text{et} \quad A = d^\square$$

sont appelés des **concepts patrons** (en anglais : **pattern concepts**) de $(G, \underline{D}, \delta)$, avec A comme extension et d comme **patron d'intension**.

2.3.3 Contexte flou

La notion de **contexte flou** a été introduite dans [71] pour présenter un contexte où tout rattachement d'un attribut à un objet serait une question de degré d'appartenance. Elle a été étendue dans [15, 17]. Des données floues permettent de représenter une relation floue entre un objet et un trait, ou, autrement dit : l'affinité d'un objet pour un trait.

Définition 27 Soit A un ensemble de **degrés de vérité** ou **degrés d'appartenance**. Un *contexte flou* est un triplet $K := (O, T, I)$ où O et T sont respectivement des ensembles d'objets et d'attributs, et $I : O \times T \rightarrow A$ est une relation floue entre O et T . Un degré $I(o, t) \in A$ est interprété comme un degré d'appartenance de l'attribut t pour l'objet o .

La table 2.5 présente un contexte flou de notre exemple. $I(\text{Alien}, \text{Horreur}) \in 40\%$ signifie que le film Alien est à 40% un film d'horreur¹⁷.

Il est à noter que la définition d'un contexte formel est un cas particulier de la définition d'un contexte flou dans le sens où A est réduit à l'ensemble binaire $\{0,1\}$.

<i>Film</i> ¹⁸	<i>Horreur</i>	<i>Comédie</i>	<i>Fant./SF</i>	<i>US</i>	<i>FR</i>	<i>GB</i>	<i>1h30</i>	<i>1h45</i>	<i>2h</i>
La cité de la peur	10%	90%			100%		40%	60%	
Alien	40%		60%	50%		50%	60%	40%	
Men in Black	5%	40%	55%	100%			47%	53%	
Stargate			100%	50%	50%				100%
Une nuit en enfer	30%	40%	30%	100%				87%	13%
Scary movie	40%	60%		100%			100%		
28 jours plus tard	90%		10%			100%		53%	47%

TABLE 2.5 – Contexte filmographique flou.

17. Ces pourcentages n'ont qu'un caractère illustratif.

2.3.4 Objets symboliques

Objets symboliques

L'analyse de données symboliques [56] a pour but d'étendre les méthodes d'analyse de données classiques à des données de type multi-valué, intervalle ou histogramme selon le formalisme des *objets symboliques* qui a pour objectif d'uniformiser.

Définition 28 *Objet symbolique.* Soit Ω un ensemble des individus w . Un objet symbolique est un triplet (aRd) où R est un vecteur de relations R_i , d est un vecteur de descriptions d_i et a une application de Ω dans $\{\text{vrai}, \text{faux}\}$.

La description d'un individu w est représentée par le vecteur $(y_1(w) \dots y_p(w))$ où chaque variable y_i , $i = 1 \dots p$ est une application de Ω dans $P(O_i)$. O_i est l'espace de description de la variable y_i : des valeurs qualitatives (nominales ou ordinales) ou quantitatives. D est l'ensemble des $y_i(w)$. La figure 2.2 représente les relations qui lient les trois ensembles Ω , D et A . Un objet symbolique peut être vu comme un concept. La fonction f permet de trouver l'objet symbolique correspondant à une description en lui associant entre autre une relation pour chaque variable. La fonction g permet de calculer l'extension d'un objet symbolique.

La hiérarchie entre les concepts suit un processus de généralisation et spécialisation. G. Polaillon présente les connexions de Galois issues des deux méthodes de généralisation : par union et par intersection. Soit A l'ensemble de toutes les assertions possibles. Les applications suivantes sont définies :

$$\begin{aligned} f : P(\Omega) &\rightarrow A \\ C &\rightarrow a(\omega) = \bigwedge_j [y_j(\omega) R \cup_{\{\omega' \in C\}} \{y_j(\omega')\}] \end{aligned}$$

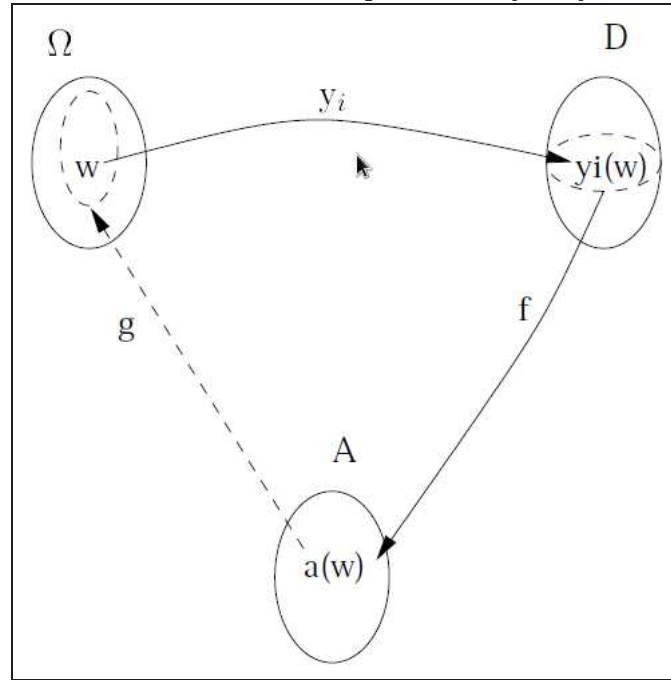
$$\begin{aligned} g : A &\rightarrow P(\Omega) \\ a(\omega) &= \bigwedge_j [y_j(\omega) R W_j] \rightarrow \text{ext}_\Omega(a) = \{\omega / y_j(\omega) R W_j, j = 1 \dots p\} \end{aligned}$$

Treillis d'union et d'intersection

G. Polaillon, définit également un ordre sur les objets symboliques permettant la construction de treillis de Galois. Deux types de treillis de Galois peuvent être obtenus : dans un premier cas il s'agit de classes (ou groupes) dont les attributs sont communs aux individus de la classe, dans l'autre, les attributs correspondent au moins à un des individus de la classe. Donnons la définition de Diday [26] :

Définition 29 *Union et intersection "symbolique" d'objets.* Soient $s_1 = (a_1 R d_1)$ et $s_2 = (a_2 R d_2)$ deux objets symboliques. L'union symbolique de s_1 et s_2 est notée $a_1 \cup a_2$ (resp. l'intersection symbolique est notée $a_1 \cap a_2$) et est définie comme la conjonction de tous les objets symboliques dont l'extension sur Ω contient $\text{ext}(a_1) \cup \text{ext}(a_2)$ (resp. $\text{ext}(a_1) \cap \text{ext}(a_2)$).

FIGURE 2.2 – Individus, descriptions et objets symboliques



2.3.5 Connexion de Galois par similarité

Dans sa thèse [52], N. Messai s'intéresse à la découverte de sources de données biologiques sur le web en associant des sources pertinentes à des besoins préalablement exprimés. Il travaille sur un annuaire de sources de données composées d'annotations sémantiques des sources. Pour obtenir un espace commun de navigation et d'interrogation, il effectue une classification en treillis et se sert de l'Analyse de Concepts Formels (ACF).

Le contexte sur lequel il travaille est multi-valué et il propose une analyse par similarité. Dans une ACF standard deux objets sont regroupés s'ils ont un même attribut. En ACF par similarité, deux objets sont regroupés s'ils ont des valeurs similaires pour un même attribut.

La fonction de similarité entre les valeurs des attributs est calculée à partir de relations sémantiques (comme une distance dans une hiérarchie) ou par un paramètre dans la définition de partage d'attributs (par un seuil de similarité).

Les relations d'ordre partiel définies pour un tel treillis sont :

- Sur les objets : l'inclusion $\subseteq$
- Sur les attributs : l'inclusion par similarité $\subseteq_S$: Soient un contexte multivalué (O, T, M, I) , t_1 et $t_2 \in T$, m_1 et $m_2 \in M$, B_1 et $B_2 \in \mathfrak{P}(T \times \mathfrak{P}(M))$.

$$B_1 \subseteq_S B_2 \text{ ssi } \forall t_{1|m_1} \in B_1, \exists t_{2|m_2} \in B_2 \text{ tel que } t_1 = t_2 \text{ et } m_1 \subseteq m_2.$$

(t_m peut s'illustrer dans le contexte filmographique par *Nationalité*_{FR} et se lit "la nationalité française").

Les opérateurs de dérivation qui en découlent sont :

- Pour tout $A \subseteq O$ l'ensemble des éléments de $T \times \mathfrak{P}(M)$ partagés par A est :

$$A^{\uparrow S} = \{t_{\{t(o), o \in A\}} \in T \times \mathfrak{P}(M) \mid \forall o_i, o_j \in A, t(o_i) \simeq t(o_j)\}$$

- Pour tout $B \subseteq T \times \mathfrak{P}(M)$, l'ensemble des objets de O partageant B est :

$$B^{\downarrow S} = \{o \in O \mid \forall t_{M_t} \in B, t(o) \simeq m \forall m \in M_t\}$$

$\uparrow S$ et $\downarrow S$ forment une connexion de Galois appelée **connexion de Galois par similarité** entre $(O, \subseteq)$ et $(T \times \mathfrak{P}(M), \subseteq_S)$

Un concept multi-valué (A, B) est tel que $A \subseteq O$, $B \subseteq T \times \mathfrak{P}(M)$, $A^{\uparrow S} = B$ et $B^{\downarrow S} = A$

2.3.6 Règles d'association et implications

L'extraction de ces règles est un problème important dans le domaine de l'extraction de connaissances et tire son origine de l'analyse de bases de données commerciales dans le but d'optimiser les ventes dans le cadre du "panier de la ménagère". Leur intérêt les a vite fait dépasser ce stade et maintenant les règles d'associations sont utilisées pour des applications très variées sortant largement du domaine commercial.

Dans [8], les auteurs définissent une règle d'association comme une relation d'implication entre deux ensembles d'articles X et Y . $X \rightarrow Y$ indique que les transactions contenant les articles de l'ensemble X ont tendance à contenir ceux de l'ensemble Y . X est appelé condition ou prémisses, et Y résultat ou conclusion.

Soit T l'ensemble des transactions t . Afin de déterminer la qualité d'une règle, on utilise généralement deux types de mesure :

- le *support* qui indique la probabilité que X et Y soient vrais en même temps. Il s'écrit formellement :

$$\text{support}(X \rightarrow Y) = \text{support}(X \cup Y) = \frac{|\{t \in T : X \subseteq t \text{ et } Y \subseteq t\}|}{|T|}.$$

Par exemple le pourcentage de fois où de la bière et des bretzels ont été achetés ensemble. On appelle support d'un itemset X le rapport entre le nombre de transactions contenant X et le nombre total de transactions de T .

- la *confiance* qui indique la probabilité que Y soit vrai sachant que X est vrai. Elle s'écrit formellement :

$$\text{confiance}(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}.$$

La confiance traduit par exemple, la probabilité qu'un client qui a acheté des bières achète aussi des bretzels.

Dans le cadre des treillis de Galois, nous pouvons extraire des *itemsets fréquents*, i.e. des ensembles d'attributs regroupés ensemble selon un support (nombre d'individus partageant ces attributs) supérieur à un seuil et des règles d'association avec une certaine confiance, soulignant les corrélations entre les ensemble d'attributs [55]. Nous nous intéressons plus particulièrement aux *implications* [29] qui sont des règles d'association pour lesquelles la confiance est de 100%.

Une implication $A \rightarrow B$ concerne les attributs A et B d'un même concept. Le support d'une implication est la taille de l'extension du concept, c'est à dire le nombre d'objets partageant à la fois les attributs A et B . Ces règles sont appliquées sur le treillis des macrophytes, et présentées page 49.

2.4 Conclusion

Dans ce chapitre, nous avons présenté un état de l'art de l'analyse de concepts formels en définissant les notions nécessaires aux treillis de Galois. Puis nous avons abordé les cas de contextes complexes et des propositions de traitement de ces contextes, puisque nos données présentées dans les parties 1.4.3 pour les macrophytes et 1.4.2 pour les invertébrés, présentent des caractéristiques de deux types de contextes complexes : les contextes multi-valués et flous. Dans le chapitre suivant nous binarisons nos données pour les rapporter à un contexte simple exploitable par les algorithmes classiques. Puis dans le chapitre 4 nous introduisons des définitions qui s'appuient sur celles de ce chapitre afin de présenter une méthode de traitement flou des contextes multi-valués flous.

Chapitre 3

Traitement binaire

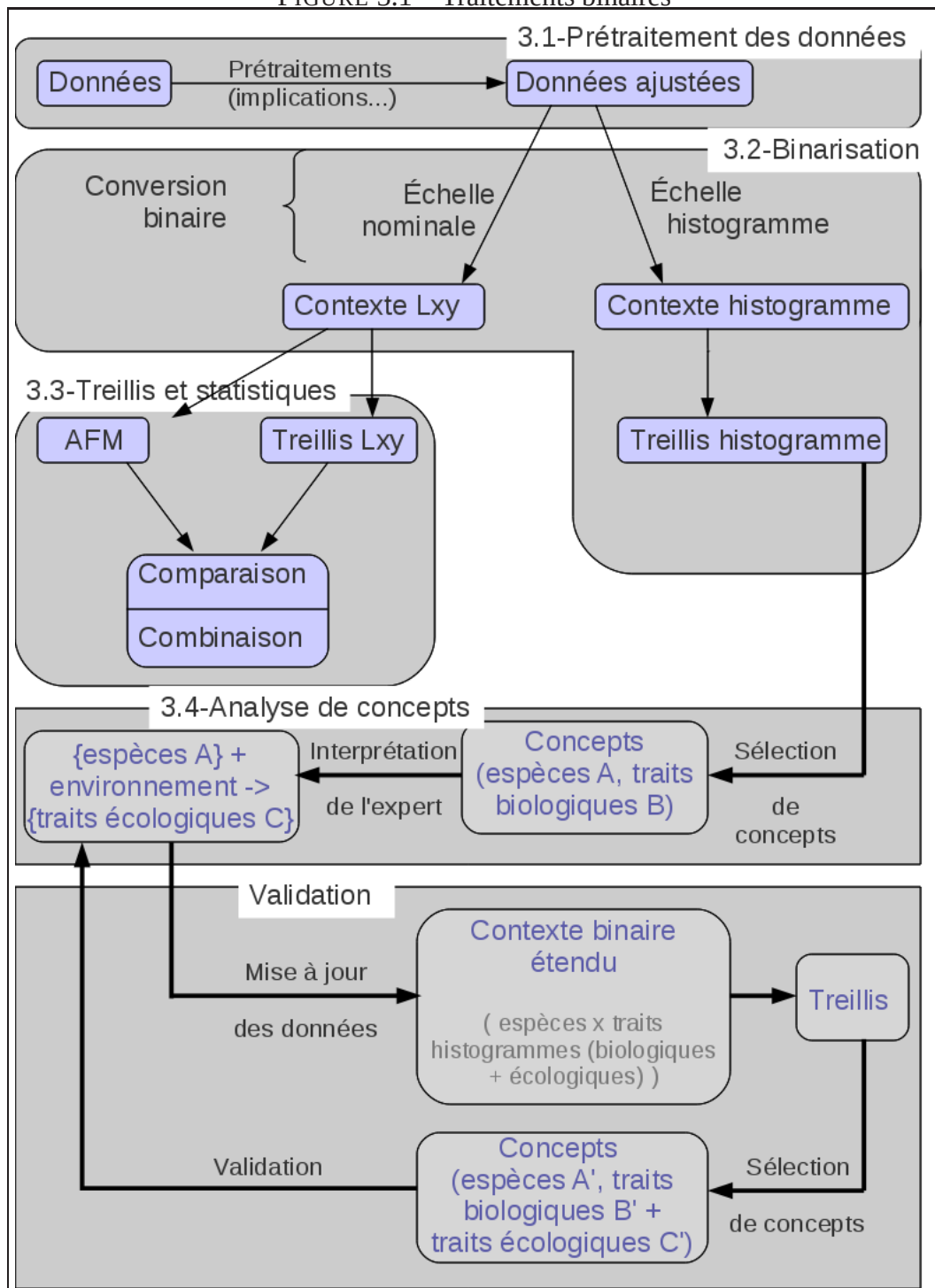
Sommaire

3.1	Les données et leurs pré-traitements	49
3.1.1	Règles d'implication	49
3.1.2	Ajustement du jeu de données	52
3.2	Méthodes de conversion en tableau binaire	53
3.2.1	Echelle nominale	56
3.2.2	Echelle histogramme	57
3.3	Comparaison et combinaison des treillis et des méthodes statistiques .	62
3.3.1	Analyse factorielle multiple	62
3.3.2	Apport des treillis	71
3.4	Démarche complète d'identification de traits écologiques	77
3.4.1	Analyse des concepts formels	77
3.4.2	Validation	80
3.5	Conclusion	84

Dans ce chapitre nous explorons diverses approches de traitement binaire de nos données afin d'explorer les possibilités offertes par les treillis et répondre à la problématique de sélection des traits écologiques. Nous commençons par décrire nos données qui ne se présentent pas sous un format binaire et les prétraitements effectués sur les données. Puis nous présentons deux méthodes de conversion des données en contextes binaires. Ensuite nous présentons une approche combinant les treillis et une analyse factorielle multiple qui est une méthode statistique couramment employée par les biologistes, et une comparaison des apports de ces méthodes. Enfin nous présentons une démarche complète allant de la conversion à sa validation, nous permettant de déterminer les traits écologiques pertinents.

La figure 3.1 montre les différentes étapes et processus décrits dans ce chapitre et indique les références aux sections correspondantes.

FIGURE 3.1 – Traitements binaires



3.1 Les données et leurs pré-traitements

Le premier jeu de données utilisé est celui décrit en 1.4.2 et concerne les macrophytes. Il comporte 21 espèces, 10 traits et 37 modalités, et a été ajusté au niveau des espèces, des traits, des modalités et des affinités. Cet ajustement a été opéré en collaboration avec Michèle Trémolières, professeure spécialiste des macrophytes.

Les espèces ont été sélectionnées selon deux critères. Le premier est de ne conserver que des plantes présentes dans la plaine d'Alsace, et le second, de posséder toutes les informations sur ces plantes concernant les traits et modalités que nous prenons en compte. Ces modifications ont été réalisées pour rendre le jeu de données exploitable notamment face à l'absence d'information.

Les traits sont ceux considérés comme pertinents par l'experte. Nous ne traitons que des traits biologiques, i.e. qui ne décrivent que la morphologie de la plante comme sa taille ou son mode de reproduction par exemple. Ces modifications ont permis de centrer l'analyse de ces données sur les traits biologiques.

Les affinités valaient entre 0 et 3 allant de l'absence d'individu (0) à une très grande quantité (3) pour chaque espèce.

La manipulation de ces données non binaires impose leur conversion en de nouvelles propriétés binaires. Le format "Lxy" est développé dans la partie 3.2.1. Dans cette partie-ci, nous avons juste besoin de savoir que 'L' représente le trait ('R' pour le potentiel de Régénération par exemple), 'x' est le numéro de la modalité, et 'y' la valeur de l'affinité de cette modalité. La conversion Lxy a généré 112 propriétés. Pour mettre à jour ce jeu de données, nous avons notamment utilisé des règles d'implication.

3.1.1 Règles d'implication

Les implications de Duquenne-Guigues [29] définies dans la partie 2.3.6 sont utiles pour interpréter un contexte car elles montrent les relations entre les propriétés. De manière générale, plus le support d'une implication est important plus elle doit être prise en compte puisqu'elle concerne davantage d'objets (pour nous des espèces). Cependant cette considération reste générale et il est parfois pertinent de s'intéresser à des cas particuliers tels que les implications négatives ou les implications ayant un faible support. Nous avons généré les implications sur notre jeu de données initial et avons interprété ces implications particulières.

Implications négatives

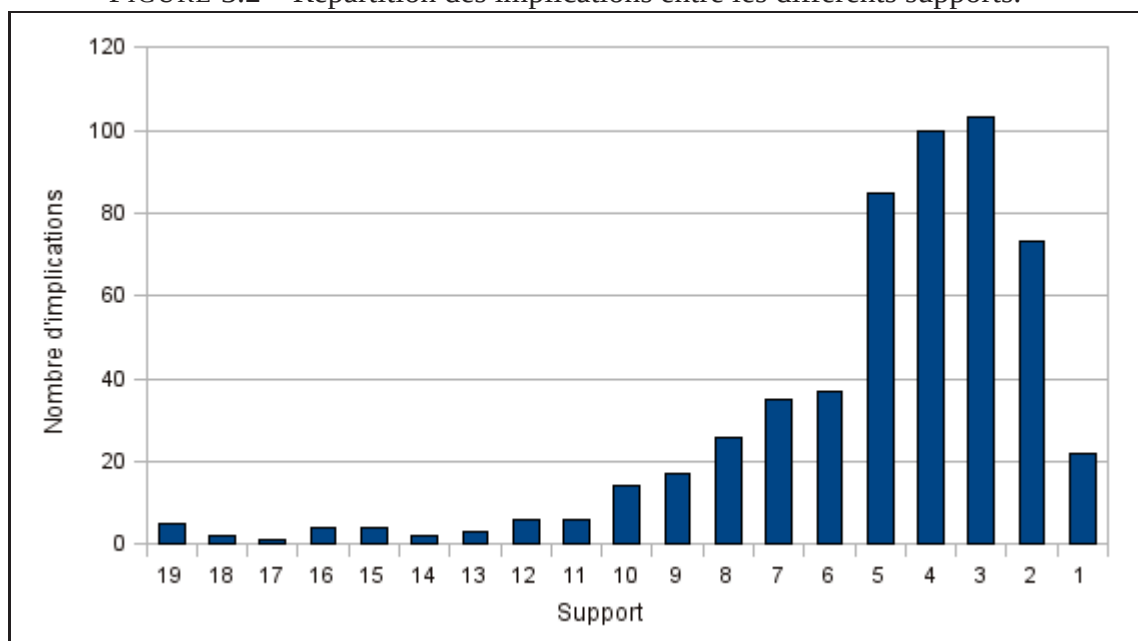
Nous avons constaté dans les 545 implications issues de nos données qu'elles sont davantage basées sur l'absence d'une propriété que sur sa présence. En effet, le jeu de données contient 55,6% d'affinités '0'.

Après conversion, les propriétés L_{xy} ayant $y=0$ représentent 33% ce qui signifie que beaucoup des affinités nulles étaient communes à plusieurs espèces. Cependant dans l'ensemble des implications, 75,9% des propriétés mentionnées sont des propriétés L_{xy} avec $y=0$.

Ceci peut paraître problématique, mais ce n'est pas obligatoirement le cas étant donné que l'absence d'une modalité représente en soi une information que l'on peut exploiter. Par exemple si on constate en médecine que l'ingestion de certains aliments implique l'absence de cholestérol, cela peut indiquer l'intérêt de ces aliments pour le réduire.

Nous présentons certaines implications afin d'illustrer leur intérêt pour les modifications et la validation du jeu de données. Le schéma de la figure 3.2 présente la répartition des implications en fonction du support compris entre 1 et 19.

FIGURE 3.2 – Répartition des implications entre les différents supports.



Nous rappelons que le support d'une implication est, dans notre contexte, le nombre d'espèces pour lesquelles cette implication est vraie. Il est fréquent que peu d'implications possèdent un support élevé. Par ce graphique, nous constatons que les implications avec un faible support (inférieur à ou égal à 5) sont très nombreuses. Nous nous intéressons donc à la signification de ces faibles supports.

Prenons le cas de l'implication

$$(1) : F10 \Rightarrow D30^{19} \quad < support = 19 >$$

19. Ces notations sont expliquées à la section 3.2.1.

qui indique que les espèces (parmi les 19 composant son support) qui n'ont pas la propriété "aucune flexibilité <10%" (autrement dit une flexibilité > 10%) n'ont pas de dispersion faible (autrement dit elles ont une dispersion intermédiaire ou forte). Cette règle seule inclut toutes celles qui se traduisent par : "une espèce flexible (propriétés F21, F22, F23, F31, F32, F33, F41, F42 ou F43) a une dispersion certaine (propriétés D11, D12, D13, D21, D22 ou D23)".

Par ailleurs, l'implication

$$(2) : C20 \Rightarrow D30 \text{ F10} \quad < \text{support} = 16 >$$

précise que sur ces 19 espèces flexibles avec une certaine dispersion (D30 F10 : BERE, CALO, CHAR, CHAH, CHAV, ELOC, ELOE, ELON, LEMM, LEMT, MENA, MYRS, NASO, PTCO, PTNO, PTPE, RANC, RANU, SEFC), 16 (support de l'implication 2 : BERE, CHAR, CHAH, CHAV, ELOC, ELOE, ELON, LEMM, LEMT, MENA, MYRS, NASO, PTPE, RANC, RANU, SEFC) n'ont pas de croissance ancrée avec des feuilles flottantes (C20). C'est à dire que soit elles sont *non ancrées* (modalité 1), soit elles le sont par du *tissu de soutien* (modalité 3) ou sont *submerses* (modalité 4).

Considérons une autre implication :

$$(3) : C30 \text{ D30 F10} \Rightarrow M50 \quad < \text{support} = 16 >$$

qui se lit : "les espèces n'ayant pas de tige rigide (C30), ni de faible potentiel de dispersion (D30), ni de flexibilité <10% (F10), ne sont pas des espèces de gros volume (M50)". En fait, cette implication négative souligne le caractère hydrophyte⁽²²⁾ des 16 espèces concernées (support de l'implication 3 : BERE, CHAR, CHAH, CHAV, ELOC, ELOE, ELON, LEMM, LEMT, MENA, MYRS, NASO, PTCO, PTPE, RANC, RANU) : un volume moyen, une flexibilité moyenne ou haute et une tige flexible immergée.

Faible support

Le cas des implications avec un support faible est intéressant car elles peuvent révéler des relations spécifiques entre les propriétés. Par exemple, cette implication

$$(4) : C30 \text{ D23} \Rightarrow F20 \quad < \text{support} = 9 >$$

signifie que "les espèces qui n'ont pas de tige rigide (C30) et ont essentiellement un niveau moyen de dispersion (D23), n'ont pas une flexibilité moyenne (F20)". Le support de cette implication est 9 (BERE, ELOC, ELOE, ELON, LEMM, LEMT, NUPL, NYMA, PTPE).

En recoupant avec l'interprétation de l'implication (3), cette seconde implication peut être transformée en

$$(5) : D23 \Rightarrow (F3 \neq 0)$$

mettant en lumière le lien entre un niveau intermédiaire de dispersion et une grande flexibilité pour certains hydrophytes.

Détaillons comment nous obtenons l'implication (5). Pour cela nous nous appuyons sur l'implication suivante :

$$(6) : F40 \Rightarrow D30 F10 M50 V10 \quad < support = 12^{20} >$$

.

(3 et 6) : Nous remplaçons D30 M50 par F40²¹ et obtenons :

$$(7) : C30 F10 \Rightarrow F40$$

(7 et 4) : Nous remplaçons C30 par F10 F40 et obtenons :

$$(5) : D23 \Rightarrow F10 F20 F40 \Leftrightarrow$$

$$D23 \Rightarrow F(3 \neq 0)$$

pour les plantes issues de l'intersection des supports des implications 1, 2, 4, et 5 (c'est à dire possédant l'union des propriétés : C30, D30, F10, M50, D23, F20, F40, V10) : BERE, ELOC, ELOE, ELON, LEMM. Notons que l'intersection des supports est un sous-ensemble de l'extension du concept ayant D23 et $F3 \neq 0$ dans son intension.

D'une manière générale, les implications ayant un faible support mettent en exergue les propriétés qui ont tendance à s'exclure. Prenons un autre exemple d'implication qui relie les propriétés : (forme de croissance) C10, C22, C31, C41, C52, (dispersion potentielle) D10, D23, D30, (flexibilité) F10, F20, F33, F40, (phénologie) G13, G20, (floraison) L11, L23, L30, (index morphologique) M10, M23, M30, M40, M50, (pérennité) P13, P20, P30, P40, (potentiel de régénération) R10, R23, R30, (taille) T10, T21, T32, T42, (reproduction végétative) V10, V20, V30, V43. Son support est de 1, ce qui signifie qu'une seule espèce (*Callitriche obtusangula*) réunit ces propriétés et qu'en dehors d'elle, ces propriétés ne sont pas compatibles.

Les implications (fournies en annexe A.4) et leurs supports ont été utilisés pour détecter des données incorrectes, inutiles ou redondantes, et ainsi mettre à jour et valider notre jeu de données, tant pour les espèces, les traits, les modalités que pour les affinités.

3.1.2 Ajustement du jeu de données

Grâce à l'analyse des données menées notamment avec les implications, nous avons pu ajuster le jeu de données pour qu'il soit à la fois plus cohérent et plus pratique pour

20. 12 espèces partagent ces attributs.

21. Si $F40 \Rightarrow D30 F10 M50 V10$ est vraie, alors $F40 \Rightarrow D30 F10 M50$ est également vraie pour le même ensemble d'espèces.

nos études. Nous avons apporté des modifications sur les affinités, les plantes, les traits et les modalités.

L'amplitude de la plage de valeurs des affinités a été revue en la ramenant à $[0;2]$ pour deux raisons. Nous avons fusionné les affinités 2 et 3, car d'une part l'affinité 2 ne représente que 8% des cas par rapport aux 3 autres affinités, et que les attributs L_{xy} avec $y=2$ dans les implications ne représentent que 3,4%. D'autre part, cela nous permet de diminuer le nombre de propriétés résultant de nos travaux, les rendant ainsi moins spécifiques.

Par ailleurs, nous avons éliminé certaines plantes dont le comportement était identique à d'autres, ou dont les informations n'étaient pas collectables, comme c'est le cas pour les *Characeae* : *Chara fragilis*, *Chara hispida* et *Chara vulgaris*.

Cependant, les principales modifications ont été opérées sur les modalités et les traits. Certaines informations étaient redondantes comme par exemple le trait "Ancrage au sol ou au substrat" avec les modalités du trait "Forme de croissance" qui sont "non ancrée", "ancrée avec feuilles flottantes", "ancrée avec tissus de soutien" et "ancrée submerse".

Des ajustements de modalités ont aussi été effectués pour limiter la dispersion d'information, comme par exemple pour le trait "période de reproduction" qui avait une modalité par mois et qui donnait des résultats inutilement spécifiques. Ces modalités ont donc été regroupées en bimestres, ce qui reste biologiquement cohérent par rapport au cycle saisonnier. Certaines modalités ont été supprimées comme la sténocécie⁽⁵⁷⁾, car elle n'avait pas de rapport suffisant avec le trait auquel elle était affectée.

Notre jeu de données final est celui présenté dans les tableaux de la figure 3.3.

Nos données étant valides, nous devons ensuite les binariser pour pouvoir les traiter.

3.2 Méthodes de conversion en tableau binaire

Afin de convertir nos données, nous nous basons dans un premier temps sur les *échelles nominales* [37], qui nous permettent d'obtenir une table disjonctive totale en générant une propriété par affinité de chacune des modalités. Dans un second temps nous nous intéressons à ce que nous avons défini comme une *échelle histogramme*, nous permettant de prendre en compte les informations intrinsèques au format particulier des données (des traits divisés en modalités ayant des affinités non binaires) en générant des propriétés pour chaque combinaison d'affinités de chaque trait.

Les échelles conceptuelles [35], [62] ont été définies dans le but de manipuler des volumes de données de plus en plus importants ainsi que pour des contextes multi-valués.

Nous utilisons le logiciel ConExp [70] pour construire et interpréter les treillis. Il nous permet d'éditer un contexte binaire, dessiner un treillis et calculer les bases d'implications de Duquenne-Guigues [29] entre les propriétés.

FIGURE 3.3 – Jeu de données actuel

TRAITS		Taille Potentielle				Forme de Croissance				Index Morphologique					Flexibilité				Pérennité		
MODALITÉS		<0.08m	0.08-0.3 m	0.3 -1m	1-5 m	non ancrée	ancrée, feuilles flottantes	ancrée, tissu de soutien	ancrée, submerse	0	1	2	3	4	aucune (< 10°)	faible (> 10 - 45°)	intermédiaire (> 45 - 300°)	forte (> 300°)	Organes pérennes (souterrains ou aériens)	Bisannuelle	annuelle
<i>Alisma plantago-aquatica</i>	ALIP	0	1	2	1	0	0	2	0	0	0	1	2	0	0	2	1	0	2	1	0
<i>Berula erecta</i>	BERE	1	2	2	0	0	0	1	2	0	0	1	2	0	0	0	2	0	2	0	0
<i>Callitriche obtusangula</i>	CALO	0	1	2	2	0	2	1	2	0	2	0	0	0	0	0	2	1	2	0	1
<i>Callitriche platycarpa</i>	CALP	0	2	2	0	0	1	0	2	0	2	0	0	0	0	0	2	0	2	0	2
<i>Carex acutiformis</i>	CARA	0	0	2	0	0	0	2	0	0	0	0	2	0	2	2	0	0	2	0	0
<i>Ceratophyllum demersum</i>	CERD	0	0	2	2	2	0	0	0	0	2	2	2	0	0	0	0	2	2	0	0
<i>Eleocharis acicularis</i>	ELEA	2	2	2	0	0	0	2	2	0	0	2	0	0	0	0	2	0	2	0	0
<i>Elodea canadensis</i>	ELOC	0	2	2	1	0	0	0	2	0	2	2	1	0	0	0	1	2	2	0	0
<i>Elodea ernstae</i>	ELOE	0	2	2	1	0	0	0	2	0	2	2	1	0	0	0	1	2	2	0	0
<i>Elodea nuttallii</i>	ELON	0	2	2	1	0	0	0	2	0	2	2	1	0	0	0	1	2	2	0	0
<i>Groenlandia densa</i>	GROD	0	1	2	0	0	0	0	2	0	1	2	0	0	2	0	2	0	2	0	0
<i>Hippuris vulgaris</i>	HIPV	0	1	2	0	0	0	0	2	0	0	1	2	2	0	2	0	0	2	0	0
<i>Hottonia palustris</i>	HOTP	0	1	2	0	0	0	0	2	0	0	2	0	0	0	2	2	0	2	0	0
<i>Hydrocharis morsus-ranae</i>	HYDM	1	2	0	0	2	0	0	0	1	2	2	0	0	1	2	2	0	2	0	0
<i>Iris pseudacorus</i>	IRIP	0	0	2	1	0	0	2	0	0	0	0	0	2	1	2	1	0	2	0	0
<i>Juncus articulatus</i>	JUNA	0	2	2	0	0	0	2	0	0	0	2	2	0	2	0	0	0	2	0	0
<i>Lemna minor</i>	LEMM	2	0	0	0	2	0	0	0	2	0	0	0	0	0	0	2	0	2	0	0
<i>Lemna trisulca</i>	LEMT	2	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0	2	2	0	0
<i>Mentha aquatica</i>	MENA	0	1	2	1	0	0	2	2	0	2	2	0	0	0	2	2	0	2	0	0
<i>Myosotis palustris</i> (= <i>scorpioides</i>)	MYOP	0	0	2	1	0	0	2	2	0	2	2	1	0	0	2	2	0	2	1	0
<i>Myriophyllum spicatum</i>	MYRS	0	2	2	2	0	0	0	2	0	2	2	1	0	0	0	0	2	2	0	0
<i>Myriophyllum verticillatum</i>	MYRV	0	2	2	0	0	0	0	2	0	2	2	0	0	0	0	0	2	2	0	0
<i>Nasturtium officinale</i>	NASO	0	2	2	0	0	0	1	2	0	0	2	0	0	0	0	2	0	2	0	0
<i>Nuphar lutea</i>	NUPL	0	0	1	2	0	2	0	0	0	0	0	0	2	2	0	0	2	2	0	0
<i>Nymphaea alba</i>	NYMA	0	1	2	2	0	2	0	0	0	0	0	1	2	2	0	0	2	2	0	0
<i>Oenanthe fluviatilis</i>	OENF	0	0	2	2	0	0	2	2	0	0	1	2	0	0	0	2	2	2	0	0
<i>Phalaris arundinacea</i>	PHAA	0	0	1	2	0	0	2	0	0	0	1	2	0	0	0	0	2	2	0	0
<i>Phragmites australis</i>	PHRG	0	0	1	2	0	0	2	0	0	0	0	1	2	2	2	2	0	2	0	0
<i>Polygonum hydropiper</i>	PLGH	0	1	2	0	0	0	2	0	0	2	1	0	0	0	2	2	0	0	0	1
<i>Potamogeton coloratus</i>	PTCO	0	0	2	0	0	1	0	2	0	0	2	0	0	0	0	2	0	2	0	0
<i>Potamogeton crispus</i>	PTCR	0	0	2	0	0	0	0	2	0	0	2	2	0	0	0	0	2	2	0	0
<i>Potamogeton lucens</i>	PTLU	0	0	1	2	0	0	0	2	0	0	2	2	0	0	0	0	2	2	0	0
<i>Potamogeton natans</i>	PTNA	0	0	2	1	0	2	0	0	0	1	2	2	0	0	0	0	2	2	0	0
<i>Potamogeton nodosus</i>	PTNO	0	0	2	2	0	2	0	2	0	0	2	2	0	0	0	0	2	2	0	0
<i>Potamogeton pectinatus</i>	PTPE	0	0	1	2	0	0	0	2	0	2	2	2	0	0	0	0	2	2	0	1
<i>Potamogeton pusillus</i>	PTPU	0	1	2	0	0	0	0	2	0	2	1	0	0	0	0	0	2	0	0	2
<i>Ranunculus circinatus</i>	RANC	0	1	2	2	0	0	0	2	0	1	2	2	0	0	0	0	2	2	0	0
<i>Ranunculus fluitans</i>	RANU	0	1	2	2	0	0	0	2	0	1	2	2	0	0	0	0	2	2	0	0
<i>Ranunculus trichophyllus</i>	RANT	0	1	2	2	0	0	0	2	0	1	2	2	0	0	0	0	2	2	2	2
<i>Roripa amphibia</i>	RORA	0	0	2	0	0	0	1	2	0	0	2	2	2	0	2	2	0	2	0	0
<i>Sagittaria sagittifolia</i>	SAGS	0	0	2	0	0	0	1	1	0	0	0	2	0	0	2	0	2	2	0	0
<i>Sparganium emersum</i>	SEFC	0	1	2	1	0	0	1	1	0	0	1	2	1	0	2	0	2	2	0	0
<i>Typha latifolia</i>	TYPL	0	0	0	2	0	0	2	0	0	0	0	1	2	1	2	0	0	2	0	0
<i>Utricularia vulgaris</i>	UTRV	0	2	2	0	2	0	0	0	0	2	1	0	0	0	0	0	2	2	0	0
<i>Veronica anagallis-aquatica</i>	VERA	0	2	2	0	0	0	2	2	0	0	2	1	0	0	2	2	0	2	1	1
<i>Veronica beccabunga</i>	VERB	0	1	2	0	0	0	2	2	0	0	2	1	0	0	2	1	0	2	0	0
<i>Zannichellia palustris</i>	ZANP	0	2	1	0	0	0	0	2	0	0	0	2	1	0	0	2	1	2	0	2

TRAITS		Phénologie		Floraison		Reproduction Végétative				Période de Reproduction				Dispersion Potentielle		
MODALITÉS		Toute l'année	Seulement durant la saison de végétation	oui (annuelle ou pas)	non (Rarement ou jamais)	Bulbes ou tubercules	Rhizomes ou stolons	Bulbilles, turions ou apex dormants	Fragments non spécialisés	Mars-mai	Juin-juillet	Avant-septembre	Octobre	Elevée (petites graines volantes) 100 et 1000	Intermédiaire (propagules flottantes, véhiculées par les flux hydriques) 10 et 100	Faible (potentialités de transport faibles: graines ou propagules non flottantes, ou de faible longévité, ou absentes) <10
<i>Alisma plantago-aquatica</i>	ALIP	0	2	2	1	0	2	0	0	0	2	2	0	1	2	0
<i>Berula erecta</i>	BERE	2	0	2	1	0	2	0	1	0	2	2	0	2	0	0
<i>Callitriche obtusangula</i>	CALO	2	0	2	0	0	0	0	2	2	2	2	0	1	2	0
<i>Callitriche platycarpa</i>	CALP	0	2	2	0	0	0	0	2	2	2	2	0	1	2	0
<i>Carex acutiformis</i>	CARA	0	2	2	0	0	2	0	0	2	2	0	0	0	2	0
<i>Ceratophyllum demersum</i>	CERD	0	2	0	2	0	0	2	2	0	2	2	0	0	2	1
<i>Eleocharis acicularis</i>	ELEA	0	2	2	0	0	2	0	0	0	1	2	2	0	2	0
<i>Elodea canadensis</i>	ELOC	2	0	2	0	0	0	2	2	0	2	2	1	0	1	2
<i>Elodea emstae</i>	ELOE	2	0	2	0	0	1	2	2	0	0	0	0	0	2	0
<i>Elodea nuttallii</i>	ELON	2	0	2	0	0	0	1	2	0	2	2	1	0	1	2
<i>Groenlandia densa</i>	GROD	0	2	0	2	2	2	0	1	0	2	2	0	2	2	0
<i>Hippuris vulgaris</i>	HIPV	0	2	2	0	0	2	0	1	0	2	1	0	0	0	2
<i>Hottonia palustris</i>	HOTP	0	2	2	0	0	2	0	2	2	2	0	0	2	1	0
<i>Hydrocharis morsus-ranae</i>	HYDM	0	2	2	0	0	2	2	0	0	2	2	0	1	2	0
<i>Iris pseudacorus</i>	IRIP	0	2	2	0	1	1	0	2	0	2	0	0	2	0	0
<i>Juncus articulatus</i>	JUNA	0	2	2	0	0	2	0	0	0	2	2	2	2	0	0
<i>Lemna minor</i>	LEMM	1	2	0	2	0	0	2	0	1	2	2	0	2	0	0
<i>Lemna trisulca</i>	LEMT	1	2	0	2	0	0	2	0	2	2	0	0	2	0	0
<i>Mentha aquatica</i>	MENA	1	0	2	0	0	2	0	0	0	2	2	0	2	2	0
<i>Myosotis palustris</i> (= <i>scorpioides</i>)	MYOP	1	0	2	0	0	2	0	1	2	2	2	1	1	2	0
<i>Myriophyllum spicatum</i>	MYRS	1	0	1	0	0	2	1	2	0	2	2	0	1	2	0
<i>Myriophyllum verticillatum</i>	MYRV	1	0	2	0	0	2	2	2	0	2	0	0	1	2	0
<i>Nasturtium officinale</i>	NASO	1	0	2	0	0	1	0	2	2	2	2	2	2	1	0
<i>Nuphar lutea</i>	NUPL	0	2	2	0	0	2	0	0	0	2	2	0	2	1	0
<i>Nymphaea alba</i>	NYMA	0	2	2	0	0	2	0	0	0	2	2	0	2	0	0
<i>Oenanthe fluviatilis</i>	OENF	0	2	0	2	0	1	0	2	0	2	2	0	2	0	0
<i>Phalaris arundinacea</i>	PHAA	0	2	2	0	0	2	0	0	0	2	2	0	2	2	0
<i>Phragmites australis</i>	PHRG	2	0	2	0	0	2	0	0	0	0	2	2	0	2	0
<i>Polygonum hydropiper</i>	PLGH	2	1	1	0	0	0	0	0	0	1	2	1	1	0	0
<i>Potamogeton coloratus</i>	PTCO	2	1	2	0	0	2	0	1	0	2	1	2	1	2	0
<i>Potamogeton crispus</i>	PTCR	0	2	2	0	0	2	2	1	2	2	2	2	2	1	0
<i>Potamogeton lucens</i>	PTLU	0	2	2	0	0	2	0	1	0	2	2	0	1	2	0
<i>Potamogeton natans</i>	PTNA	0	2	2	0	0	2	1	1	2	2	2	0	2	0	0
<i>Potamogeton nodosus</i>	PTNO	1	2	2	0	0	2	0	1	0	0	2	1	2	1	0
<i>Potamogeton pectinatus</i>	PTPE	1	2	1	2	2	2	0	2	2	2	2	0	2	1	0
<i>Potamogeton pusillus</i>	PTPU	0	2	2	0	0	0	2	0	0	2	2	2	2	1	0
<i>Ranunculus circinatus</i>	RANC	1	2	1	0	0	0	0	2	0	2	2	0	2	0	0
<i>Ranunculus fluitans</i>	RANU	1	2	2	1	0	0	0	2	1	2	2	0	0	2	0
<i>Ranunculus trichophyllus</i>	RANT	1	1	2	0	0	0	0	2	2	2	0	0	2	0	0
<i>Rorippa amphibia</i>	RORA	1	2	2	0	2	0	0	2	2	2	2	2	2	0	0
<i>Sagittaria sagittifolia</i>	SAGS	0	2	2	0	0	2	2	0	0	2	2	2	2	0	0
<i>Sparganium emersum</i>	SEFC	0	2	2	0	0	2	0	0	0	2	0	0	2	0	0
<i>Typha latifolia</i>	TYPL	0	2	1	0	0	2	0	0	2	2	0	0	2	0	0
<i>Utricularia vulgaris</i>	UTRV	0	2	2	0	0	0	2	2	0	1	2	2	2	2	1
<i>Veronica anagallis-aquatica</i>	VERA	0	2	2	0	0	1	0	2	1	1	2	0	2	0	0
<i>Veronica beccabunga</i>	VERB	0	2	2	0	0	2	0	2	2	2	2	2	2	0	0
<i>Zannichellia palustris</i>	ZANP	0	2	2	0	0	2	0	1	0	2	2	0	1	2	0

3.2.1 Echelle nominale

Notre contexte peut être décrit comme un contexte multi-valué tel que défini dans la partie 2.3.1. Pour obtenir les données sous forme d'une table binaire telle que la table 3.1, nous avons appliqué une *échelle nominale*²² [37] sur le contexte de traits biologiques. Nous rappelons que le nom des nouveaux attributs suit un modèle de la forme 'Lxy' où 'L' représente le trait ('R' pour le potentiel de Régénération par exemple), 'x' est le numéro de la modalité, et 'y' la valeur de l'affinité. Pour éviter les confusions, ces nouveaux attributs seront appelés *propriétés*. La table 3.1 présente les propriétés issues du trait "taille potentielle" pour un sous-ensemble de 15 plantes. Par exemple, T21 signifie que "peu d'individus (affinité 1) ont une taille potentielle (T) entre 0,08 et 0,3 mètre (seconde modalité)". La figure 3.4 présente le treillis correspondant au contexte de la table 3.1.

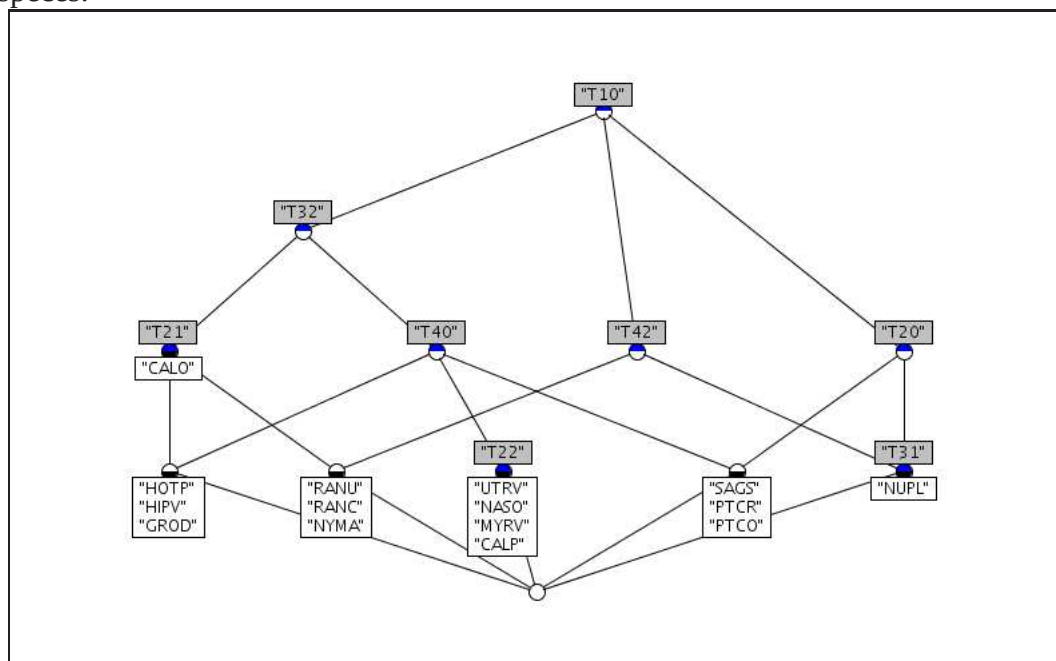
	T10	T20	T21	T22	T31	T32	T40	T42
CALO	×		×			×		
CALP	×			×		×	×	
GROD	×		×			×	×	
HIPV	×		×			×	×	
HOTP	×		×			×	×	
MYRV	×			×		×	×	
NASO	×			×		×	×	
NUPL	×	×			×			×
NYMA	×		×			×		×
PTCO	×	×				×	×	
PTCR	×	×				×	×	
RANC	×		×			×		×
RANU	×		×			×		×
SAGS	×	×				×	×	
UTRV	×			×		×	×	

TABLE 3.1 – Table disjonctive totale pour le trait "taille potentielle" pour un sous ensemble de 15 espèces.

L'échelonnage nominal nous permet d'obtenir une table binaire et répond donc à nos besoins pour la manipulation des treillis de Galois. Cependant, une telle conversion nous donne un contexte dont le treillis (présenté par la figure 3.5), même limité à 21 objets et 10 traits, est beaucoup trop volumineux pour être exploitable. De plus, dans le cadre

22. *Nominal scale* en anglais. Il s'agit d'une échelle conceptuelle telle que les valeurs (modalités) s'excluent mutuellement.

FIGURE 3.4 – Treillis pour le trait de "taille potentielle" pour un sous-ensemble de 15 espèces.



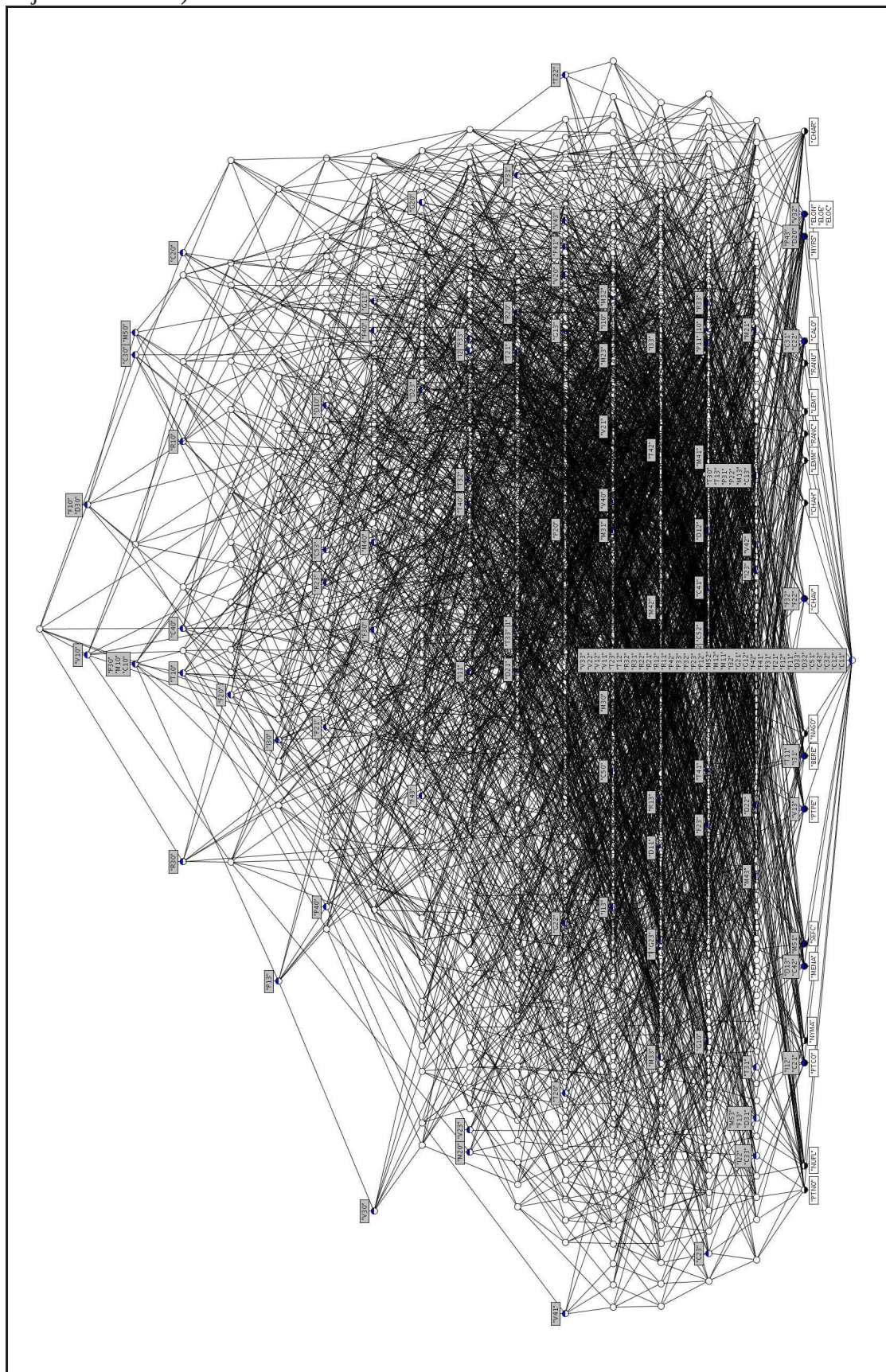
de nos données, les traits sont divisés en modalités, et chaque espèce possède plus ou moins d'individus (affinité) pour chacune de ces modalités. La répartition des individus dans les modalités de chaque trait représente un schéma semblable à un histogramme qui est une information en elle-même pour les biologistes. Pour conserver cette information, une conversion par échelle nominale n'est pas suffisante car elle brise les liens entre les modalités. Ainsi pour conserver ensemble les différentes modalités d'un même trait dans l'opération de binarisation du contexte, nous avons recours à l'*échelonnage histogramme* que nous présentons.

3.2.2 Echelle histogramme

Signification biologique des affinités

Préalablement, pour garder la notion de répartition, nous considérons que chaque espèce a 100% de ses individus répartis dans chaque trait. Nous remplaçons donc par un pourcentage chacune des affinités proportionnellement à sa valeur par rapport à la somme des affinités de toutes les modalités de chaque trait. Ainsi, par exemple, si un trait de 4 modalités était pourvu des affinités 0, 1, 2 et 1, les affinités deviendraient 0%, 25%, 50% et 25%. On ne dit alors plus que beaucoup d'individus (affinité 2) d'une

FIGURE 3.5 – Treillis complet pour le contexte échelonné nominalement (limité à 21 objets et 10 traits).



espèce possèdent la troisième modalité, mais qu'il s'agit de la moitié²³ des individus de cette espèce. On situe alors immédiatement la tendance générale de l'espèce pour un trait. Ainsi si on applique cet exemple au trait "taille potentielle" on pourrait comprendre immédiatement que l'espèce concernée est plutôt grande.

Par ailleurs, on peut considérer cette répartition comme une distribution de probabilités, nous rapportant aux contextes flous.

Contexte multi-valué flou

La structure de nos données peut être apparentée aux contextes multi-valués de par les traits et leurs modalités ainsi qu'aux contextes flous de par les affinités (section 2.3, page 38). Nous introduisons donc la notion de contexte multi-valué flou car nous sommes dans le cas où $(o, t, m) \in I$ et $(o, t, n) \in I$ n'impliquent pas forcément $m = n$: les modalités m et n du trait t appartiennent à l'objet o avec différents degrés.

Définition 30 *Un contexte multi-valué flou K est un quintuplet (O, T, M, A, I) où O, T, M et I sont définis tels que pour un contexte multi-valué (partie 2.3.1) : O est un ensemble d'objets (les espèces), T est un ensemble d'attributs (les traits), M est un ensemble de modalités. A est un ensemble d'affinités et correspond à un ensemble de degrés de vérité (ou d'appartenance) tels que définis pour les contextes flous (partie 2.3.3). La notation $I(o, t, m) \in A$ ou $I(o, t, m) = a$ signifie que l'objet o a l'affinité a pour la modalité m du trait t .*

Echelonnage histogramme

Nous manipulons des données complexes qui ne sont pas utilisables en l'état par les algorithmes de construction de treillis. Nous devons nous ramener à une table binaire. Utiliser une *échelle nominale* [37] (table disjonctive totale) est évident, cependant les résultats ne satisfont pas les biologistes parce que les modalités d'un même trait ont été éparpillées dans plusieurs propriétés, faisant perdre la vision globale de la répartition de l'espèce dans les différentes modalités. C'est pourquoi nous proposons une *échelle histogramme*. Les différentes affinités d'un même trait sont concaténées en un seul attribut multi-valué et un contexte multi-valué (non flou) est ainsi construit, sur lequel on applique un échelonnage nominal. La conversion complète est un *échelonnage histogramme*.

Définition 31 *Soit $K := (O, T, M, A, I)$ un contexte multi-valué flou. Pour chaque trait $t \in T$ ayant m_t ²⁴ modalités, on associe des traits-histogrammes $h_t \in H_t$ ²⁵ avec $|H_t| \leq$*

23. Le faible écart de valeur entre les affinités implique que 50% est une estimation.

24. m_t est le nombre de modalités du trait t

25. h_t est un histogramme représentant une seule instance du trait t . H_t est l'ensemble des histogrammes du trait t , et H est l'ensemble de tous les histogrammes de tous les traits.

$|A|^{m_t}$ (appelés plus simplement histogrammes). Un histogramme $h_t \in H_t$ est tel que $h_t = \{a_1, \dots, a_i, \dots, a_{m_t}\}$ où $a_i \in A$. Soit $H = \cup_{t \in T} H_t$ l'ensemble de tous les histogrammes. Le contexte multi-valué flou K peut être représenté comme un contexte binaire (O, H, I_H) . Pour un objet $o \in O$ et un histogramme $h_t \in H$, $(o, h_t) \in I_H \Leftrightarrow I(o, t, m_i) = a_i$ pour tout $i \in [1, m_t]$.

Un histogramme $h_t \in H$ est composé d'une lettre qui qualifie le trait t considéré (par exemple V pour "reproduction végétative"), et des m_t affinités de l'objet considéré pour les m_t modalités de t . La figure 3.6 illustre les modalités de la taille potentielle des espèces CALO (*Callitriche obtusangula*) et NASO (*Nasturtium officinale*)

- Les deux espèces n'ont aucune plante d'une taille potentielle inférieure à 8 centimètres.
- 20% des plantes de l'espèce CALO et 50% de NASO mesurent entre 8 et 30 cm.
- CALO possède 40% d'affinité avec la troisième modalité (entre 30 cm et 1 m), contre 50% pour NASO
- Enfin 40% des plantes CALO mesurent plus d'un mètre alors que NASO n'a aucune affinité avec cette dimension.

Ainsi leurs histogrammes "taille potentielle" sont T0-20-40-40 pour CALO et T0-50-50-0 pour NASO. La figure 3.6 illustre ces histogrammes qui permettent de se rendre compte immédiatement des différences ou similitude entre les espèces.

FIGURE 3.6 – Exemples d'histogrammes pour le trait "taille potentielle".

	0	20	40	40
		CALO		
	0	50	50	0
		NASO		
< 0,08 m				
0,08-0,3 m				
0,3-1 m				
1-5 m				

Treillis d'histogrammes

Le contexte issu de l'*échelonnage histogramme* est basé sur une relation binaire I_H qui associe pour chaque trait un unique $h_t \in H_t$ à chaque espèce puisque les histogrammes s'excluent mutuellement. Ainsi chaque espèce possède 10 histogrammes, i.e. un par trait. La table 3.2 présente le trait "taille potentielle" (pour un sous-ensemble de 15 espèces) du nouveau contexte obtenu par l'*échelonnage histogramme*. Le contexte intégral est composé de 124 histogrammes.

	T0-20-40-40	T0-50-50-0	T0-0-100-0	T0-33-66-0	T0-0-33-66
CALO	×				
CALP		×			
GROD				×	
HIPV				×	
HOTP				×	
MYRV		×			
NASO		×			
NUPL					×
NYMA	×				
PTCO			×		
PTCR			×		
RANC	×				
RANU	×				
SAGS			×		
UTRV		×			

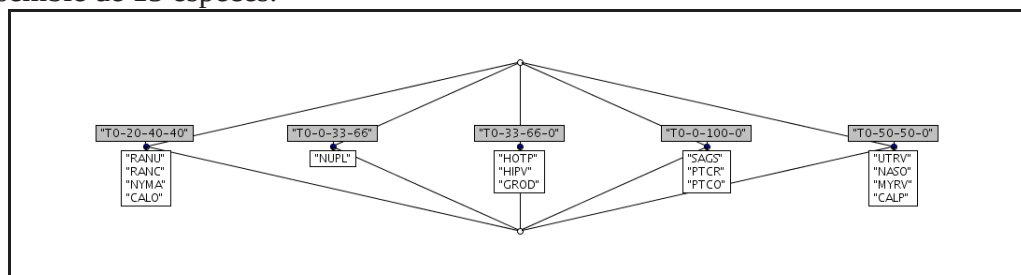
TABLE 3.2 – Histogrammes "taille potentielle" pour un sous-ensemble de 15 espèces.

La figure 3.7 présente le treillis correspondant au contexte de la table 3.2.

Par comparaison des treillis des figures 3.4 et 3.7, on constate que les histogrammes sont moins nombreux et concernent donc individuellement plus d'espèces. La lecture est aussi plus aisée. Prenons par exemple l'espèce NUPL. De la figure 3.4 on apprend qu'elle possède les propriétés T10, T20, T31 et T42. Comme toutes les autres espèces, elle possède 4 propriétés de taille potentielle (une pour chaque modalité). Dans le treillis de la figure 3.7 elle possède seulement l'histogramme T-0-0-33-66 qui les résume. On constate également rapidement la différence de taille et de lisibilité des deux treillis.

Le treillis complet, obtenu par une connexion de Galois ordinaire à partir du contexte d'histogrammes (O, H, I_H) est présenté par la figure 3.8 (47 macrophytes $\times$ 124 histogrammes).

FIGURE 3.7 – Treillis pour le trait histogramme de "taille potentielle" pour un sous ensemble de 15 espèces.



3.3 Comparaison et combinaison des treillis et des méthodes statistiques

Les biologistes travaillent classiquement avec des méthodes statistiques multivariées d'analyse de données. On trouve notamment les méthodes descriptives permettant de structurer et résumer l'information telles que l'analyse en composantes principales ou l'analyse factorielle des correspondances. Nous nous intéressons ici à l'analyse factorielle multiple (AFM) [31, 51] qui est basée sur l'analyse en composante principale (ACP). Cette analyse traite des tables de données binaires, telles que nous les obtenons par conversion de notre contexte, afin d'obtenir une représentation graphique permettant de voir les rapprochements et les oppositions entre les caractéristiques d'individus. Nous nous intéressons à évaluer l'intérêt des approches statistique et latticielle.

Nous commençons par présenter des résultats issus d'une AFM puis nous combinons et comparons les deux approches.

3.3.1 Analyse factorielle multiple

L'intérêt d'une table binaire est qu'elle permet l'analyse des données non seulement par la construction d'un treillis de Galois, mais également par l'application des méthodes statistiques habituellement utilisées par les hydrobiologistes. Le jeu de données binarisé a donc été analysé avec une AFM [31]. Cette méthode permet de comparer k groupes de variables définies sur le même ensemble d'individus et se base sur des projections mathématiques dans le but de visualiser les caractéristiques principales d'un jeu de données. Ces caractéristiques forment un nuage de points (i.e. un point par caractéristique) dans un espace de dimension d (avec $d < \text{nombre d'attributs}$) représentés par des axes factoriels qui sont des axes virtuels (aussi appelés dimensions) issus d'une synthèse entre les variables de l'analyse. Ces dimensions sont ordonnées par ordre décroissant de leur représentativité, i.e. leur inertie (variance) dont la somme (l'inertie totale) est égale à 100%. Le choix du nombre d'axes à prendre en compte dépend de la somme de leur

FIGURE 3.8 – Treillis de Galois pour le contexte histogramme complet (15 traits).

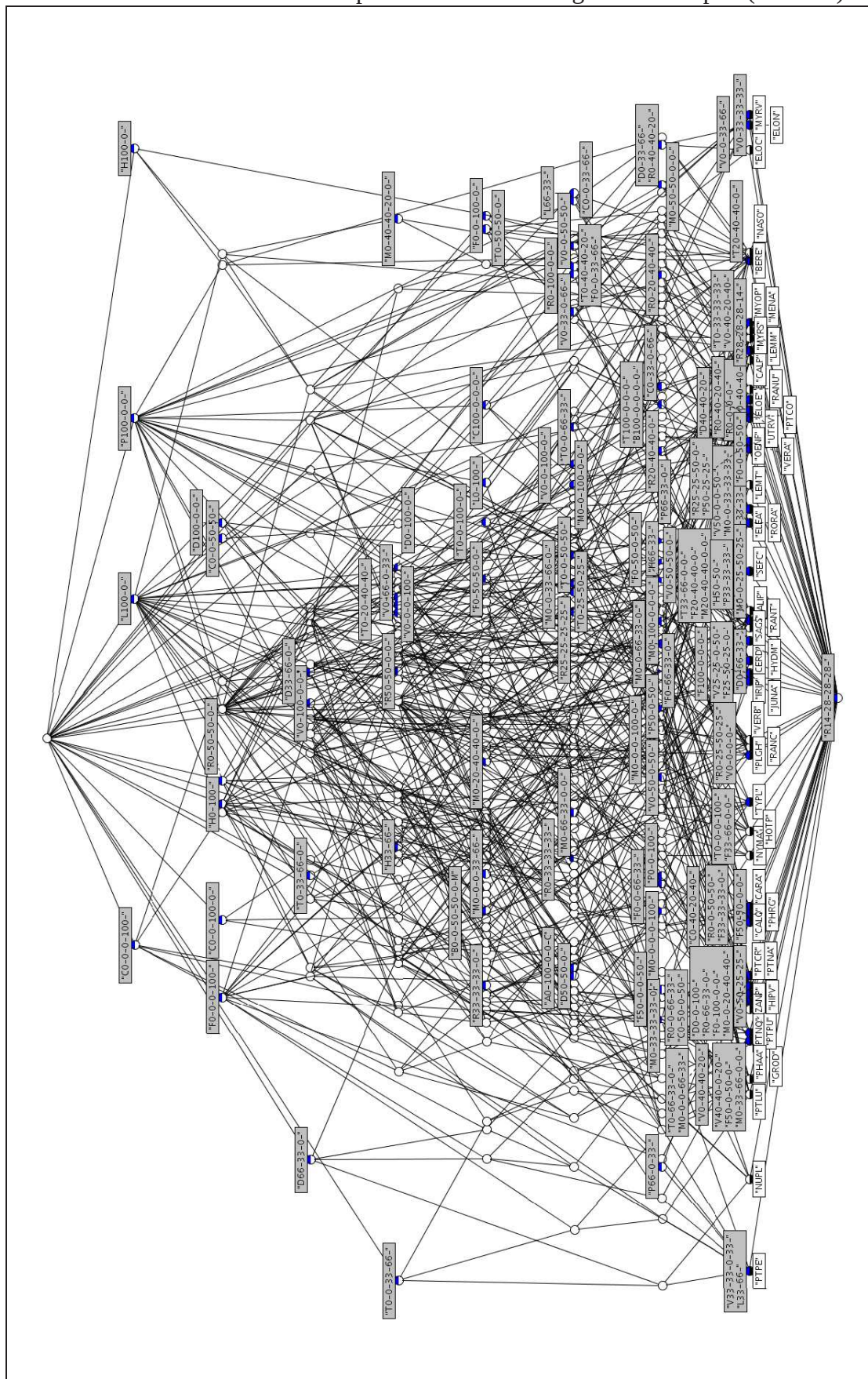
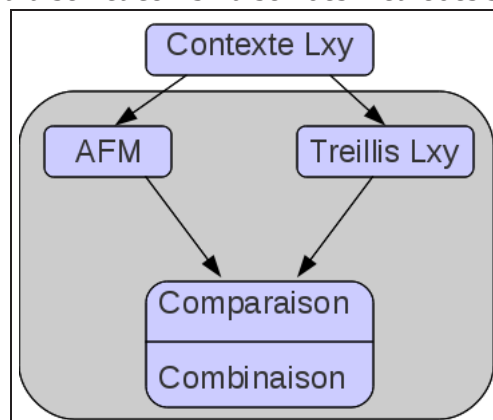


FIGURE 3.9 – Comparaison et combinaison des méthodes statistique et latticielle



inertie qui doit représenter une part importante de l'inertie totale (environ 60 à 80 %). Les points les plus intéressants à consulter sont ceux à la fois proches d'un des axes et loin de l'origine, car ils en sont les plus corrélés (les plus explicatifs). Cette corrélation est exprimée par une valeur allant de 0 (aucune corrélation) à 1 (très corrélé).

Nous analysons les trois figures²⁶ 3.10, 3.11 et 3.12. La première (3.10) présente les résultats obtenus en appliquant l'AFM sur les traits biologiques (nous y ferons référence par la suite sous le terme AFM_{traits}), la seconde (3.11) les résultats sur les propriétés de la table disjonctive totale (appelée $AFM_{propriétés}$), et enfin la troisième (3.12) présente les espèces du jeu de données (appelée $AFM_{espèces}$). Ces schémas nous donnent une représentation graphique des données, cependant le logiciel (R) nous fournit des données chiffrées telles que présentées dans les tableaux 3.3, 3.4, 3.5 et 3.6 nous permettant plus facilement de sélectionner les données.

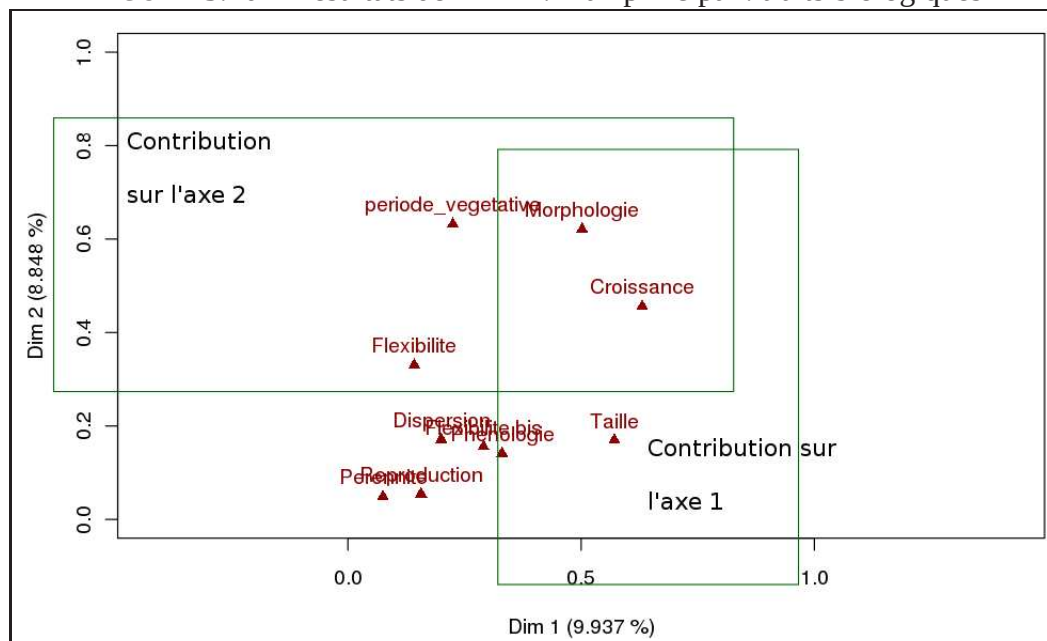
La lecture des trois figures repose sur le fait qu'elles présentent sur chacun de leurs axes des *dimensions* successives. Nous nous intéressons ici aux deux premières (les plus représentatives) des 46 dimensions de chacune des AFM. Nous pouvons cependant noter que la dimension 1 représente 9,937% de l'information globale et la seconde : 8,848%, ce qui souligne une faible contribution.

La figure 3.10 nous présente les traits *index morphologique* (Morphologie), *forme de croissance* (Croissance), *taille potentielle* (Taille) et *phénologie*⁽³⁹⁾ (Phénologie) comme étant les plus contributifs à la première dimension. Quant à la seconde, on y trouve les traits : *reproduction végétative* (periode_vegetative), *index morphologique* (Morphologie), *forme de croissance* (Croissance) et *flexibilité* (Flexibilité). Nous avons conservé les traits caractérisant 60% de l'information totale. Ces traits sont donc les plus influents sur les données.

La figure 3.11 présente les attributs de la table disjonctive totale. La lecture d'un

26. Il s'agit de la même analyse dont nous étudions les différentes conclusions qu'elle nous fournit.

FIGURE 3.10 – Résultats de l'AFM. Plan principal : traits biologiques



cercle de corrélation est très similaire à celle des plans factoriels : la projection sur les axes nous permet cette fois-ci de déterminer les propriétés les plus contributives pour chaque axe. Par ailleurs plus une propriété est proche du cercle plus elle est influente. On s'aperçoit ici que beaucoup de propriétés ont un vecteur court et sont donc peu pertinentes. On sélectionne les points (propriétés) les plus éloignés de l'axe considéré. Le cercle de corrélation est un graphe nous permettant d'obtenir des indications par une représentation visuelle.

Parallèlement à ce schéma, nous avons des données numériques telles que présentées par le tableau 3.3 indiquant des informations comme les coordonnées, les contributions, les $\cos^2$... Ce tableau nous permet d'extraire aisément, pour la première dimension (dans l'ordre d'importance de leur contribution), 9 propriétés sur la partie négative de l'axe (T30, C40, C12, M12, H22, L10, V40, T12 et L22) et 7 propriétés sur la partie positive (C42, C10, M10, T32, H20, T10 et H12).

Pour la seconde dimension (tableau 3.4) nous obtenons sur la partie négative 9 propriétés (V30, C10, M20, M52, V22, C32, F22, H10 et D30) et sur la partie positive 8 propriétés (V20, V32, M50, C12, M22, C30, F20 et F10).

Par recoupement avec la figure 3.10, on retrouve en effet les mêmes traits influents plus détaillés sur les modalités et affinités. En plus de ces traits déjà identifiés, on retrouve pour l'axe 1 des propriétés des traits *floraison* et *reproduction végétative*, et pour l'axe 2 des propriétés du trait *dispersion*. Ainsi les différents graphes peuvent nous renseigner à différents niveaux : soit au niveau des traits, soit des propriétés et donc

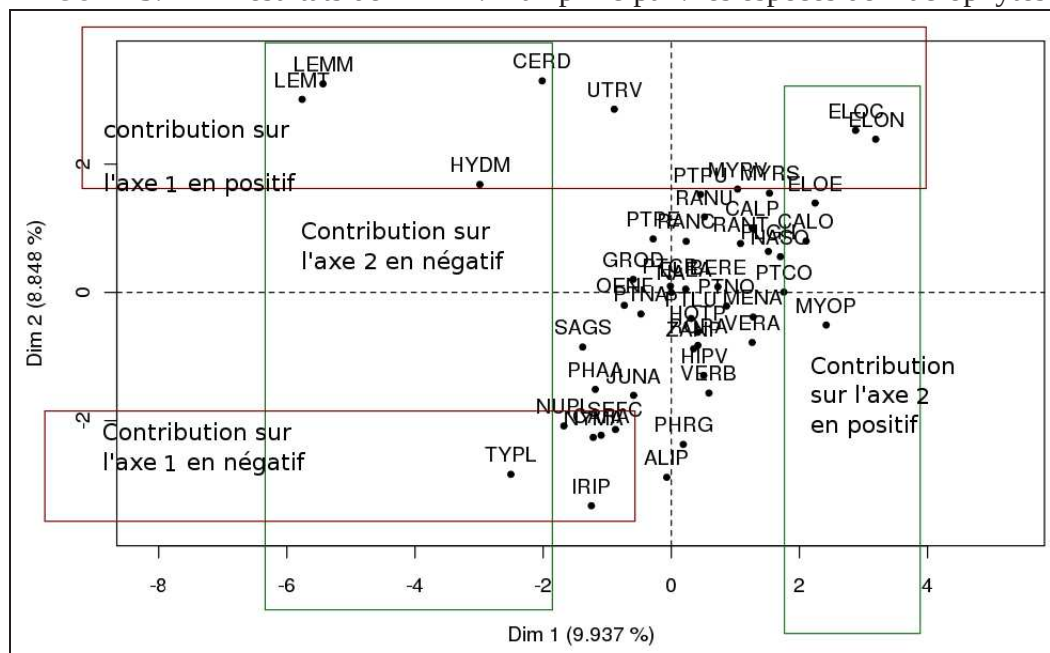
Propriétés	% contribution	% cumulé	coordonnées	Signification
T30	4,99	4,99	-0,71	Aucune plante avec une taille potentielle entre 0,3 et 1 mètre
C42	4,92	9,9	0,68	Beaucoup de plantes avec une forme de croissance ancrée, submerse
C40	4,92	14,82	-0,68	Aucune plante avec une forme de croissance ancrée, submerse
C10	4,65	19,47	0,66	Aucune plante avec une forme de croissance non ancrée
C12	4,65	24,12	-0,66	Beaucoup de plantes avec une forme de croissance non ancrée
M10	4,34	28,47	0,69	Aucune plante avec un index morphologique de 1
M12	3,97	32,44	-0,66	Beaucoup de plantes avec un index morphologique de 1
H22	3,82	36,26	-0,62	Beaucoup de plantes avec une phénologie seulement durant la saison de végétation
T32	3,35	39,61	0,58	Beaucoup de plantes avec une taille potentielle entre 0,3 et 1 mètre
L10	3,13	42,75	-0,56	Aucune plante avec une floraison (annuelle ou pas)
H20	2,97	45,72	0,54	Aucune plante avec une phénologie seulement durant la saison de végétation
V40	2,95	48,67	-0,52	Aucune plante avec une reproduction végétative par fragments non spécialisés
T12	2,81	51,48	-0,53	Beaucoup de plantes avec une taille potentielle inférieure à 0,08m
L22	2,78	54,26	-0,53	Beaucoup de plantes sans floraison (Rarement ou jamais)
T10	2,57	56,82	0,51	Aucune plante avec une taille potentielle inférieure à 0,08m
H12	2,16	58,98	0,46	Beaucoup de plantes avec une phénologie toute l'année

TABLE 3.3 – Propriétés les plus contributives à l'axe 1 de l' $AFM_{propriétés}$

Propriétés	% contribution	% cumulé	coordonnées	Signification
V30	5,3	5,3	-0,66	Aucune plante avec une reproduction végétative par bulbilles, turions ou apex dormants
V20	4,96	10,26	0,64	Aucune plante avec une reproduction végétative par rhizomes ou stolons
V32	4,47	14,73	0,61	Beaucoup de plantes avec une repro. végétative par bulbilles, turions ou apex dormants
M50	4,24	18,97	0,64	Aucune plante avec un index morphologique de 4
C12	3,99	22,96	0,58	Beaucoup de plantes avec une forme de croissance non ancrée
C10	3,99	26,95	-0,58	Aucune plante avec une forme de croissance non ancrée
M20	3,77	30,73	-0,61	Aucune plante avec un index morphologique de 1
M52	3,68	34,41	-0,6	Beaucoup de plantes avec un index morphologique de 4
V22	3,48	37,88	-0,54	Beaucoup de plantes avec une reproduction végétative par rhizomes ou stolons
C32	3,48	41,36	-0,54	Beaucoup de plantes avec une forme de croissance ancrée, avec tissus de soutien
M22	3,37	44,73	0,57	Beaucoup de plantes avec un index morphologique de 1
C30	2,88	47,61	0,49	Aucune plante avec une forme de croissance ancrée, avec tissus de soutien
F22	2,58	50,19	-0,54	Beaucoup de plantes avec une flexibilité faible ($> 10 - 45^\circ$)
F20	2,58	52,78	0,54	Aucune plante avec une flexibilité faible ($> 10 - 45^\circ$)
H10	2,13	54,91	-0,43	Aucune plante avec une phénologie toute l'année
F10	2	56,91	0,48	Aucune plante avec aucune flexibilité ($< 10^\circ$)
D30	1,93	58,84	-0,4	Aucune plante avec une dispersion potentielle faible

TABLE 3.4 – Propriétés les plus contributives à l'axe 2 de l' $AFM_{propriétés}$

FIGURE 3.12 – Résultats de l'AFM. Plan principal : les espèces de macrophytes



Espèces	% contribution	% cumulé	coordonnées	nom latin
LEMT	22,14	22,14	-5,7	Lemna trisulca
LEMM	19,71	41,85	-5,38	Lemna minor
ELON	6,79	48,63	3,16	Elodea nuttallii
HYDM	5,96	54,59	-2,96	Hydrocharis morsus-ranae
ELOC	5,51	60,1	2,85	Elodea canadensis
TYPL	4,18	64,28	-2,48	Typha latifolia
MYOP	3,9	68,18	2,39	Myosotis palustris (= scorpioides)
ELOE	3,36	71,54	2,22	Elodea ernstae
CALO	2,95	74,49	2,08	Callitriche obtusangula
CERD	2,71	77,2	-1,99	Ceratophyllum demersum
PTCO	2,06	79,25	1,74	Potamogeton coloratus

TABLE 3.5 – Espèces les plus contributives à l'axe 1 de l'AFM_{espèces}

Espèces	% contribution	% cumulé	coordonnées	nom latin
IRIP	8,28	8,28	-3,29	Iris pseudacorus
CERD	8,15	16,43	3,26	Ceratophyllum demersum
LEMM	7,92	24,35	3,22	Lemna minor
LEMT	6,78	31,14	2,98	Lemna trisulca
ALIP	6,22	37,36	-2,85	Alisma plantago-aquatica
UTRV	6,11	43,47	2,83	Utricularia vulgaris
TYPL	6,02	49,49	-2,81	Typha latifolia
ELOC	4,78	54,27	2,5	Elodea canadensis
ELON	4,27	58,54	2,36	Elodea nuttallii
PHRG	4,21	62,75	-2,34	Phragmites australis
NYMA	3,83	66,58	-2,24	Nymphaea alba
CARA	3,71	70,29	-2,2	Carex acutiformis
SEFC	3,42	73,71	-2,11	Sparganium emersum
NUPL	3,25	76,96	-2,06	Nuphar lutea
HYDM	2,12	79,08	1,67	Hydrocharis morsus-ranae

TABLE 3.6 – Espèces les plus contributives à l'axe 2 de l' $AFM_{especies}$

- Coordonnées négatives : LEMT, LEMM, HYDM, TYPL et CERD sont concernées par les traits "taille potentielle" (T30 et T12), "forme de croissance" (C40 et C12), "index morphologique" (M12), "phénologie" (H22), "floraison" (L10 et L22) et "reproduction végétative" (V40).
- Coordonnées positives : ELON, ELOC, MYOP, ELOE, CALO et PTCO sont concernées par les traits "taille potentielle" (T32 et T10), "forme de croissance" (C42 et C10), "index morphologique" (M10), "phénologie" (H20 et H12).
- Axe 2 des $AFM_{especies}$ et $AFM_{proprietes}$:
 - Coordonnées négatives : IRIP, ALIP, TYPL, PHRG, NYMA, CARA, SEFC et NUPL sont concernées par les traits "reproduction végétative" (V30 et V22), "index morphologique" (M20 et M52), "forme de croissance" (C10 et C32), "flexibilité" (F22), "dispersion" (D30) et "phénologie" (H10)
 - Coordonnées positives : HYDM, ELON, ELOC, UTRV, LEMT, LEMM et CERD sont concernées par les traits "reproduction végétative" (V20 et V32), "index morphologique" (M50 et M22), "forme de croissance" (C12 et C30) et "flexibilité" (F20 et F10).

Les axes de ces analyses sont construits sur l'opposition des variables. En effet plus les variables sont éloignées l'une de l'autre par rapport à l'origine de l'axe : donc les plus extrêmes en positif et les plus extrêmes en négatif, plus cela signifie que leur informations s'opposent. Il est intéressant de noter ici, que nous retrouvons cette informa-

tion, non pas uniquement grâce au seul positionnement des propriétés sur les axes mais également dans leur signification. On peut noter que sur chacun des axes se retrouvent très souvent en opposition les mêmes traits avec les mêmes modalités, mais lorsqu'en positif l'affinité stipule "beaucoup de plantes", en négatif elle précise "aucune plante", et inversement. Par exemple sur l'axe 1 avec en négatif T30 et T12 et en positif T32 et T10. La cohérence entre les propriétés en opposition nous conforte dans les choix qui ont été faits sur le jeu de données.

Ces associations entre les espèces et les traits sont issues de l'analyse, mais les limites auxquelles nous arrêter dans la lecture de cette analyse restent floues : 60% ou 80% sont des notions arbitraires. Nous les avons choisies car elles semblaient les plus pertinentes aux vues des données. Nous nous intéressons à savoir si les treillis de Galois confirmeront, infirmeront ou compléteront ces choix. Pour cela nous comparons les associations listées ci-dessus avec les conclusions d'un treillis.

3.3.2 Apport des treillis

Dans le cas de notre contexte binarisé, 47 espèces et 105 propriétés engendrent un treillis de 190 964 concepts. La représentation graphique de ce treillis est trop dense pour être exploitée, et la quantité d'informations portée par ce treillis est elle-même trop importante [12]. Un tel problème est souvent soulevé, notamment dans [42] où est proposé un calcul de χ^2 permettant de tester l'indépendance de deux variables qualitatives. Dans notre cas, afin de choisir les attributs multi-valués les plus pertinents sur lesquels construire le treillis, nous utilisons les résultats obtenus par l'AFM pour sélectionner des sous-ensembles de plantes ou traits *a priori* intéressants.

Afin de comparer les approches statistique et laticielle, nous :

1. réduisons notre contexte aux espèces et propriétés de l'axe 1 de l'AFM (depuis les graphes $AFM_{especes}$ et $AFM_{proprietes}$)
2. construisons le treillis de ce contexte
3. analysons le treillis par rapport aux conclusions établies sur l'axe 1 de l'AFM.

Puis nous reproduisons cette démarche pour l'axe 2.

Treillis sur l'axe 1

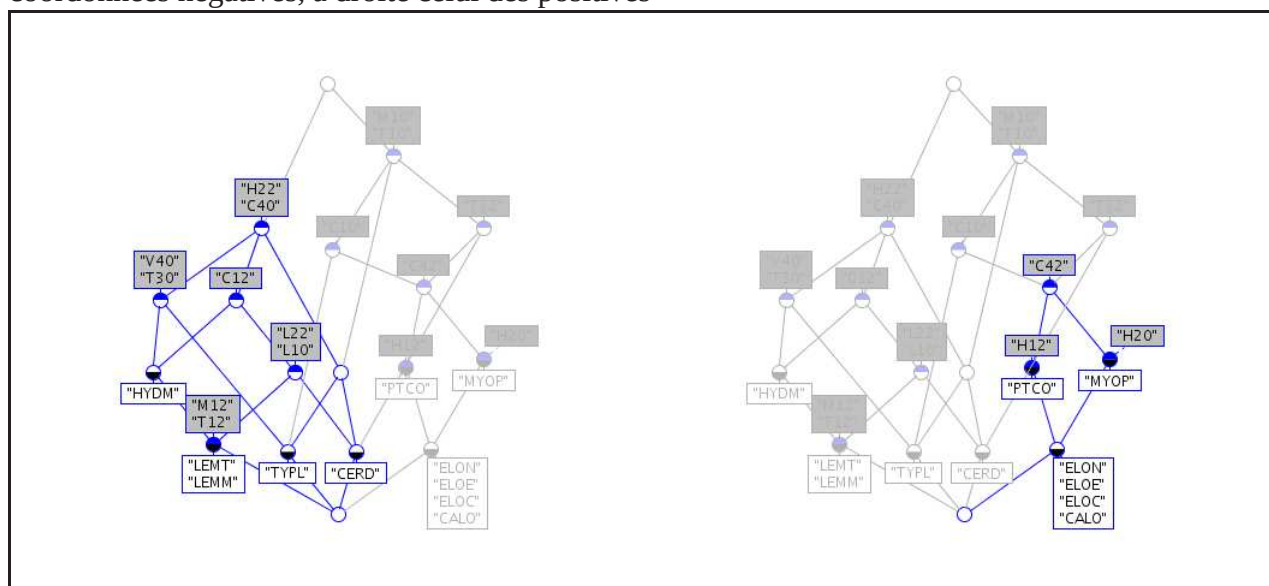
La table 3.7 présente le contexte réduit aux espèces de macrophytes et aux propriétés extraits de l'axe 1.

Le treillis 3.13 issu de ce contexte de 11 objets et 16 propriétés contient 18 concepts. On constate qu'il se divise en deux sous-treillis correspondant aux coordonnées négatives et positives de l'axe 1.

Espèces	T10	T12	T30	T32	C10	C12	C40	C42	M10	M12	H12	H20	H22	L10	L22	V40
CALO	1	0	0	1	1	0	0	1	1	0	1	1	0	0	0	0
CERD	1	0	0	1	0	1	1	0	1	0	0	0	1	1	1	0
ELOC	1	0	0	1	1	0	0	1	1	0	1	1	0	0	0	0
ELOE	1	0	0	1	1	0	0	1	1	0	1	1	0	0	0	0
ELON	1	0	0	1	1	0	0	1	1	0	1	1	0	0	0	0
HYDM	0	0	1	0	0	1	1	0	0	0	0	0	1	0	0	1
LEMM	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1
LEMT	0	1	1	0	0	1	1	0	0	1	0	0	1	1	1	1
MYOP	1	0	0	1	1	0	0	1	1	0	0	1	0	0	0	0
PTCO	1	0	0	1	1	0	0	1	1	0	1	0	0	0	0	0
TYPL	1	0	1	0	1	0	1	0	1	0	0	0	1	0	0	1

TABLE 3.7 – Contexte réduit aux données issues de l'axe 1

FIGURE 3.13 – Treillis du contexte de l'axe 1. A gauche le sous-treillis des plantes à coordonnées négatives, à droite celui des positives

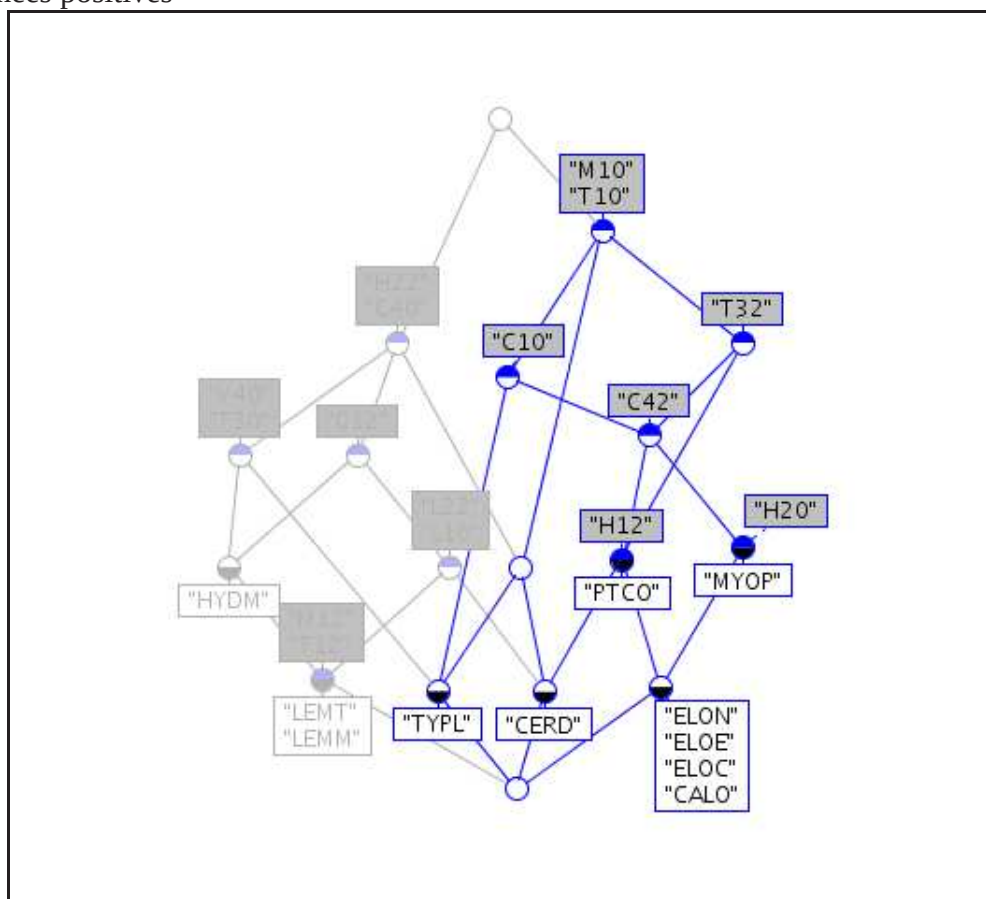


Le sous-treillis contenant les plantes à coordonnées négatives contient toutes les plantes et propriétés de coordonnées négatives et uniquement elles. Le treillis confirme donc exactement l'analyse statistique.

Nous retrouvons de même un sous-treillis contenant toutes les plantes à coordonnées positives. Là encore il existe avec toutes les plantes attendues et uniquement celles-ci. Cependant il manque les propriétés T32, T10, C10 et M10. La figure 3.14 présente le sous-treillis qui les inclut. Cependant l'inclusion de ces propriétés implique la présence des espèces TYPL et CERD qui sont à coordonnées négatives (et fortement contributives comme indiqué dans le tableau 3.6).

Le treillis nous permet ici de cerner plus précisément quelles sont les propriétés qui sont uniquement associées aux plantes "positives", et celles qui sont partagées entre les deux groupes de plantes. Ceci est une information complémentaire à l'AFM car dans sa représentation graphique, un point (espèce) ne peut se trouver fortement contributif en même temps en coordonnées positives et négatives.

FIGURE 3.14 – Concepts du treillis du contexte de l'axe 1 avec les propriétés à coordonnées positives



Treillis sur l'axe 2

Considérons maintenant le contexte réduit aux plantes et propriétés issues de l'axe 2 de l'AFM (la table 3.8).

Espèce	C10	C12	C30	C32	M20	M22	M50	M52	F10	F20	F22	H10	V20	V22	V30	V32	D30
ALIP	1	0	0	1	1	0	0	1	1	0	1	1	0	1	1	0	1
CARA	1	0	0	1	1	0	1	0	0	0	1	1	0	1	1	0	1
CERD	0	1	1	0	0	1	1	0	1	1	0	1	1	0	0	1	0
ELOC	1	0	1	0	0	1	1	0	1	1	0	0	1	0	0	1	0
ELON	1	0	1	0	0	1	1	0	1	1	0	0	1	0	0	0	0
HYDM	0	1	1	0	0	1	1	0	0	0	1	1	0	1	0	1	1
IRIP	1	0	0	1	1	0	0	1	0	0	1	1	0	0	1	0	1
LEMM	0	1	1	0	1	0	1	0	1	1	0	0	1	0	0	1	1
LEMT	0	1	1	0	1	0	1	0	1	1	0	0	1	0	0	1	1
NUPL	1	0	1	0	1	0	0	1	0	1	0	1	0	1	1	0	1
NYMA	1	0	1	0	1	0	0	1	0	1	0	1	0	1	1	0	1
PHRG	1	0	0	1	1	0	0	1	0	0	1	0	0	1	1	0	1
SEFC	1	0	1	0	1	0	0	0	1	0	1	1	0	1	1	0	1
TYPL	1	0	0	1	1	0	0	1	0	0	1	1	0	1	1	0	1
UTRV	0	1	1	0	0	1	1	0	1	1	0	1	1	0	0	1	0

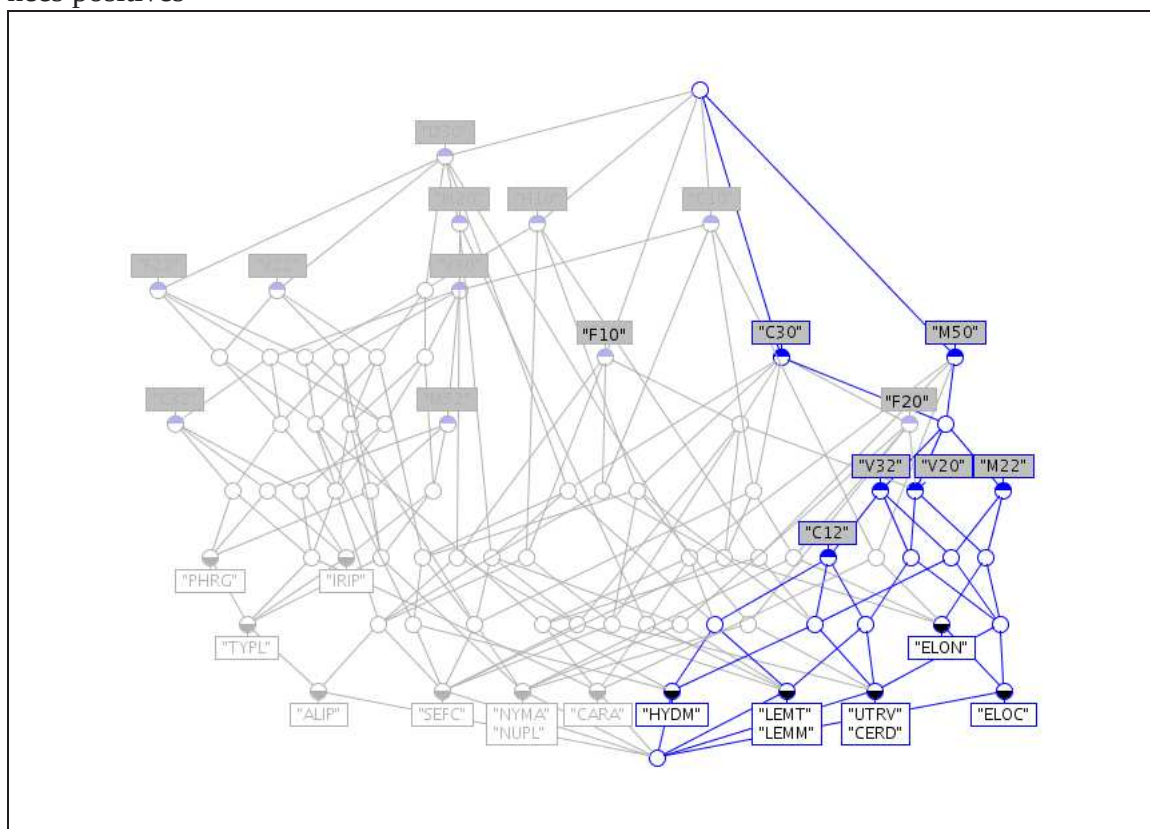
TABLE 3.8 – Contexte réduit aux données issues de l'axe 2

La petitesse du premier treillis 3.13 pourrait expliquer les résultats obtenus puisque le nombre de concepts reste assez faible. Le treillis issu du contexte de l'axe 2 de 15 objets et 17 propriétés contient 81 concepts. La figure 3.15 présente le sous-treillis contenant toutes les plantes à coordonnées positives de l'axe 2. Toutes les propriétés attendues sont présentes exceptées F20 et F10. La propriété F20 qui est partagée par toutes les plantes "positives" est aussi une propriété de SEFC, NYMA et NUPL qui sont des plantes "négatives". La propriété F10 est aussi partagée par toutes les plantes "positives", mais s'étend aux espèces "négatives" ALIP et SEFC.

Ainsi le treillis nous permet de confirmer les résultats statistiques mais également de les étendre. En effet, nous savons avec le treillis exactement quelles espèces et quelles propriétés sont associées et lesquelles ne le sont pas, une information que l'AFM ne fournit pas puisqu'elle permet d'opérer des regroupements sur la base de projections, "oubliant" une partie de l'information.

La figure 3.16 présente le sous-treillis contenant toutes les plantes à coordonnées négatives de l'axe 2. Là encore nous retrouvons toutes les propriétés attendues sauf V22, F22 et H10. La particularité ici, est que si l'on recherche les concepts intégrant ces propriétés, alors il faut sélectionner le treillis dans son intégralité. En effet, aucune

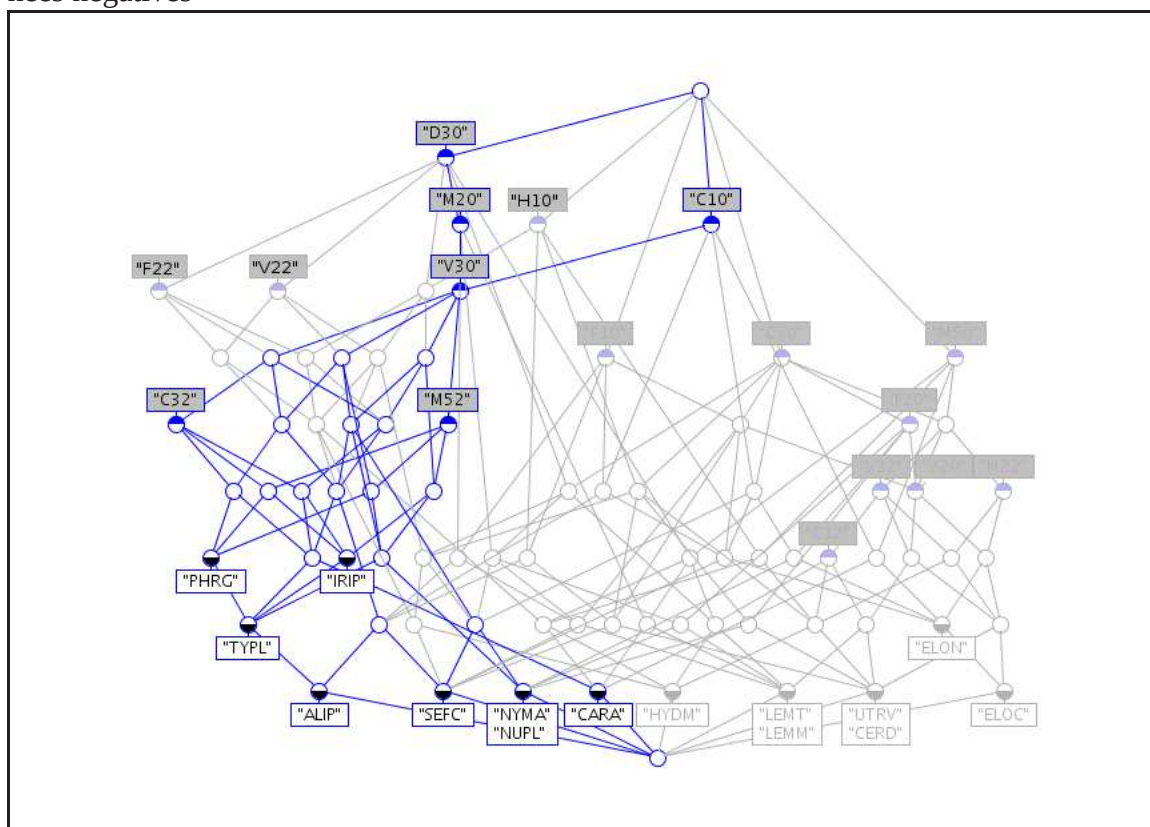
FIGURE 3.15 – Sous-treillis du treillis du contexte de l'axe 2 avec les plantes à coordonnées positives



de ces propriétés n'est partagée par toutes les plantes de ce sous-treillis. Ainsi V22 ne concerne pas IRIP, F22 ne concerne pas NUPL ni NYMA et enfin H10 ne concerne pas PHRG. En outre, ces des propriétés concernent des plantes "positives" : HYDM pour V22 et F22, et CERD, HYDM et UTRV pour H10.

Dans ce cas le treillis contredit l'analyse générale statistique qui associe les plantes et propriétés "positives" ensemble et "négatives" ensemble.

FIGURE 3.16 – Sous-treillis du treillis du contexte de l'axe 2 avec les plantes à coordonnées négatives



Cette étude montre que les treillis de Galois aboutissent à des conclusions qui rejoignent celles de l'analyse statistique et qu'ils sont des outils intéressants et pertinents. Cependant ils ont l'inconvénient d'être difficilement analysable lorsqu'il y a beaucoup de données à traiter. Ainsi pour se concentrer sur les données les plus pertinentes, nous faisons appel aux statistiques qui permettent de sélectionner ces données. La combinaison de ces deux méthodes permet donc d'un côté d'alléger les treillis pour une meilleure interprétation, et d'un autre coté d'obtenir des résultats plus précis que par les analyses statistiques pures.

3.4 Démarche complète d'identification de traits écologiques

Nous recherchons le moyen de nous abstraire des espèces pour déterminer l'état écologique d'un cours d'eau. Pour cela nous souhaitons utiliser des traits écologiques caractéristiques de l'environnement dans lequel les espèces évoluent, en nous basant sur les caractéristiques biologiques de ces espèces. Ainsi, en connaissant les relations entre les espèces et leurs traits biologiques d'une part, et les relations entre les espèces et leurs traits écologiques d'autre part, nous pourrions déterminer et ne conserver que les relations entre les traits biologiques et écologiques.

Pour cela, nous avons établi une démarche d'identification des traits écologiques qui est présentée par la figure 3.17 et qui se divise en 3 étapes : la conversion du contexte, l'analyse de concepts et la validation. La conversion du contexte se fait par échelonnage histogramme afin d'obtenir un contexte binaire à partir duquel nous construisons un treillis. Dans la seconde phase de cette méthode les concepts potentiellement intéressants sont sélectionnés par l'expert, puis ils sont interprétés en associant des espèces et des traits écologiques conformément aux connaissances de l'expert sur les milieux de vie (environnement) des espèces. Ensuite vient la troisième phase, celle de la validation des choix des traits écologiques. Pour cela les traits écologiques sont ajoutés au contexte initial puis un nouveau treillis est construit. Les concepts sont sélectionnés à partir du treillis étendu afin de retrouver les concepts ayant la même extension que ceux sélectionnés lors de l'étape précédente. Ainsi leurs intensions devraient être similaires en incluant les traits écologiques mentionnés par l'expert.

Nous développons ces étapes successives dans cette section.

3.4.1 Analyse des concepts formels

Dans cette partie, nous interprétons les concepts grâce aux connaissances d'experts sur l'environnement des espèces dans le but de sélectionner les traits écologiques qui relient les espèces et les caractéristiques de l'environnement où ils vivent. Nous présentons l'étape de la méthode (figure 3.18) permettant de sélectionner ces traits écologiques : le choix des concepts pertinents, puis leur analyse et les conclusions qui en découlent en terme de traits écologiques.

Sélection de concepts

Parmi les concepts obtenus, les concepts sur les couches centrales du treillis sont les plus intéressants pour les biologistes, i.e. ceux possédant entre 3 et 5 histogrammes, ce qui correspond à des extensions ayant entre 3 et 7 espèces. Les concepts ayant plus d'espèces sont trop généraux et ne peuvent être reliés à des conditions environnementales spécifiques, tandis que ceux avec moins de 3 espèces sont trop spécifiques et ne

FIGURE 3.17 – Méthode complète d'identification des traits écologiques

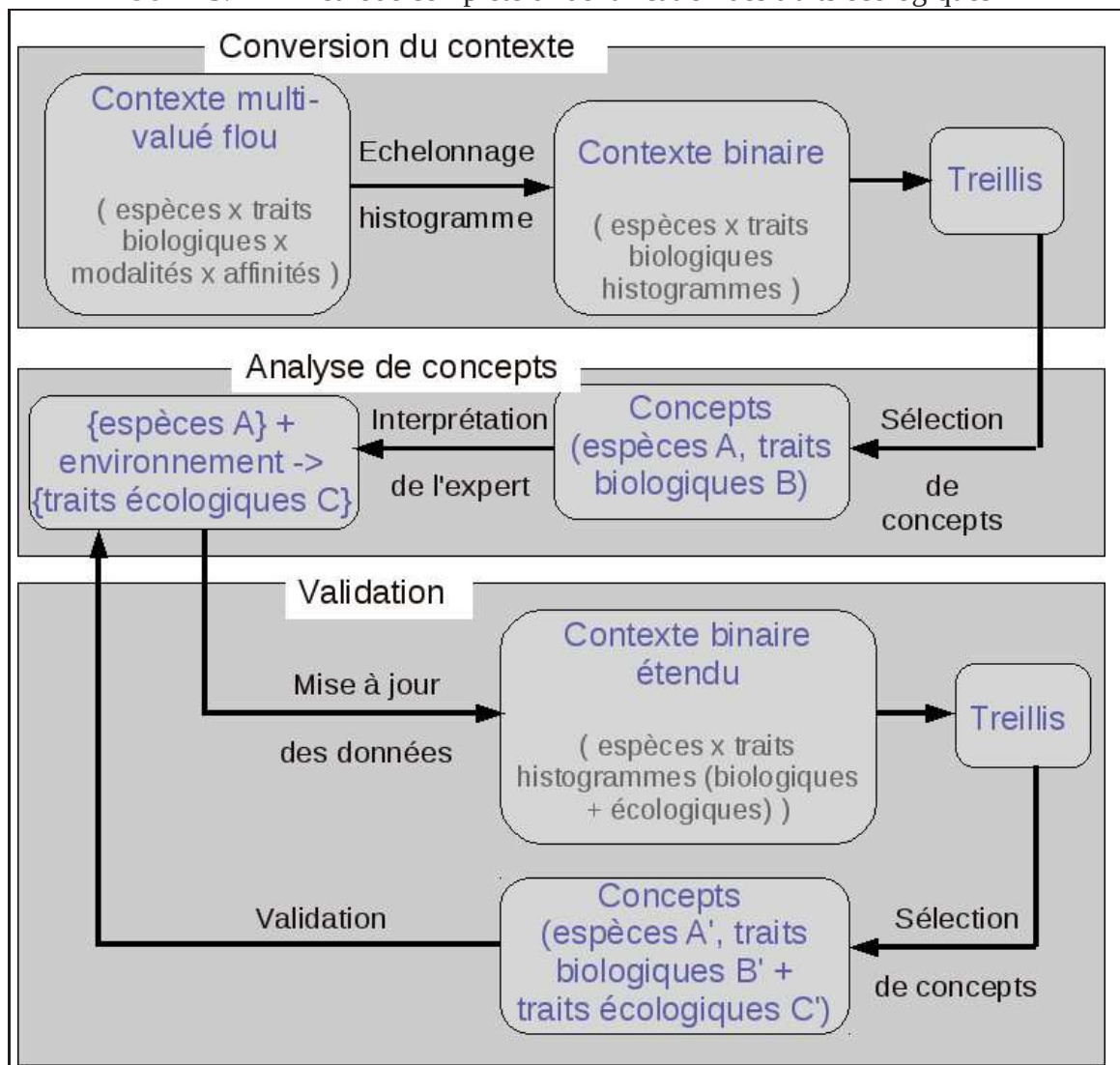
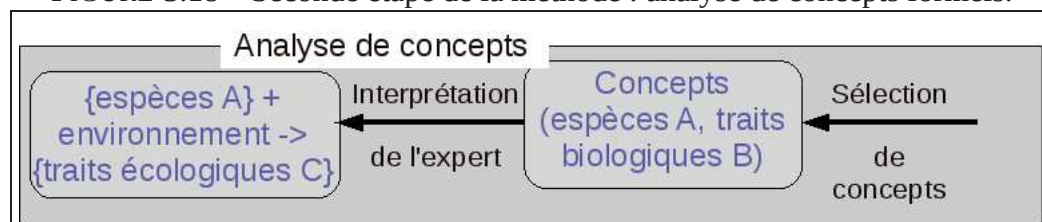


FIGURE 3.18 – Seconde étape de la méthode : analyse de concepts formels.



fournissent pas d'information pertinente. Prenons deux exemples parmi ces concepts "centraux" issus du treillis de la figure 3.8 :

- Concept 1 : (JUNA SEFC TYPL, L100-0 H0-100 P100-0-0 V0-100-0-0 D100-0-0). Les espèces considérées sont *Juncus articulatus*, *Sparganium emersum*, *Typha latifolia*. Les traits qu'elles partagent sont une floraison annuelle (L100-0), une phénologie⁽³⁹⁾ durant la période végétative seulement (H0-100), des organes pérennes (aériens ou souterrains) (P100-0-0), une reproduction végétative par rhizomes ou stolons (V0-100-0-0) et une grande dispersion (avec de petites graines volantes) (D100-0-0).
- Concept 2 : (CERD PHAA PTCR PTLU PTNA UTRV, H0-100 F0-0-0-100 P100-0-0). Ce concept concerne les macrophytes *Ceratophyllum demersum*, *Phalaris arundinacea*, *Potamogeton crispus*, *Potamogeton lucens*, *Potamogeton natans*, *Utricularia vulgaris*. Leurs propriétés communes sont une phénologie⁽³⁹⁾ pendant la période végétative seulement (H0-100), une grande flexibilité (> 300°) (F0-0-0-100) et des organes pérennes (aériens ou souterrains) (P100-0-0).

Interprétation de l'expert : des traits biologiques aux traits écologiques

Grâce au treillis, nous obtenons des concepts associant les traits biologiques et les espèces, et par ailleurs les biologistes nous informent des relations entre les espèces et les caractéristiques des milieux dans lesquels les espèces vivent. Pour les deux concepts précédents, les experts indiquent les liens entre espèces et milieux et associent des traits écologiques :

- Concept 1 : ces espèces vivent dans des eaux mésotrophes⁽³¹⁾ à eutrophes⁽¹⁷⁾. Ce qui souligne l'intérêt du trait *tolérance à la trophie*.
- Concept 2 : ces espèces vivent dans des eaux calmes. Ce qui souligne l'intérêt du trait *tolérance aux variations d'humidité*.

Ces conclusions attirent notre attention sur la pertinence de traits écologiques décrivant les caractères de tolérance des espèces au niveau trophique et aux inondations.

Dans la littérature, il existe de nombreux traits écologiques portant tous une information pertinente biologiquement, mais pas forcément pour notre étude. Notre approche permet donc de sélectionner les plus adéquats (ainsi que leurs modalités) afin de décrire les environnements des plantes de la plaine d'Alsace. Cette approche part des informations sans a priori que nous fournissent les macrophytes et leurs propriétés biologiques. Ce sont les regroupements intrinsèques aux données qui sont mis en évidence par la méthode des treillis, et permettent donc de faire émerger certains traits écologiques et d'en ignorer d'autres. L'expertise intervient ultérieurement pour interpréter la connaissance inhérentes aux données dite "guidée par les données".

Actuellement à partir des caractéristiques communes indiquées par les concepts, 5 traits écologiques émergent comme étant utiles à l'évaluation de l'état écologique de l'eau parmi tous ceux présents dans la bibliographie. Le tableau 3.9 présente ces 5 traits

Traits	Stabilité de l'eau (i.e. variation du niveau de l'eau)	Résistance aux inondations	Tolérance aux matières organiques	Tolérance à la sédimentation (dépôt ou accumulation de matière organique ou minérale, probabilité d'enterrement et demande d'ajustement de l'enracinement)	Niveau trophique (importance du milieu en nutriments)
Modalités	stable fluctuations occasionnelle	aucune tolérance tolérance faible tolérance moyenne tolérance forte	<10% 10-40% >40%	enracinement faible enracinement moyen enracinement fort	oligotrophe mésotrophe eutrophe hypertrophe

TABLE 3.9 – Traits écologiques sélectionnés et leurs modalités.

et leurs modalités.

3.4.2 Validation

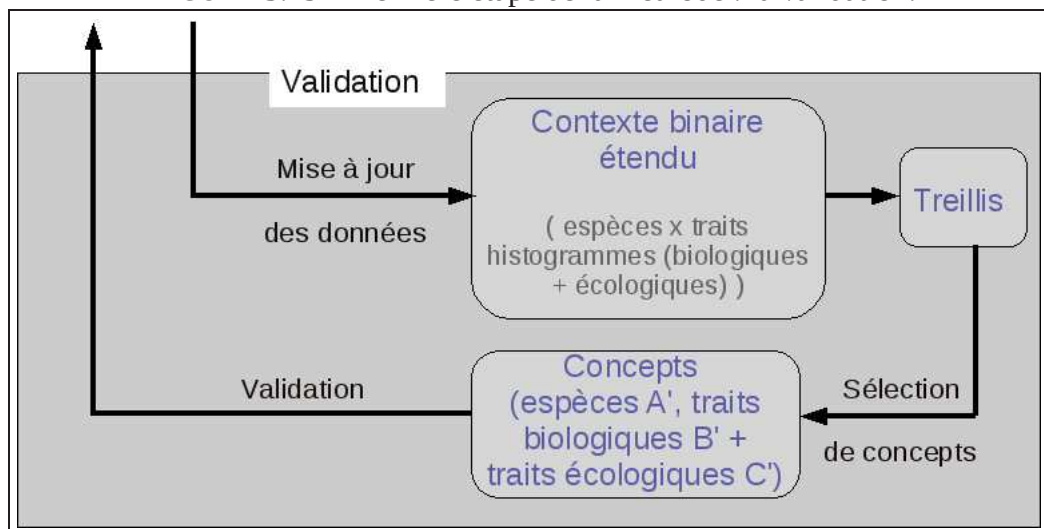
Les 5 traits écologiques mis en évidence par les deux premières étapes de la méthode ont besoin d'être validés comme des choix pertinents. Cette validation s'opère dans la troisième phase de la démarche telle que décrite dans cette partie et présentée par la figure 3.19.

Intégration des traits écologiques dans le jeu de données

Les concepts ({macrophytes},{traits biologiques}) ont mis en évidence des traits écologiques. Ces traits correspondent donc aux macrophytes du concept qui l'a révélé. Cependant, lorsque nous intégrons ces traits écologiques dans le jeu de données, nous les renseignons pour tous les macrophytes et pas uniquement ceux qui les ont révélés et vont donc tous être concernés par tous les traits écologiques. Ces traits écologiques et leurs modalités sont ajoutés au jeu de données et sont renseignés pour les différentes espèces en terme d'affinités selon [69]. Jusqu'à présent 47 espèces ont été pleinement complétées.

Un treillis issu d'un sous-ensemble de 25 espèces (pour des questions de lisibilité)

FIGURE 3.19 – Dernière étape de la méthode : la validation.



du contexte étendu (comportant les traits biologiques et écologiques) est présenté par la figure 3.20. Le concept mis en évidence dans le treillis correspond au concept 2 présenté plus tôt.

Validation de la méthode

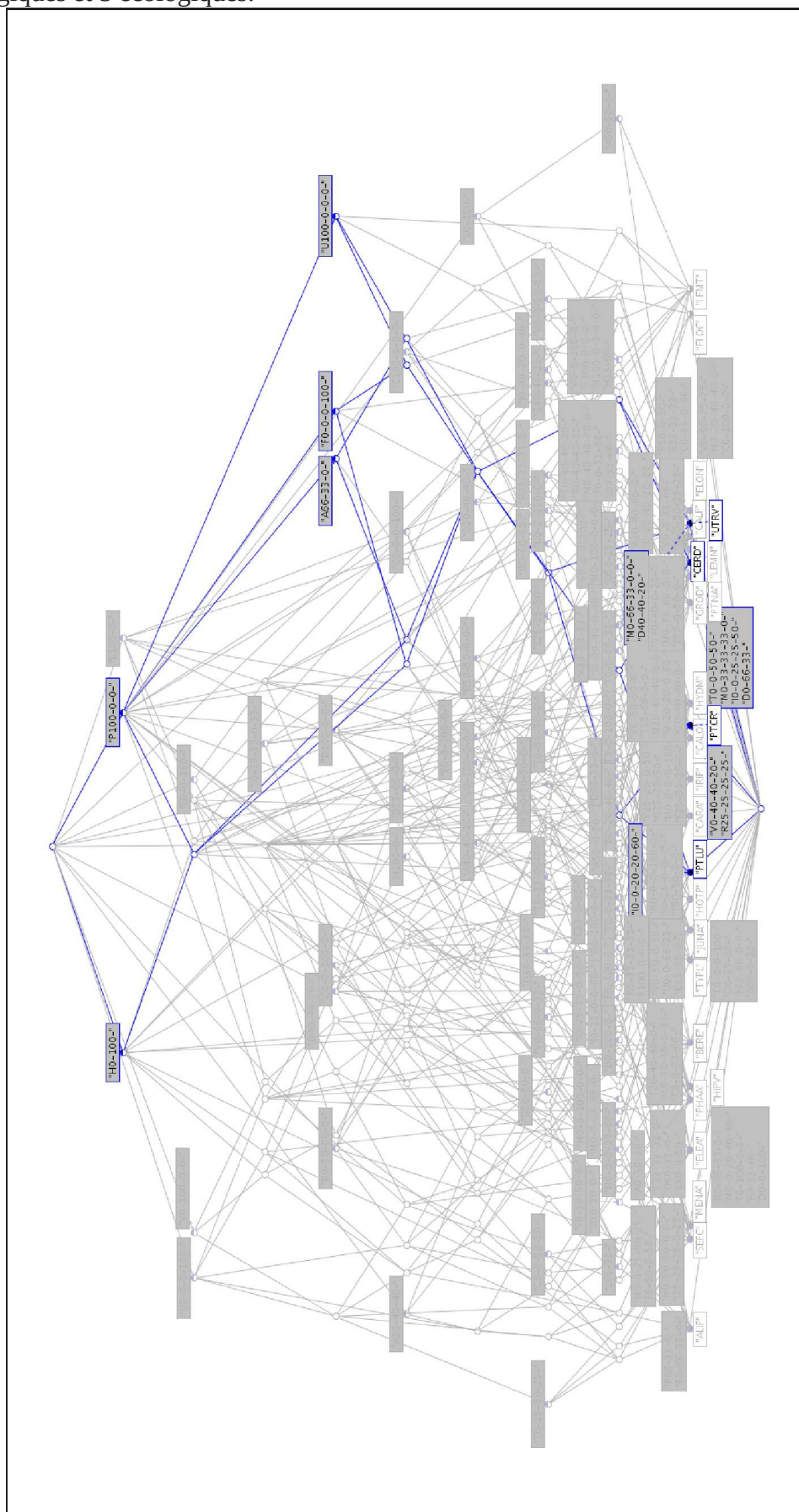
Nous disposons de sources d'information. La première vient du treillis et concerne les relations entre les espèces et les traits biologiques. La seconde provient des spécialistes et concerne les relations entre les espèces et les traits écologiques.

Dans le but de valider cette démarche, nous cherchons à vérifier le raisonnement suivant. En identifiant les mêmes espèces venant de chacune de ces deux sources d'information, nous pouvons relier directement les traits biologiques de la première source et les traits écologiques de la deuxième, et faire abstraction de leurs espèces communes. Reprenons nos deux exemples :

- Concept 1 : on peut mettre en relation une floraison annuelle, une phénologie durant la période végétative seulement, des organes pérennes, une reproduction végétative par rhizomes ou stolons et une haute dispersion (traits biologiques issus du treillis) avec des eaux mésotrophes à eutrophes (trait écologique issu de l'expertise).
- Concept 2 : on peut mettre en relation une phénologie durant la période végétative seulement, une haute dispersion et des organes pérennes (traits biologiques issus du treillis) avec le besoin de vivre en eaux calmes (trait écologique issu de l'expertise).

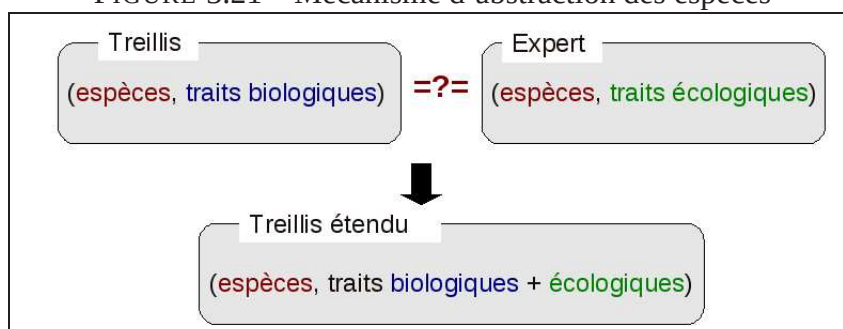
Nous analysons donc les concepts issus du treillis étendu pour vérifier les relations

FIGURE 3.20 – Treillis de Galois pour un sous-ensemble de 25 macrophytes, 10 traits biologiques et 5 écologiques.



entre les traits biologiques et écologiques telles qu'illustrées par le schéma 3.21. Si nous avons les mêmes espèces (en rouge) provenant du treillis et de l'expert, nous devrions les retrouver dans le treillis étendu dans un concept qui regroupe les mêmes traits biologiques (en bleu) du treillis et les mêmes traits écologiques (en vert) de l'expert à la manière d'une factorisation sur les espèces.

FIGURE 3.21 – Mécanisme d'abstraction des espèces



Application

Nous illustrons cette étape à l'aide des deux concepts déjà présentés :

- Concept 1 : (JUNA SEFC TYPL, L100-0 H0-100 P100-0-0 V0-100-0-0 D100-0-0). L'expertise pour ce concept indique que ces espèces vivent dans un milieu mésotrophe à eutrophe. Dans le treillis étendu aux traits écologiques, nous retrouvons ce concept avec l'intension incluant les histogrammes : I0-50-50-0 et E0-0-100, confirmant exactement l'expertise.
 - l'histogramme I0-50-50-0 signifie que les individus de chacune des espèces sont équitablement répartis entre des besoins en eaux mésotrophes et eutrophes.
 - Le treillis indique que ces espèces partagent également l'histogramme E0-0-100 qui signifie que ces espèces ont une implantation constante. Cette information est révélée exclusivement par le treillis.
- Concept 2 : (CERD PHAA PTCR PTLU PTNA UTRV, H0-100 F0-0-0-100 P100-0-0). L'expertise de ce concept établit que ces espèces vivent dans des eaux calmes. En observant le treillis étendu on constate que ces 6 espèces ne partagent pas les mêmes traits écologiques et se retrouvent donc dans des concepts séparés. Ainsi seulement 4 des 6 espèces partagent les attributs A66-33-0 et U100-0-0-0, comme le montre la figure 3.20.
 - L'attribut A66-33-0 signifie que chaque espèce a 66% de ses plantes qui vivent dans des eaux stables et 33% qui vivent dans des eaux ayant des fluctuations. Ce trait correspond à l'expertise (vivre dans des eaux calmes) pour les 4 espèces CERD, PTCR, PTLU et UTRV.

PTNA a l'attribut A40-40-20 indiquant qu'elle vit principalement en eaux calmes pour 80% de ses plantes ce qui reste proche de l'expertise.

En revanche, PHAA possède l'attribut A0-50-50 ce qui ne correspond pas à l'expertise.

- L'attribut U100-0-0-0 indique que 4 espèces (CERD, PTCR, PTLU et UTRV) n'aiment pas les eaux instables. PTNA confirme globalement ce trait (U0-100-0-0) en supportant de faibles variations contrairement à PHAA qui peut supporter facilement d'importantes variations (U0-0-0-100).

3.5 Conclusion

Les analyses du treillis étendu ne révèlent pas de traits inadéquats, ils sont tous pertinents et influents sur les associations entre les espèces et les traits biologiques. En comparant les treillis avec l'analyse des experts, nous retrouvons des informations similaires ce qui nous conforte dans notre méthode. En effet, l'ajout de 5 traits par rapport aux 10 traits d'origine (soit 50%) aurait pu bouleverser les concepts. Or comme cela n'est pas le cas, cela signifie que les associations entre espèces et traits biologiques conservent leur cohérence vis à vis des traits écologiques. Par ailleurs le treillis nous permet d'aller au-delà de l'expertise. En effet, comme nous l'illustre le concept 1, le trait E-0-0-100 n'était pas attendu. Mais le treillis l'a souligné comme étant également caractéristique des espèces et des traits biologiques de ce concept. Par ailleurs, en plus de compléter l'expertise, les treillis permettent de la préciser et la nuancer. Comme pour le concept 2 avec l'espèce PHAA dont le comportement dans son environnement écologique se distingue des autres espèces du même concept initial. Cela tient du fait que les experts indiquent des traits écologiques généraux pour les groupes d'espèces venant des concepts. Lorsque les données ont été collectées dans [69], toutes les espèces ont été renseignées pour tous les traits et pas uniquement pour celles qui les ont mis en exergue. Le treillis confirme non seulement les indications de l'expert (les traits biologiques et écologiques sont associés tel qu'indiqué par l'expert) mais nous apportent plus de précision comme dans le cas de l'espèce PHAA. Ils révèlent davantage de relations entre les espèces et les traits écologiques. Une méthode de traitement flou des données est développée dans le chapitre suivant où nous présentons également des algorithmes capable de les mettre en œuvre et des techniques de seuillage pour les optimiser. Cette approche est mise en application sur un contexte de macrophytes et un contexte d'invertébrés dont nous présentons les démarches et résultats dans le chapitre 5.

Chapitre 4

Traitement flou par histogrammes

Sommaire

4.1	Le format histogramme	86
4.1.1	Echelonnage histogramme	86
4.1.2	Fermeture de Galois sur les histogrammes	86
4.1.3	Connexion min-max seuillée	88
4.1.4	Illustration d'un contexte multi-valué flou	88
4.1.5	Structure de patron d'histogramme	91
4.2	Algorithmes de construction de treillis	92
4.2.1	Algorithme NextClosure de Ganter	93
4.2.2	Algorithme Charm	96
4.3	Algorithmes adaptés	97
4.3.1	MinMaxNC	97
4.3.2	MinMaxC	100
4.4	Test de validation des algorithmes	103
4.4.1	Données	104
4.4.2	Concepts	104
4.4.3	Délai d'exécution	105
4.5	Seuillage des histogrammes	106
4.5.1	Seuil brut par trait	107
4.5.2	Seuil fondé sur les écarts réels entre les affinités	108
4.5.3	Seuil par la variance et l'écart-type	108
4.5.4	Seuil par les quantiles	111
4.5.5	Comparatif des deux algorithmes seuillés par quantiles . . .	112
4.6	Conclusion	115

Les données que nous manipulons ont une structure particulière et complexe de répartition entre modalités, ce qui est une information à part entière que nous voulons préserver. Pour cela, il faut étendre le modèle des treillis de Galois, comme cela a été proposé pour d'autres structures complexes dans [17] [56] [62] [53]. Notre travail se situe dans cette lignée, et s'attache au traitement de données multi-valuées floues.

La structure des données est prise en considération par la conversion du triplet (attribut, modalité, affinité) en "histogrammes", conversion appelée *échelonnage histogramme*. Cette conversion est abordée page 59 et nous la définirons formellement dans ce chapitre. Nous présentons ensuite des correspondances de Galois spécifiques pour manipuler les histogrammes et les différentes interprétations que nous permet ce format par rapport au format d'origine. Enfin nous présentons les algorithmes que nous utilisons pour mettre en œuvre ces correspondances de Galois.

4.1 Le format histogramme

Dans le chapitre 2.3 nous avons présenté les définitions des contextes multi-valués, des contextes flous, ainsi que les structures de patrons et les objets symboliques.

Nous avons également expliqué l'intérêt que revêt la répartition des affinités selon les modalités pour les biologistes à la page 57. Pour cette raison, nous avons choisi de convertir chaque triplet (attribut, modalités, affinités) en un *attribut histogramme* représentant la distribution des affinités entre les modalités de l'attribut d'origine. Nous appelons cet échelonnage spécifique un "*échelonnage histogramme*". Nous avons présenté l'idée ainsi que la construction d'un histogramme de manière informelle dans le chapitre précédent, nous les définissons ici de manière formelle.

4.1.1 Echelonnage histogramme

Définition 32 *Attribut histogramme (ou histogramme)*. Considérons un contexte multi-valué flou (O, T, M, A, I) . Soit $t \in T$ un attribut flou possédant n_t modalités m_i dans M , on associe à t un ensemble $H_t = \{h_t\}$ d'histogrammes $h_t = \{a_1, \dots, a_{n_t}\}$ où $a_i \in A$ et $|H_t| \leq |A|^{n_t}$.

Définition 33 *Echelonnage histogramme*. Soit $H = \cup_{t \in T} H_t$ l'ensemble de tous les attributs histogrammes. Alors (O, H, I_H) est un contexte binaire avec $(o, h_t) \in I_H \Leftrightarrow I(o, t, m_i) = a_i$ pour tout $i \in [1, n_t]$.

4.1.2 Fermeture de Galois sur les histogrammes

Le contexte (O, H, I_H) est un contexte binaire sur lequel peut être classiquement définie une correspondance de Galois (définition 17). Toutefois, du fait du grand nombre

d'attributs générés, les concepts obtenus sont peu nombreux et ont une petite extension. Afin de pouvoir regrouper davantage d'objets dans un concept, nous utilisons une méthode proche de celles proposées par [56], méthodes qui associent à un ensemble d'objets symboliques l'union ou l'intersection des valeurs de leurs attributs. Considérons alors la relation d'ordre suivante sur les attributs histogrammes :

$$(h_t^1 \leq h_t^2) \Leftrightarrow (a_1^1 \leq a_1^2) \wedge (a_2^1 \leq a_2^2) \wedge \dots \wedge (a_{n_t}^1 \leq a_{n_t}^2)$$

où h_t^1 et h_t^2 sont deux histogrammes dérivant d'un même attribut t . Le minimum et le maximum entre deux histogrammes h_t^1, h_t^2 sont définis de la façon suivante :

$$\begin{aligned} \min(h_t^1, h_t^2) &= \{\min(a_1^1, a_1^2), \min(a_2^1, a_2^2), \dots, \min(a_{n_t}^1, a_{n_t}^2)\} \\ \max(h_t^1, h_t^2) &= \{\max(a_1^1, a_1^2), \max(a_2^1, a_2^2), \dots, \max(a_{n_t}^1, a_{n_t}^2)\} \end{aligned}$$

Dans la suite on considère qu'un objet possède un histogramme pour chaque attribut d'origine, l'ensemble des histogrammes d'un objet formant une séquence ordonnée. On note $\theta \in \Theta = \times_{t \in T} H_t$ une séquence d'attributs histogrammes ainsi ordonnée. $\theta(o)$ est la séquence particulière d'attributs histogrammes de l'objet $o \in O$.

La comparaison entre deux séquences θ^1 et θ^2 s'effectue simplement par comparaison deux à deux des histogrammes h_t^1, h_t^2 correspondant à chaque attribut t . Sur cette base, nous pouvons définir un couple $(\uparrow, \downarrow)$ de fonctions sur les ensembles Θ et O :

Définition 34

$$\begin{aligned} \uparrow: X \subseteq O &\rightarrow X^\uparrow = \{\theta \in \Theta \mid \min_{o \in X} \theta(o) \leq \theta \leq \max_{o \in X} \theta(o)\} \\ \downarrow: Y \subseteq \Theta &\rightarrow Y^\downarrow = \{o \in O \mid \min_{\theta \in Y} \theta \leq \theta(o) \leq \max_{\theta \in Y} \theta\} \end{aligned}$$

Le couple $(\uparrow, \downarrow)$ constitue une **pseudo correspondance de Galois**, que nous appellerons **correspondance min-max**.

En effet, $\uparrow$ et $\downarrow$ sont deux applications monotones croissantes : soient $\forall X, Y \subseteq O, X \subseteq Y, X^\uparrow = \{\theta \in \Theta \mid \min_{o \in X} \theta(o) \leq \theta \leq \max_{o \in X} \theta(o)\} \subseteq Y^\uparrow = \{\theta \in \Theta \mid \min_{o \in Y} \theta(o) \leq \theta \leq \max_{o \in Y} \theta(o)\}$, car

$$\begin{cases} \min_{o \in X} \theta(o) \geq \min_{o \in Y} \theta(o) \\ \max_{o \in X} \theta(o) \leq \max_{o \in Y} \theta(o) \end{cases}$$

Ces deux applications sont également extensives :

$$X^{\uparrow\downarrow} = [\min_{o \in X} \theta(o), \max_{o \in X} \theta(o)]^\downarrow \supseteq X$$

et idempotentes :

$$\begin{aligned} (X^{\uparrow\downarrow})^{\uparrow\downarrow} &= \{o \in O \mid \min_{\theta \in Y} \theta \leq \theta(o) \leq \max_{\theta \in Y} \theta\}^{\uparrow\downarrow} \text{ avec } Y = X^\uparrow \\ &= [\min_{o \in X} \theta(o), \max_{o \in X} \theta(o)]^\downarrow = \{o \in O \mid \min_{\theta \in Y} \theta \leq \theta(o) \leq \max_{\theta \in Y} \theta\} \end{aligned}$$

Les applications composées $\uparrow\downarrow$ et $\downarrow\uparrow$ sont donc monotones croissantes, extensives et idempotentes, et sont donc des fermetures dans O et H respectivement.

4.1.3 Connexion min-max seuillée

Par ailleurs, pour limiter le nombre de concepts et pour obtenir des concepts avec des intensions plus spécifiques et pertinentes, nous introduisons une contrainte de seuil dans la connexion min-max c'est-à-dire avec un écart entre les attributs min et max inférieur à une certaine limite :

Définition 35 La connexion min-max seuillée est une connexion min-max avec le couple $(\uparrow^l, \downarrow^l)$ tel que :

$$\begin{aligned} \uparrow^l: X \subseteq O &\rightarrow X^{\uparrow^l} \\ X^{\uparrow^l} &= \begin{cases} \{\theta \in \Theta \mid \min_{o \in X} \theta(o) \leq \theta \leq \max_{o \in X} \theta(o)\}, \\ \text{si } distance(\max_{o \in X} \theta(o), \min_{o \in X} \theta(o)) \leq \text{seuil} \\ \emptyset \text{ sinon} \end{cases} \\ \downarrow^l: Y \subseteq \Theta &\rightarrow Y^{\downarrow^l} \\ Y^{\downarrow^l} &= \begin{cases} \{o \in O \mid \min_{\theta \in Y} \theta \leq \theta(o) \leq \max_{\theta \in Y} \theta\}, \\ \text{si } distance(\max_{\theta \in Y} \theta, \min_{\theta \in Y} \theta) \leq \text{seuil} \\ \emptyset \text{ sinon} \end{cases} \end{aligned}$$

Il s'agit d'une fermeture "floue" au sens où on ne regroupe pas des objets qui ont des attributs identiques, mais des objets qui ont des attributs *suffisamment* proches, similaires. Plus précisément, cette fermeture nous permet de déterminer des groupes d'objets compris entre des valeurs minimale et maximale d'attributs, comme nous l'illustrons.

4.1.4 Illustration d'un contexte multi-valué flou

Nous illustrons la notion de contexte multi-valué flou avec un exemple simple extrait de nos données. Nous nous intéressons ici à deux traits écologiques : la tolérance à la sédimentation (que nous référençons par la lettre E) et la tolérance à la trophie⁽⁶⁴⁾ (référéncée par la lettre I). Nous avons limité le nombre d'objets à 3 : les espèces ALIP (*Alisma plantago-aquatica*), BERE (*Berula erecta*) et CALO (*Callitriche obtusangula*). La table 4.1 présente cet extrait des données.

Conversion en histogrammes.

Les attributs histogrammes correspondant aux triplets (attribut, modalités, affinités) de la table 4.1 sont constitués avec une lettre, 'E' pour la tolérance à la sédimentation et 'I' pour la tolérance à la trophie, à laquelle on accole des affinités en nombre et dans l'ordre correspondant aux modalités de l'attribut "tolérance à la sédimentation" (3 modalités) ou "tolérance à la trophie" (4 modalités). La relation binaire indique si une plante possède le profil décrit par l'histogramme. La table 4.2 présente le nouveau

Traits	Tolérance à la sédimentation			Tolérance à la trophie			
	aucune (plante étouffée)	moyenne (enracinement variable)	fort (enracinement constant)	oligotrophe	mesotrophe	eutrophe	hypertrophe
Modalités							
ALIP	0	0	2	0	1	1	0
BERE	0	2	0	1	2	1	0
CALO	0	2	0	0	2	1	0

TABLE 4.1 – Extrait des données macrophytes pour deux traits écologiques.

contexte binaire constitué avec les attributs histogrammes. Seuls les histogrammes utiles (i.e. associés à des individus) sont conservés. La figure 4.1 présente le treillis obtenu par application d’une correspondance de Galois classique à ce jeu de données binaire.

Histogramme	E002	E020	I0110	I1210	I0210
ALIP	1	0	1	0	0
BERE	0	1	0	1	0
CALO	0	1	0	0	1

TABLE 4.2 – Données histogrammes de l’extrait des données sur les traits écologiques.

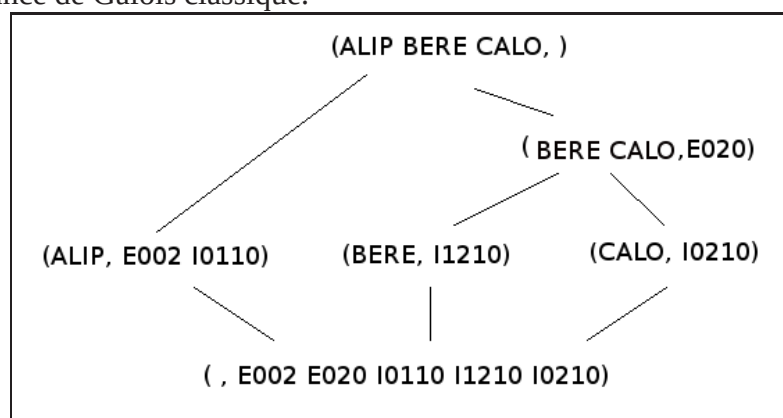
Correspondance min-max.

La correspondance min-max ($\uparrow, \downarrow$) permet de construire un treillis de Galois de la façon suivante. Initialement nous avons les couples $(o, \theta(o))$:

- A : $(\{\text{ALIP}\}, \{\text{E002 I0110}\})$
- B : $(\{\text{BERE}\}, \{\text{E020 I1210}\})$
- C : $(\{\text{CALO}\}, \{\text{E020 I0210}\})$

Chacun des couples A, B, C est fermé maximal pour la correspondance min-max. Par exemple pour ALIP : $(\{\text{ALIP}\})^{\uparrow\downarrow} = (\{\text{E002 I0110}\})^{\downarrow} = \{\text{ALIP}\}$. Comparons maintenant ces couples deux à deux. Nous indiquons pour chaque comparaison, quels sont les histogrammes minimal et maximal issus de f_H et la fermeture qui en découle.

FIGURE 4.1 – Treillis de Galois de l'extrait de données sur les traits écologiques en correspondance de Galois classique.



- A avec B :

$$(\{ALIP BERE\})^{\uparrow, \downarrow} = (\{\theta | [E000 I0110] \leq \theta \leq [E022 I1210]\})^{\downarrow} = \{ALIP BERE CALO\}$$

On obtient le concept D : **({ALIP BERE CALO}, {[E002 I0110][E020 I1210][E020 I0210]}).**

- A avec C :

$$(\{ALIP CALO\})^{\uparrow, \downarrow} = (\{\theta | [E000 I0110] \leq \theta \leq [E022 I0210]\})^{\downarrow} = \{ALIP CALO\}$$

On obtient le concept E : **({ALIP CALO}, {[E002 I0110][E020 I0210]}).**

- B avec C :

$$(\{BERE CALO\})^{\uparrow, \downarrow} = (\{\theta | [E020 I0210] \leq \theta \leq [E020 I1210]\})^{\downarrow} = \{BERE CALO\}$$

On obtient le concept F : **({BERE CALO}, {[E020 I1210][E020 I0210]}).**

La figure 4.2 présente le treillis des concepts ainsi déterminés. L'intension des concepts est symbolisée par leurs minima et maxima communs. On constate que l'étendue des intensions croît avec le nombre d'objets de l'extension.

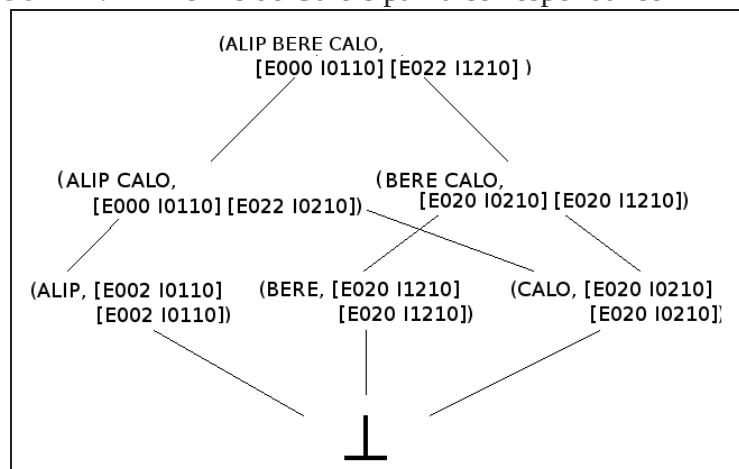
Comparaison des treillis.

Intéressons nous à la comparaison des deux treillis. La figure 4.1 montre que BERE et CALO ont toutes les deux une forte propension (affinité valant 2) à un enracinement variable et exclusivement cette sorte de tolérance à la sédimentation : concept (BERE CALO, E020).

Le concept F : (BERE CALO, [E020 I0210] [E020 I1210]) nous renseigne non seulement sur la tolérance à la sédimentation de BERE et CALO mais nous précise qu'ils ont également une préférence trophique oligotrophe à eutrophe, mais ne tolèrent pas l'hypertrophie (Ixxx0).

L'hypertrophie est également rejetée (Ixxx0) par le concept E (ALIP CALO, [E000 I0110] [E022 I0210]) de même que la possibilité de vivre dans un milieu oligotrophe

FIGURE 4.2 – Treillis de Galois par la correspondance min-max.



(I0xxx). Cependant il autorise une tolérance à la sédimentation plus large, car il accepte toute forme de tolérance à la sédimentation (E0xx : pas de plante étouffée).

On voit donc bien les liens qui se créent entre les propriétés. Dans notre exemple, si une plante n'est pas adaptée aux milieux hypertrophes, elle a toute sa place avec l'un ou l'autre de ces concepts. Cependant elle sera placée dans les concepts en fonction de ses autres traits, en effet si elle a une forte tolérance à la sédimentation par un enracinement constant, selon ce treillis, elle ne peut pas vivre dans un environnement oligotrophe.

Enfin la correspondance min-max sur les histogrammes permet de se rendre compte d'un point commun entre ALIP, BERE et CALO grâce au concept D (ALIP BERE CALO, [E000 I0110][E022 I1210]) : aucun n'est intolérant à la sédimentation (E0xx) et ils peuvent tous vivre dans des environnements non hypertrophes (Ixxx0), ce qui n'apparaît pas avec la correspondance classique puisque le concept top (ALIP BERE CALO,) a une intensité vide.

4.1.5 Structure de patron d'histogramme

La définition des structures de patrons appelées en anglais "pattern structures" comme présenté à la page 41 proposée par [35] permet une généralisation d'écriture et un formalisme commun aux contextes non binaires. C'est donc pour cela que nous proposons une écriture du formalisme des histogrammes sous la forme de structures de patrons.

Une structure de patrons est un ensemble d'objets ayant une "description" (patron) avec une opération de similarité $\sqcap$ qui correspond à une contrainte permettant d'associer deux éléments qui ne sont pas égaux mais sont semblables, par exemple "bleu" et "noir" sont "foncés". La relation de subsomption entre les patrons est telle que $x \sqsubseteq y \Leftrightarrow x \sqcap y = x$ car l'opération de similarité respecte les trois lois :

- idempotence : $x \sqcap x = x$

- commutativité : $x \sqcap y = y \sqcap x$
- associativité : $(x \sqcap y) \sqcap z = x \sqcap (y \sqcap z)$

Définition 36 Structure de patron d’histogramme. Soient Θ l’ensemble de toutes les séquences d’histogrammes possibles (de longueur fixe), O un ensemble d’objets, $(\Theta \times \Theta, \sqcap)$ un inf-demi-treillis et $\delta : O \rightarrow \Theta \times \Theta$ une application. On définit une structure de patrons par le triplet $(O, \underline{\Theta}^2, \delta)$ avec $\underline{\Theta}^2 = (\Theta \times \Theta, \sqcap)$ à condition que

$$\delta(O) = \{\delta(o) \mid o \in O\}$$

Dans notre cas, nous nous intéressons à une opération fournissant des descriptions minimale et maximale communes pour chaque ensemble d’objets. Pour cela nous manipulons des couples (ou intervalles) de séquences d’histogrammes définis par (θ^i, θ^s) et tels que $\theta^i \leq \theta^s$ (au sens défini précédemment, page 87). En particulier, la description d’un objet est $\delta(o) = (\theta^i, \theta^s)$ avec $\theta^i(o) = \theta^s(o)$. La relation de similarité entre deux couples de séquences d’histogrammes (θ_1^i, θ_1^s) et (θ_2^i, θ_2^s) est alors :

$$(\theta_1^i, \theta_1^s) \sqcap (\theta_2^i, \theta_2^s) = (\min(\theta_1^i, \theta_2^i), \max(\theta_1^s, \theta_2^s))$$

Les connexions de Galois min-max pour les histogrammes appliquées aux structures de patrons $(O, \underline{\Theta}^2, \delta)$ sont

$$A^\square := \bigcap_{o \in A} \delta(o) \quad \text{pour } A \subseteq O \Leftrightarrow$$

$$A^\square := (\min_{o \in A} \theta^i(o), \max_{o \in A} \theta^s(o)) \quad \text{pour } A \subseteq O$$

$$d^\square := \{o \in O \mid \theta_d^i \leq \theta^i(o) \leq \theta^s(o) \leq \theta_d^s\} \text{ avec } d = (\theta_d^i, \theta_d^s) \Leftrightarrow$$

$$d^\square := \{o \in O \mid (\theta_d^i, \theta_d^s) \sqsubseteq \delta(o)\}$$

Un concept patron se définit par (A, d) en satisfaisant $A \subseteq O$, $d \in \Theta^2$, $A^\square = d$, $A = d^\square$, avec A l’extension et d l’intension (patron).

4.2 Algorithmes de construction de treillis

Notre objectif est de construire un treillis et plus particulièrement d’extraire les fermés maximaux pour un contexte multi-valué flou. Pour cela nous avons besoin des connexions min-max. Nous nous sommes intéressés à des algorithmes performants que nous pourrions adapter. Nous présentons ici les algorithmes NextClosure [34] et Charm [72] que nous avons retenus.

4.2.1 Algorithme NextClosure de Ganter

Les études comparatives [33, 41] concluent que l'algorithme NextClosure de Ganter et celui de Bordat sont les plus rapides, cependant ces méthodes se distinguent sur leur utilisation. En effet, celle de Bordat sera choisie plutôt pour construire le diagramme de Hasse du treillis, et celle de Ganter pour les contextes volumineux dont la taille de la relation est importante, autrement dit un contexte dont le nombre de croix est conséquent. "L'algorithme de Ganter est le seul utilisable pour des correspondances portant sur des ensembles de plus de, disons, 15 éléments" [41]. Comme nous n'avons pas besoin du diagramme de Hasse, et que notre contexte est volumineux, nous choisissons l'algorithme NextClosure. De plus cet algorithme s'applique à une relation de fermeture quelconque, ce qui nous permet d'y adapter la correspondance min-max. Cet algorithme s'appuie sur un ordre lexicographique pour parcourir le contexte. Nous présentons cet ordre avant l'algorithme lui-même.

Ordre lexicographique

Dans l'algorithme NextClosure, Ganter propose de s'appuyer sur un ordre lexicographique [34] pour optimiser les calculs. Guénoche [41] reprend cette définition ainsi :

Considérons le contexte (O, T, I) , n le nombre d'objets dans O et $A, B \subseteq T$ deux ensembles distincts. $A = (a_1, a_2, \dots, a_n)$ et $B = (b_1, b_2, \dots, b_n)$ peuvent s'écrire comme deux vecteurs référençant tous les objets, tels que chaque valeur a_i (resp. b_i) vaut 0 (ou 1) pour indiquer que l'objet est considéré (ou non) dans le sous-ensemble A (resp. B). Par exemple, $(0,0,0,0) < (0,0,0,1) < (0,0,1,0) < (0,0,1,1) < (0,1,0,0)$ etc... Soit $A < B$ et i l'indice tel que pour tout $j < i$ on ait $a_j = b_j$ et $a_i < b_i$ (on a donc $a_i = 0$ et $b_i = 1$). On note $A <_i B$. Maintenant si dans $(a_1, a_2, \dots, a_n)$ on a $a_j = 0$, on note $A_j^+ = (a_1, \dots, a_{j-1}, 1, 0, \dots, 0)$ et on définit $A \oplus j = (A_j^{+\uparrow\downarrow})$. B. Ganter en déduit trois lemmes et une proposition que nous ne redémontrons pas :

- Si $i \notin A$ alors $A < A \oplus i$
- Si $A <_i B$ et B fermé alors $A \oplus i \subseteq B$ et en particulier $A \oplus i \leq B$
- Si $A <_i B$ et B fermé alors $A <_i A \oplus i$

Le plus petit fermé placé après A pour l'ordre lexicographique est $A \oplus i$ où i est le plus grand indice tel que $A <_i A \oplus i$. De plus, Ganter énonce la proposition suivante : le plus petit fermé placé après A pour l'ordre lexicographique, noté A^+ est $A \oplus i$, avec i le plus grand indice tel que $A <_i A \oplus i$ et

$$A \oplus i = ((A \cap \{1, \dots, i-1\}) \cup \{i\})^{\uparrow\downarrow}$$

Considérons l'exemple suivant pour illustrer cette notion : Soit $E = \{1, 2, 3, 4, 5, 6\}$, et les sous-ensembles de E $\{1\}, \{2\}, \{2, 3\}, \{3, 4, 5\}, \{3, 6\}, \{1, 4, 5\}$.

Nous avons alors :

- $\{2\} <_1 \{1\}$ car $1 \in (\{1\} \setminus \{2\}) = \{1\}$ et $A \cap \emptyset = B \cap \emptyset$
 Ecrit de manière vectorielle nous avons : $010000 <_1 100000$.
- $\{3, 6\} <_4 \{3, 4, 5\}$ car $4 \in (\{3, 4, 5\} \setminus \{3, 6\}) = \{4, 5\}$ et $A \cap \{1, 2, 3\} = B \cap \{1, 2, 3\}$
 vectoriellement nous avons : $001001 <_4 001110$
- Tous les ensembles sont rangés lexicographiquement ainsi : $\{3, 6\} <_4 \{3, 4, 5\} <_2$
 $\{2\} <_3 \{2, 3\} <_1 \{1\} <_4 \{1, 4, 5\}$
 vectoriellement nous avons : $001001 <_4 001110 <_2 010000 <_3 011000 <_1 100000 <_4$
 100110 . On peut noter que cela correspond à l'ordre entre les nombres binaires.

Algorithme NextClosure

Le nombre de parties d'un ensemble est exponentiel : 2^n avec n le nombre d'éléments de l'ensemble. Par ailleurs plusieurs sous-ensembles peuvent avoir des fermetures identiques, et la vérification de la redondance des fermetures est d'une complexité exponentielle. C'est pourquoi NextClosure est intéressant car il génère toutes les extensions d'un contexte donné (O, T, I) et il propose de générer des fermetures dans un ordre prédéfini pour éviter d'avoir à faire la vérification de doublons.

Pour un sous-ensemble $A \subset T$, on trouve l'intension "suivante" (selon l'ordre lexicographique) en balayant tous les éléments i de $T \setminus A$, en commençant par le plus grand et en décroissant jusqu'à ce que $A <_i A \oplus i$. $A \oplus i$ est alors l'intension "suivante" recherchée²⁷. NextClosure est présenté dans le pseudo code Algorithm 1. Un algorithme plus détaillé est fourni en annexe par l'algorithme 8.

Algorithm 1 NextClosure

```

1:  $A \leftarrow \emptyset^{\uparrow\downarrow}$ 
2: Mémoriser (A);
3: while  $A \neq T$  do
4:    $A \leftarrow A^+$ ;
5:   Mémoriser (A);
6: end while

```

Exemple de déroulement de l'algorithme NextClosure

Pour comprendre le fonctionnement de l'algorithme de Ganter, nous le déroulons sur le contexte suivant (tableau 4.3) [16] :

1. $A = \emptyset^{\uparrow\downarrow} = \emptyset$

27. Par des propriétés de dualité, entre les objets et les attributs, inhérentes aux treillis de Galois, le même principe peut être adopté pour les objets en remplaçant T par O .

I	1	2	3
o1	x	x	x
o2	x		x
o3		x	x
o4	x		

TABLE 4.3 – Exemple de contexte pour illustrer le fonctionnement de l’algorithme de Ganter

2. Rechercher $A^+ = \emptyset^+ = A \oplus i$, i devant être la plus grande des valeurs possibles telle que $A <_i A \oplus i$. Nous testons chacune des valeurs $i = 3, 2, 1$ et testons si $A <_i A \oplus i$:

$i = 3 : A \oplus i = ((\emptyset \cap \{1, 2\}) \cup \{3\})^{\uparrow\downarrow} = \{3\}^{\uparrow\downarrow} = \{3\}$ et $\emptyset <_3 \{3\} = A \oplus i$, ainsi $A^+ = \{3\}$.

3. Rechercher $\{3\}^+ :$

$i = 3 : A \oplus i = ((\{3\} \cap \{1, 2\}) \cup \{3\})^{\uparrow\downarrow} = \{3\}^{\uparrow\downarrow} = \{3\}$ et $\{3\} \not<_3 \{3\} = A \oplus i$, ainsi nous testons la valeur $i = 2$.

$i = 2 : A \oplus i = ((\{3\} \cap \{1\}) \cup \{2\})^{\uparrow\downarrow} = \{2\}^{\uparrow\downarrow} = \{2, 3\}$ et $\{3\} <_2 \{2, 3\} = A \oplus i$, ainsi $A^+ = \{2, 3\}$

4. Rechercher $\{2, 3\}^+ :$

$i = 3 : A \oplus i = \{2, 3\}^{\uparrow\downarrow} = \{2, 3\}$ et $\{2, 3\} \not<_3 \{2, 3\} = A \oplus i$, ainsi nous testons la valeur $i = 2$.

$i = 2 : A \oplus i = \{2\}^{\uparrow\downarrow} = \{2, 3\}$ et $\{2, 3\} \not<_2 \{2, 3\} = A \oplus i$, ainsi nous testons la valeur $i = 1$.

$i = 1 : A \oplus i = \{1\}^{\uparrow\downarrow} = \{1\}$ et $\{2, 3\} <_1 \{1\} = A \oplus i$, ainsi nous avons $A^+ = \{1\}$.

5. Rechercher $\{1\}^+ :$

$i = 3 : A \oplus i = \{1, 3\}^{\uparrow\downarrow} = \{1, 3\}$ et $\{1\} <_3 \{1, 3\} = A \oplus i$, ainsi nous avons $A^+ = \{1, 3\}$.

6. Rechercher $\{1, 3\}^+ :$

$i = 3 : A \oplus i = \{1, 3\}^{\uparrow\downarrow} = \{1, 3\}$ et $\{1, 3\} \not<_3 \{1, 3\} = A \oplus i$, ainsi nous testons la valeur $i = 2$. – $i = 2 : A \oplus i = \{1, 2\}^{\uparrow\downarrow} = \{1, 2, 3\}$ et $\{1, 3\} <_2 \{1, 2, 3\} = A \oplus i$, ainsi nous obtenons $A^+ = \{1, 2, 3\} = E$.

Les intensions sont donc (par ordre lexicographique) : $\emptyset < \{3\} < \{2, 3\} < \{1\} < \{1, 3\} < \{1, 2, 3\}$.

4.2.2 Algorithme Charm

Présentation générale de l'algorithme

Charm est un algorithme efficace pour l'extraction d'ensembles de fermés fréquents grâce à un parcours en profondeur des données en entrée sous forme d'arbre. Charm innove en proposant une exploration à la fois de l'espace des *itemsets* $\mathfrak{I}$ (qui correspond à notre ensemble d'objets) et de celui des *tidsets* $\mathfrak{T}$ (qui correspond à notre ensemble de propriétés). Cette double exploration lui permet d'identifier plus rapidement les itemsets fermés fréquents. Par ailleurs, Charm utilise une stratégie d'élagage en deux volets. Le premier élimine les sous-ensembles non-fréquents (dont le nombre d'objets de l'extension soit supérieur à un seuil, page 45) et le second élimine tous les sous-ensembles non fermés. Charm ne se base que sur les deux opérations fondamentales d'union des itemsets et d'intersection de leurs tidsets. Charm construit un graphe en combinant chaque nœud avec chacun de ses "frères cadets" (issu du même père selon l'ordre choisi). Dans le cas des treillis, le père initial est bottom et ses fils sont chacun des objets (ou top et ses fils sont chacune des propriétés). Les fils sont rangés selon un ordre qui peut être par exemple lexicographique ou lié à la valeur des supports. Ainsi 3 nœuds A, B, C, seraient combinés ainsi : AB, AC et BC.

Ainsi Charm détermine tous les concepts fréquents de la forme (X, Y) avec $X = \text{itemset}(Y)$ et $Y = \text{tidset}(X)$ (écrit plus simplement $t(X)$). La recherche des concepts peut tout aussi bien se mener à partir de l'espace des objets que celui des attributs.

Propriétés des couples itemsets-tidsets

Considérons f , une application définissant un ordre total sur tous les itemsets. Pour deux itemsets X_1 et X_2 , $X_1 \leq X_2$ si et seulement si $f(X_1) \leq f(X_2)$. Supposons que nous voulons combiner $X_1 \times t(X_1)$ avec son frère cadet $X_2 \times t(X_2)$, alors Charm s'appuie sur les 4 propriétés suivantes :

1. Si $t(X_1) = t(X_2)$, alors $t(X_1 \cup X_2) = t(X_1) \cap t(X_2) = t(X_1) = t(X_2)$. Ainsi on peut simplement remplacer chaque occurrence de X_1 par $X_1 \cup X_2$, et enlever X_2 puisque sa fermeture est égale à la fermeture de $X_1 \cup X_2$.
2. Si $t(X_1) \subset t(X_2)$, alors $t(X_1 \cup X_2) = t(X_1) \cap t(X_2) = t(X_1) \neq t(X_2)$. Dans ce cas, on remplace chaque occurrence de X_1 par $X_1 \cup X_2$, puisque si on a X_1 alors on a aussi X_2 . Cependant puisque $t(X_1) \neq t(X_2)$ on ne peut pas enlever X_2 car il génère une fermeture différente.
3. Si $t(X_1) \supset t(X_2)$, alors $t(X_1 \cup X_2) = t(X_1) \cap t(X_2) = t(X_2) \neq t(X_1)$. Ici, on remplace chaque occurrence de X_2 par $X_1 \cup X_2$, puisque quand on a X_2 , on a toujours X_1 . Cependant X_1 produit une fermeture et doit donc être conservé.
4. Si $t(X_1) \neq t(X_2)$, alors $t(X_1 \cup X_2) = t(X_1) \cap t(X_2) \neq t(X_2) \neq t(X_1)$. Dans ce cas, rien ne peut être éliminé car X_1 et X_2 produisent tous les deux des fermetures

différentes.

algorithm Charm

Cet algorithme récursif (Algorithm 2) est construit sur trois fonctions que nous présentons. La première, notée *Charm*, est la fonction principale appelant la fonction récursive *Charm-Extend* pour chaque sous-ensemble $[P]$ de concepts créé. La dernière fonction *Charm-Property* sert à élaguer les fermés qui ont été calculés.

L'algorithme démarre par l'initialisation des nœuds à examiner tels qu'ils correspondent à chaque objet et leurs propriétés si le support (nombre d'objets) est supérieur au seuil *minsup*.

Charm-Extend a pour objectif de rechercher si chaque branche du graphe est viable. Il extrait chaque couple itemset-tidset ($X_i \times t(X_i)$) de l'ensemble des nœuds courants : P (ligne 4) et le combine aux couples suivants ($X_j \times t(X_j)$), (ligne 6) selon l'ordre total du contexte. La combinaison de deux couples itemset-tidset est traitée ligne 6. La fonction *Charm-Property* teste l'ensemble résultant en fonction du support requis et applique les 4 propriétés vues plus tôt. On remarque que cette fonction peut modifier le nœud courant en supprimant des couples itemset-tidset qui seraient contenus dans d'autres nœuds. Il inclut également le nouveau couple dans l'ensemble P_i des nouveaux nœuds. Si cet ensemble ne reste pas vide, le processus est réitéré récursivement selon une recherche en profondeur (ligne 10). Ensuite on insère l'itemset X_i (possiblement étendu), dans l'ensemble des itemsets fermés, puisqu'il arrive au terme des traitements qui le concernent, et puisqu'à cette étape n'importe quel itemset fermé contenant X_i a déjà été généré. Enfin nous rebouclons à la ligne 4 pour traiter la branche non-élaguée suivante.

La fonction *Charm-Property* teste simplement si le nouveau couple est fréquent et le supprime s'il ne l'est pas. Ensuite elle teste les 4 propriétés des couples itemset-tidset : étendre les itemsets existants, supprimer des branches subsumées de l'ensemble de nœuds courant, ou insérer de nouveaux couples dans l'ensemble de nœuds, pour la prochaine étape de la recherche en profondeur.

4.3 Algorithmes adaptés

Les algorithmes NextClosure et Charm utilisent une connexion de Galois classique, c'est-à-dire binaire. Nous présentons dans cette partie les modifications que nous avons apportées à ces algorithmes pour intégrer les correspondances min-max : MinMaxNC et MinMaxC.

4.3.1 MinMaxNC

Algorithm 2 CharmCharm

Require: minsup est le support minimal ou fréquence, $\mathfrak{S}$ est un ensemble d'éléments (objets en FCA). $\sigma(A)$ est le support de A tel que $\sigma(A) = |t(A)|$.

Ensure: C est l'ensemble de tous les fermés.

- 1: $[P] = X_i \times t(X_i) : X_i \in \mathfrak{S} \wedge |t(X_i)| \geq \text{minsup}$
- 2: Charm-Extend($[P]$, $C = \emptyset$)
- 3: **return** C

Charm-extend

Require: $[P]$ et C

- 4: **for** chaque $X_i \times t(X_i)$ de $[P]$ **do**
- 5: $[P_i] = \emptyset$ et $X = X_i$
- 6: **for** chaque $X_j \times t(X_j)$ de $[P]$, avec $X_j \geq_f X_i$ **do**
- 7: $X = X \cup X_j$ et $Y = t(X_i) \cap t(X_j)$
- 8: Charm-Property($[P]$, $[P_i]$)
- 9: **end for**
- 10: **if** $P_i \neq \emptyset$ **then**
- 11: Charm-Extend ($[P_i]$, C)
- 12: **end if**
- 13: supprimer $[P_i]$
- 14: $C = C \cup X$ { //si X n'est pas subsumé }
- 15: **end for**

Charm-Property

Require: $[P]$ et $[P_i]$

- 16: **if** $\sigma(X) \geq \text{minsup}$ **then**
- 17: **if** $t(X_i) = t(X_j)$ **then** { //Propriété 1 }
- 18: Retirer X_j de $[P]$
- 19: Remplacer tous les X_i par X
- 20: **else**
- 21: **if** $t(X_i) \subset t(X_j)$ **then** { //Propriété 2 }
- 22: Remplacer tous les X_i par X
- 23: **else**
- 24: **if** $t(X_i) \supset t(X_j)$ **then** { //Propriété 3 }
- 25: Retirer X_j de $[P]$
- 26: Ajouter $X \times Y$ à $[P_i]$ { //en utilisant l'ordre f }
- 27: **else**
- 28: **if** $t(X_i) \neq t(X_j)$ **then** { //Propriété 4 }
- 29: Ajouter $X \times Y$ à $[P_i]$ { //en utilisant l'ordre f }
- 30: **end if**
- 31: **end if**
- 32: **end if**
- 33: **end if**
- 34: **end if**

Algorithm 3 Fonction min-max

Require: La liste des objets concernés**Ensure:** Etablit l'intension du concept : le min (desc-min) et le max (desc-max) communs de chaque propriété pour tous les objets

```

1: for chaque objet  $o$  du concept do
2:   for chaque propriété  $p$  do
3:     if  $p < \text{desc-min}[p]$  then
4:        $\text{desc-min}[p] = p$ 
5:     end if
6:     if  $p > \text{desc-max}[p]$  then
7:        $\text{desc-max}[p] = p$ 
8:     end if
9:   end for
10: end for
11: intension du concept  $\leftarrow$  desc-min et desc-max

```

Algorithm 4 Fonction fermeture

Require: le concept à fermer**Ensure:** le concept fermé maximal

```

1: for chaque objet  $o \notin$  concept do
2:   booléen appartient = false
3:   for chaque propriété  $p$  do
4:     if  $(p < \text{desc-min}[p])$  or  $(p > \text{desc-max}[p])$  then
5:       appartient = true
6:     end if
7:   end for
8:   if ! appartient then
9:     inclure  $o$  dans concept
10:  end if
11: end for
12: return concept fermé

```

Nous présentons les modifications inhérentes à l'intégration des connexions de Galois min-max dans l'algorithme NextClosure (algorithme 8). Ces modifications sont notifiées en gras dans l'algorithme 5. En commentaire sont indiquées les lignes de l'algorithme 8 correspondant à la ligne en cours dans l'algorithme 5.

Notons au préalable que les vecteurs *objets-à-fermer*, *objets-traités* et *concept-à-fermer* sont tous de même longueur, à savoir le nombre d'objets du contexte : *nombre-d'objets*. Il est donc important que les objets soient toujours rangés de la même façon pour les trois vecteurs. L'ordre lexicographique peut alors être choisi.

La première modification est le calcul des descriptions (ensemble de propriétés) minimale (*desc-min*) et maximale (*desc-max*) de l'ensemble d'objets en cours de traitement : *objets-à-traiter* (ligne 12). Cette opération s'effectue par l'appel de la fonction *min-max* décrite par l'algorithme 3.

La seconde modification est la fermeture de l'ensemble d'objets en cours de traitement : *objets-à-fermer* (ligne 13) qui appelle la fonction *fermeture* décrite par l'algorithme 4. Cette fonction considère chaque objet n'étant pas déjà présent dans l'ensemble d'objets considérés. Si les propriétés de cet objet sont supérieures au minimum commun (*desc-min*) du concept en construction et inférieures au maximum commun (*desc-max*), alors l'objet est inclus dans le concept. Cette fonction fournit donc un fermé maximal : *concept-fermé*.

La ligne 12 de l'algorithme prend une liste d'objets : *objets-à-fermer* et détermine les minimum et maximum communs qui vont former l'intension du concept que nous recherchons. C'est pourquoi une fois cette opération terminée, un concept est formé : *concept-à-fermer* car il contient une intension et une extension. Cependant l'extension n'est pas maximale. Ainsi la ligne 13 de l'algorithme cherche l'extension maximale du concept pour obtenir le *concept-fermé-maximal*. Enfin la dernière modification tient aux particularités de la fermeture min-max. En effet, dans la version binaire de l'algorithme, l'ordre lexicographique implique que les fermetures des combinaisons sont uniques. Or dans le cas de la correspondance min-max, deux objets A et B peuvent avoir les mêmes minimum et maximum que deux autres objets C et D. Alors C et D seront inclus dans le concept obtenu par la fermeture de A et B. Et réciproquement A et B seront inclus dans le concept issu de la fermeture de C et D. Ceci implique deux concepts doublons. Cette contrainte est donc vérifiée à la ligne 14 de l'algorithme.

4.3.2 MinMaxC

Cet algorithme (algorithme 6) est une extension de l'algorithme 2 présenté plus tôt. Les différences portent sur l'intégration de nos correspondances de Galois. Mais également sur la révision des conditions d'élagage. L'objectif est de remplir un tableau nommé *Closed* avec tous les concepts fermés maximaux. Dans une étape préalable (le *main*), l'algorithme construit un tableau *U* des concepts unaires, autrement dit des concepts ne contenant qu'un seul objet et leurs propriétés associées.

Algorithm 5 NextClosure adapté : MinMaxNC

Require: le concept à fermer**Ensure:** le concept fermé maximal

```

1: booléen test-fin = false
2: while ! test-fin do {ligne 2}
3:   for chaque objet  $o$  do {ligne 4}
4:     if ! objets-traités[ $o$ ] then
5:       for  $i \leftarrow 0$  to  $o$  do {ligne 5}
6:         objets-à-fermer[ $i$ ]=false
7:       end for
8:       objets-à-fermer[ $o$ ]=false{ligne 9}
9:       for  $i \leftarrow o + 1$  to nombre-d'objets do {ligne 10}
10:        objets-à-fermer[ $i$ ]=objets-traités[ $i$ ]
11:      end for
12:      concept-à-fermer = min-max(objets-à-fermer)
13:      concept-fermé-maximal = fermeture(concept-à-fermer) {ligne 13}
14:      if ! doublon(concept-fermé-maximal) then
15:        mémoriser concept-fermé-maximal
16:      end if
17:    end if
18:  end for
19:  for chaque objet  $n$  do {ligne 14}
20:    if  $n \in$  concept-fermé-maximal then {ligne 15}
21:      objets-traités[ $n$ ] = true{ligne 16}
22:    end if
23:  end for
24:  test-fin = true
25:  for chaque objet  $n$  and test-fin do
26:    if ! objets-traités[ $n$ ] then
27:      test-fin = false
28:    end if
29:  end for
30: end while

```

Dans l'algorithme 6 nommé *MinMaxC-Extend* la ligne 4 crée un concept appelé *XY*. Pour cela l'algorithme fait l'union des extensions des deux concepts *i* et *j* considérés et l'intersection de leurs intensions. Celle-ci se fait par le calcul du minimum et du maximum de chacun des attributs. Nous rappelons que le nombre d'attributs est fixe pour notre application puisque chaque objet possède un histogramme par trait.

La ligne 5 appelle la fonction de fermeture (algorithme 7) qui recherche dans le tableau de concepts unaires *U* et compare pour chaque concept, chacun de ses attributs à ceux du concept *XY* afin de vérifier s'ils sont compris entre le minimum et le maximum qui ont été calculés. Si oui alors l'objet est inclus dans le concept *XY*.

Le test de la ligne 6 vérifie que le concept n'a pas déjà été trouvé par la comparaison aux autres concepts déjà établis, et s'il est nouveau il sera rajouté à *Closed* et *[Pi]*.

Lorsque tout le tableau passé en paramètre a été analysé, on traite le nouveau tableau *Pi* qui contient les concepts qui viennent d'être calculés. Lors du premier appel de cette fonction par le *main*, le tableau passé en paramètre est *U*.

Algorithm 6 MinMaxC-Extend

Require: tableau de concepts [P] et taille de P

```

1: for chaque concept i de [P] do
2:   créer un tableau de concepts [Pi] vide
3:   for chaque concept  $j \geq i$  de [P] do
4:     créer un concept  $XY = \text{min-max}(i, j)$ 
5:     fermeture(XY)
6:     if  $XY \notin \text{Closed}$  then
7:       ajouter dans Closed
8:       ajouter dans [Pi]
9:     end if
10:  end for
11:  if Pi non vide then
12:    MinMaxC-Extend([Pi], nombre-concept-Pi)
13:  end if
14: end for

```

Dans le cadre de la connexion min-max (définition 34, page 87) certaines propriétés d'élagage ne sont plus appropriées. En effet les propriétés 2 et 3 (issues des 4 propriétés listées page 96) considèrent qu'une intension (respectivement une extension puisque Charm peut fonctionner selon les deux approches) qui serait incluse dans une autre peut être remplacée par l'intension qui la subsume (au sens binaire de la subsumption). Or dans notre jeu de données, les propriétés sont toujours de même longueur car il y a une propriété (histogramme) par trait. De plus, la correspondance min-max s'établit sur une relation de subsumption qui range tous les concepts selon la "largeur" (différence entre max et min) de ses propriétés. Et ainsi deux concepts *A* et *B* tels que $A \subseteq B$ existent et

Algorithm 7 Fermeture**Require:** un concept XY

```

1: for chaque concept  $c$  de  $U$  do
2:   booleen ferme = true
3:   for chaque propriété  $p$  do
4:     if  $((U[c]_{min}[p] < XY_{min}[p]) \text{ or } (U[c]_{max}[p] > XY_{max}[p]))$  then
5:       ferme = false
6:     end if
7:   end for
8:   if ferme then
9:     ajouter extension de  $c$  à  $XY$ 
10:  end if
11: end for

```

ne peuvent être fusionnés.

Définition 37 *Subsommation entre deux concepts fermés par une connexion min-max.*

Soit $A(ext_A, int_A)$ et $B(ext_B, int_B)$ deux concepts, alors $A \subseteq B$ si

- $ext_A \subseteq ext_B$
- ou
- $min_{int_A} \geq min_{int_B}$ et $max_{int_A} \leq max_{int_B}$

La propriété 1 consiste à rechercher et trouver un doublon du concept qui vient d'être calculé parmi les concepts calculés préalablement. Dans ce cas, nous ignorons ce concept. La propriété 4 considère qu'aucun doublon n'a été trouvé, ainsi dans ce cas, nous mémorisons ce concept. Ces deux propriétés sont traitées par les lignes 14 et 15 de l'algorithme 5, et par les lignes 6 et 7 de l'algorithme 6.

Enfin, le critère basé sur le support ne nous intéresse pas. En effet, nous considérons que tous les concepts sont pertinents quelle que soit la longueur de leur extension.

4.4 Test de validation des algorithmes

Afin de vérifier la justesse des modifications apportées aux algorithmes NextClosure et Charm, nous avons constitué un jeu de données simple dont le treillis peut être aisément déterminé manuellement. Nous présentons ces données binaires et le contexte histogramme correspondant, puis nous présenterons la liste des concepts que nous avons calculés manuellement, et enfin les résultats obtenus par les algorithmes.

4.4.1 Données

Contexte binaire

Le jeu de données utilisé pour valider les algorithmes comporte 6 objets et deux attributs dont le premier (A) est découpé en 3 modalités et le second (B) en 2 modalités. Les affinités valent 0, 1 ou 2. Ces données sont présentées par la table 4.4.

	A			B	
o1	0	1	2	0	1
o2	2	1	0	1	1
o3	2	2	2	0	2
o4	1	1	1	2	1
o5	0	2	0	1	2
o6	0	1	0	2	2

TABLE 4.4 – Données binaires pour la validation des algorithmes.

Contexte histogramme

Les affinités du contexte binaire sont rapportées en pourcentage puis on applique l'échelonnage histogramme. Les données ainsi obtenues sont présentées par la table 4.5.

	A0-33-66	A66-33-0	A33-33-33	A0-100-0	B0-100	B50-50	B66-33	B33-66
o1	1	0	0	0	1	0	0	0
o2	0	1	0	0	0	1	0	0
o3	0	0	1	0	1	0	0	0
o4	0	0	1	0	0	0	1	0
o5	0	0	0	1	0	0	0	1
o6	0	0	0	1	0	1	0	0

TABLE 4.5 – Données en histogrammes pour la validation des algorithmes.

4.4.2 Concepts

De ce contexte nous calculons l'ensemble des concepts (annexe A.6). Ce calcul a été opéré manuellement d'une part, puis avec chacun des deux algorithmes. Le contexte est assez conséquent pour que des erreurs dans les algorithmes entraînent une liste erronée soit sur le nombre de concepts, leur extension, ou les bornes de leur extension ; mais il reste de taille suffisamment raisonnable pour être traité à la main. Ce contexte n'a pour vocation que de valider la justesse des algorithmes. Les calculs humains et informatiques aboutissent à la même liste de concepts. Par la suite, les algorithmes s'accordent sur le nombre de concepts générés à partir de contextes réels beaucoup plus volumineux.

4.4.3 Délai d'exécution

Ces concepts ont été calculés sur un noyau linux 2.6.31-20 disposant de 1 Go de RAM et de deux CPU Intel Core2 T5600 cadencés à 1.83 GHz.

Les différents paramètres qui ont été évalués sur ces algorithmes sont : l'utilisation des connexions binaires et des connexions min-max, l'emploi ou non des seuils, et la recherche ou non de doublons. Cette recherche implique des balayages supplémentaires des données, i.e. des concepts déjà calculés.

Nous regardons dans quelle mesure l'ajout de cette recherche de doublons modifie l'exécution des algorithmes. La petite taille du jeu de test est intéressante pour la validation manuelle des algorithmes, mais elle est un inconvénient par ailleurs car ce jeu s'exécute trop rapidement pour évaluer une quelconque différence, nous avons donc utilisé un jeu de données plus grand contenant 15 objets et 110 histogrammes (issus de 15 traits). La table 4.6 présente les différences de temps d'exécution entre MinMaxNC et MinMaxC en fonction de la recherche ou non de doublons.

	MinMaxNC	MinMaxC
Doublons autorisés temps	32479 concepts 36,06 secondes	processus arrêté -
Doublons éliminés temps	32431 concepts 57,64 secondes	32431 concepts 82,21 secondes

TABLE 4.6 – Délai d'exécution des algorithmes en fonction de la prise en charge de la recherche de doublons.

On constate que l'algorithme MinMaxNC génère un nombre de doublons relativement faible : 48. Dans ce cas précis, la raison tient dans la structure des données qui génère ce faible nombre. En effet, si on traite un jeu de données de 21 objets (ceux du contexte ayant produit le treillis 3.5 page 58) et 42 histogrammes (correspondant aux 5 traits écologiques de ces objets) pour lequel les espèces ont donc davantage de caractéristiques communes, nous obtenons 38 486 concepts uniques (en 67, 26 secondes) et 70 470 concepts avec doublons (en 15,31 secondes) tel que présenté dans la table 4.7. Dans ce cas le nombre de doublons devient très important.

	nombre de concepts	délais
Doublons autorisés	38486 concepts	67,26 secondes
Doublons éliminés	70470 concepts	15,31 secondes

TABLE 4.7 – Délai d'exécution des algorithmes en fonction de la prise en charge de la recherche de doublons pour un contexte de 21 objets et 42 histogrammes

La recherche des doublons est un processus qui allonge considérablement le temps d'exécution des algorithmes, mais cette opération est nécessaire. Notamment pour Min-MaxC dont le nombre de doublons explose rapidement et engorge la mémoire capable de calculer les concepts.

4.5 Seuillage des histogrammes

La construction des treillis par une connexion de Galois min-max entraîne un traitement de toutes les paires d'histogrammes possibles et génère des treillis très volumineux. De plus, les temps de calculs pour la fermeture et la recherche de doublons sont proportionnels au nombre de fermés, et deux fermés identiques peuvent s'obtenir par la combinaison de deux intensions différentes ayant la même fermeture, à savoir le même couple (min,max).

De plus, pour rester en accord avec les préférences des biologistes, nos histogrammes sont dans un format en pourcentage, où la somme des affinités d'un trait est égale à 100%. Ainsi les variations de valeurs entre les affinités d'une même modalité entre les histogrammes min et max peut être importante : jusqu'à 100%. Le fermé est alors très général et n'est pas forcément pertinent.

Il existe plusieurs options de seuillage car, dans notre contexte, chaque objet possède une instance de chaque trait. On peut définir un seuil global (*seuil*), un seuil par trait (*seuil_t*) ou un seuil par modalité de chaque trait (*seuil_{t,i}*). L'idée est d'autoriser une différence qui soit "partageable" entre toutes les modalités d'un trait, de telle sorte que les modalités qui ont un faible écart puissent laisser une autre avoir un écart plus important. Sans cette flexibilité nous aurions des concepts trop restreints, avec des extensions de peu d'objets, et même très peu de concepts :

- seuillage sur les modalités. La contrainte impose que chaque modalité respecte la valeur du seuil : $\forall t, \forall i, a_{t,i}^{max} - a_{t,i}^{min} \leq \text{seuil}_{t,i}$
- seuillage sur les traits : la somme des écarts entre les modalités min et max d'un trait doit être inférieure à un seuil proportionnel au nombre de modalités du trait : $\forall t, \sum_{i=1}^{n_t} (a_{t,i}^{min} - a_{t,i}^{max}) \leq \text{seuil}_t$
- seuillage sur la totalité des traits de l'intension : $\sum_{t=1}^{|T|} \sum_{i=1}^{n_t} (a_{t,i}^{min} - a_{t,i}^{max}) \leq \text{seuil}$ avec n_t le nombre de modalités pour un trait t , $a_{t,i}$ l'affinité de la i^{me} modalité du trait t , et $|T|$ le nombre de traits.

Pour nos travaux, nous considérons le même seuil *seuil_{t,i}* pour chaque modalité, ce qui implique que le seuil pour un trait dépend du nombre de modalités : $\text{seuil}_t = \text{seuil}_{t,i} \times n_t$. Nous obtenons le seuil global sur le même principe : $\text{seuil} = \text{seuil}_t \times (\sum_{t=1}^{|T|} n_t)$.

Le cas le plus pertinent est d'envisager un seuillage au niveau du trait. En effet le seuillage sur les modalités est trop restrictif et celui sur la totalité de l'intension, trop laxiste. De plus le seuillage sur les traits permet de conserver la signification de la

conversion des histogrammes en pourcentage.

Détaillons cette conversion en nous appuyant sur un exemple.

4.5.1 Seuil brut par trait

Description et tests

Cette approche autorise l'utilisateur à choisir un seuil fondé sur un écart moyen par modalité. Prenons l'exemple suivant : un trait de 3 modalités. L'utilisateur choisit un seuil de 20%, cela fait $3 \times 20 = 60\%$ de marge d'erreur à répartir entre les histogrammes min et max.

Ainsi, si l'on prend deux histogrammes : (min).(33).(33).(33) et (max).(0).(50).(50), nous obtenons : (max-min).(33-0).(50-33).(50-33) = (max-min).(33).(17).(17) $\Rightarrow 33 + 17 + 17 = 67 \geq 60$ donc ce concept n'est pas retenu. Il devient alors inutile de calculer les écarts pour les autres traits (car tous les traits d'un concept doivent correspondre à ce seuil), ni de calculer la fermeture. Le gain de temps de calcul est alors très significatif puisqu'il permet même de faire aboutir des calculs qui ne terminent pas par des méthodes standard à cause du volume de concepts générés (rappelons que les fonctions de fermeture sont croissantes monotones). De plus, par le seuillage, les concepts sont plus en adéquation avec ce que recherche l'utilisateur.

Nous avons effectué des tests sur le jeu de données complet : 47 macrophytes, 10 traits biologiques et 5 traits écologiques. La table 4.8 présente le nombre de concepts et le temps d'exécution pour les obtenir en fonction du seuil.

Seuil (en %)	Nombre de concepts	Temps d'exécution
20	47	0,02 sec
30	49	0,02 sec
40	68	0,03 sec
50	611	0,13 sec
60	966	0,21 sec
65	1710	0,4 sec
70	17152	21,39 sec

TABLE 4.8 – Calcul des concepts en fonction d'un seuil par modalité.

Limites posées par le nombre de modalités

En fait, le seuillage est biaisé par le nombre de modalités d'un trait. En effet un trait ayant 2 modalités va avoir des valeurs d'affinités d'une grande variance²⁸ puisque leur somme doit faire 100%. Par ailleurs les traits avec un nombre important de modalités

28. Elle permet de caractériser la dispersion des valeurs par rapport à la moyenne.

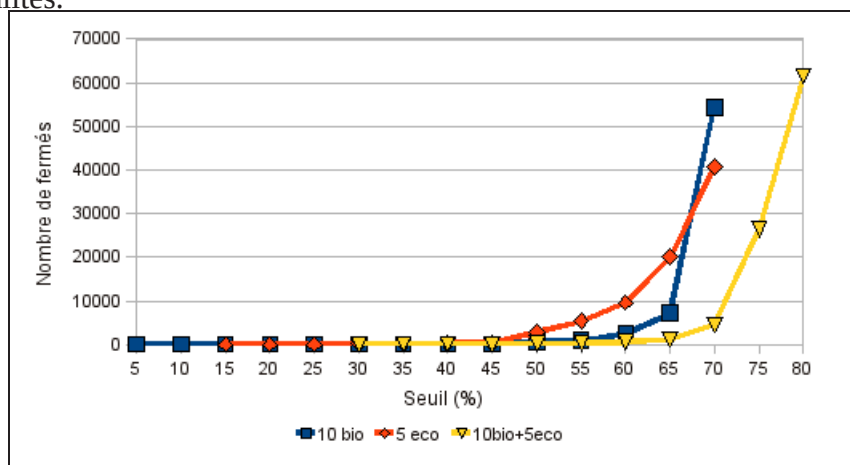
vont permettre d'accumuler une grande marge d'erreur possible. Cette approche n'est donc pas optimale. De plus le seuil est fixé arbitrairement par l'utilisateur, et n'est pas forcément en adéquation avec les valeurs réelles du jeu de données.

4.5.2 Seuil fondé sur les écarts réels entre les affinités

Afin d'ajuster le seuil aux valeurs réelles des données, nous nous intéressons aux écarts qui existent pour chaque modalité. Nous regardons donc pour chaque trait, modalité par modalité quelle est la plus petite et la plus grande valeur que chaque modalité peut prendre. Nous calculons la différence entre ces deux valeurs pour chaque modalité et sommes ces différences pour toutes les modalités d'un trait. Cet écart est donc le seuil réel maximal qui existe entre les modalités pour chaque trait. Nous autorisons l'utilisateur à définir un pourcentage de cet écart afin de doser la "largeur" des concepts et leur nombre. Le tableau 4.9 illustre ce seuillage par un exemple simple.

Nous avons testé ce seuillage sur trois contextes de différentes tailles : un contexte avec les 10 traits biologiques, un contexte avec les 5 traits écologiques, et enfin un contexte avec tous les traits de vie. La table 4.10 présente le nombre de concepts obtenus en fonction du pourcentage de l'écart réel autorisé entre les valeurs des modalités. Le tableau 4.11 présente les délais d'exécution correspondant à la table 4.10. Ces deux tables sont représentées par les graphes 4.3 et 4.4 respectivement.

FIGURE 4.3 – Nombre de concepts en fonction d'un seuil selon la variance des valeurs des modalités.



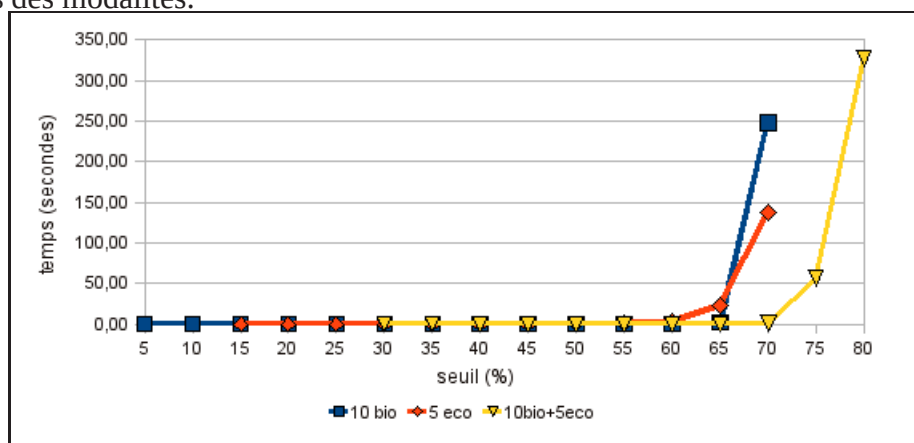
4.5.3 Seuil par la variance et l'écart-type

Le fait que l'utilisateur puisse paramétrer le seuil permet de doser le résultat à sa convenance. Cependant l'utilisateur se trouve généralement démuné lorsqu'il doit faire

	A			B	
o1	0	1	2	0	1
o2	0	1	1	2	1
o3	2	2	2	2	2
o1(%)	0	33	66	0	100
o2(%)	0	50	50	66	33
o3(%)	33	33	33	50	50
min	0	33	50	0	33
max	33	50	66	66	100
max-min	33	17	16	66	66
$\Sigma(max - min)$	66			132	
Exemple pour fermer o1 avec o2					
%écart=50%	$66 \times 0,5 = 33$			$132 \times 0,5 = 66$	
min(o1-o2)	0	33	50	0	33
max(o1-o2)	0	50	66	66	100
$\Sigma(max(o1 - o2) - min(o1 - o2))$	$0+17+16=33$ $33 \leq 33 \Rightarrow$ accepté			$66+66=132$ $132 \geq 66 \Rightarrow$ rejeté	

TABLE 4.9 – Illustration de fermeture par un seuillage basé sur les écarts réels entre les affinités.

FIGURE 4.4 – Délai de calcul des concepts en fonction d'un seuil selon la variance des valeurs des modalités.



Données	10 traits bio	5 traits eco	10 traits bio + 5 traits eco
Nb histogrammes	49	123	172
% de l'écart			
5	47		
10	48		
15	48	45	
20	48	53	
25	50	71	
30	53	113	47
35	63	182	48
40	97	282	58
45	179	365	70
50	616	2875	218
55	951	5389	279
60	2499	9570	623
65	7233	20074	1134
70	54327	40811	4574
75			26345
80			61541

TABLE 4.10 – Nombre de concepts selon un pourcentage de l'écart réel entre les valeurs des modalités.

Données	10 traits bio	5 traits eco	10 traits bio + 5 traits eco
% de l'écart			
5	0,01		
10	0,01		
15	0,02	0,00	
20	0,01	0,00	
25	0,01	0,00	
30	0,01	0,00	0,03
35	0,02	0,01	0,02
40	0,02	0,01	0,03
45	0,03	0,02	0,02
50	0,07	0,36	0,06
55	0,12	1,02	0,05
60	0,44	3,02	0,13
65	2,39	22,81	0,23
70	248,12	137,26	1,40
75			56,57
80			326,75

TABLE 4.11 – Délai de calcul des concepts en fonction du pourcentage d'écart réel entre les valeurs des modalités.

un choix de ce genre car définir un seuil arbitraire est toujours délicat. Ainsi nous cherchons à fournir à l'utilisateur une liste de concepts qui soit optimale sans qu'il ait à intervenir. Nous sommes concernés par les données les plus moins spécifiques donc les moins dispersées par rapport à leur valeur médiane. Pour cela, nous nous intéressons à calculer la variance qui est une mesure caractérisant la dispersion d'un échantillon.

$$1/n \sum_{i=0}^n (v_i - m)^2$$

où n est le nombre d'éléments, v est la valeur de chaque élément et m est la moyenne des valeurs de tous les éléments. Nous calculons la variance pour chaque modalité, et le seuil que nous considérons pour le trait est égal à la somme des variances des modalités du trait.

Les seuils établis sur la variance que nous obtenons sur le jeu de données des 5 traits écologiques sont pour chacun des traits : 295, 1047, 641 180 et 1636 ; alors que les écarts maximaux réels pour ces traits sont 200, 250, 400, 157, 400. La variance va donc calculer tous les concepts, ce qui n'est pas ce que nous recherchons.

Nous nous interrogeons donc bien évidemment sur les résultats provenant de l'écart-type puisqu'il mesure la dispersion d'une série de valeurs autour de leur moyenne. L'écart-type correspond à la racine carrée de la variance, les seuils sont donc : 17, 32, 25, 13, 40. De tels écarts sont trop faibles, et les seuls concepts que nous obtenons, sont les concepts $\vee$ -irréductibles.

La solution statistique est pertinente, car elle est la plus à même de fournir un résultat optimal, et la variance et l'écart-type sont les approches les plus évidentes pour aborder ce genre de problématique. Cependant si en théorie elles sont prometteuses, leurs résultats ne répondent pas à nos attentes. Ainsi nous nous cherchons une approche statistique plus quantitative qui s'établisse à partir d'une quantité proportionnelle de valeurs à prendre en compte et qui en déduise le seuil correspondant : les quantiles.

4.5.4 Seuil par les quantiles

Nous présentons les quantiles sur un cas précis : les quartiles, à partir duquel nous illustrons la démarche avec un exemple simple. les quartiles qui sont des 4-quantiles, constituent une mesure de statistique descriptive : l'échantillon de la population est trié et divisé en 4 parts égales. Un quartile est chacune des 3 valeurs qui divisent cet échantillon. Le deuxième quartile est donc celui du centre et représente la valeur médiane (celle séparant l'échantillon en deux sous-échantillons de même quantité). Ce qui nous intéresse est de calculer l'écart interquartile qui est la différence entre le troisième quartile (Q3) et le premier (Q1). Entre ces deux quartiles se trouve donc la moitié des individus de la population : ceux qui sont les plus proches de leur valeur centrale.

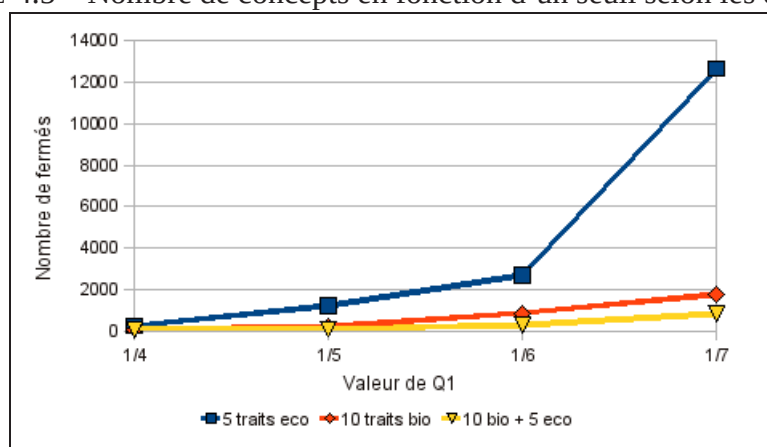
Considérons l'exemple concret de la table 4.12 pour illustrer les quartiles. Il s'agit de la modalité "plante étouffée" du trait "tolérance à la sédimentation". Les 46 taxons sont

rangés par ordre croissant de la valeur de leur affinité. On note que la valeur médiane vaut 0 qui sépare l'ensemble en deux sous-ensemble de même taille. Q1 (vaut 0) et Q3 (vaut 33) sont les valeurs médianes de chacun de ces deux sous-ensembles.

Pour chaque trait nous cherchons, modalité par modalité les valeurs de Q3 et Q1, et nous sommions les écarts interquartiles de chaque modalité pour définir le seuil pour le trait. Les résultats obtenus sont 244 concepts pour le jeu de données des 5 traits écologiques, 74 concepts pour celui des 10 traits biologiques et 53 concepts pour celui des 5 traits écologiques et 10 traits biologiques regroupés.

Les quartiles sont trop restrictifs, en effet, ils écartent trop de concepts. Nous devons donc ajuster les bornes que nous appelons toujours Q1 et Q3 par abus de langage. Nous avons effectué les tests en choisissant d'écarter 1/5, 1/6 et enfin 1/7 des valeurs à chaque extrémité. Le graphique de la figure 4.5 présente les résultats obtenus.

FIGURE 4.5 – Nombre de concepts en fonction d'un seuil selon les quantiles.



De ces tests, résumés dans le tableau 4.13, le seuillage par quantile ressort comme le plus efficace, car il est simple à paramétrer pour l'utilisateur, et que la qualité de ses résultats sont meilleurs car il rejette les informations les moins pertinentes.

4.5.5 Comparatif des deux algorithmes seuillés par quantiles

Nous avons implémenté la connexion min-max et les seuillages dans deux algorithmes : MinMaxNC et MinMaxC.

Les tests effectués sur un sous-ensemble du contexte biologique contenant 27 objets et 5 traits écologiques (soit 44 histogrammes) montrent que MinMaxNC est plus efficace pour traiter les données sans seuillage. En effet, MinMaxNC génère 33917 concepts en 32,05 secondes, alors que MinMaxC termine en 82,75 secondes.

ALIP	0	< 1/4 des données >	< 1/7 >
BERE	0		
CALO	0		
HIPV	0		
HYDM	0		
IRIP	0		
JUNA	0		
LEMM	0	< 1/4 des données >	Q1 (quantile 1/7)
LEMT	0		>
MYOP	0		
NUPL	0	Q1 (quartile)	
NYMA	0	< données utilisées >	< données utilisées >
OENF	0		
PHAA	0		
PHRG	0		
PLGH	0		
RANU	0		
SAGS	0		
SEFC	0		
TYPL	0		
UTRV	0		
VERA	0		
VERB	0	Médiane	Médiane
ZANP	0	< données utilisées >	>
PTCO	20		
PTCR	20		
PTLU	20		
PTNA	20		
PTNO	20		
PTPE	20		
RANC	25		
RANT	25		
CALP	33		
CERD	33		
HOTP	33	Q3 (quartile)	
MENA	33	< données >	< données utilisées >
ELOC	40		
ELOE	40		
ELON	40		Q3 (quantile 1/7)
GROD	40		< 1/7 >
PTPU	40		
ELEA	50		
MYRS	50		
MYRV	50		
NASO	50		
RORA	50		

TABLE 4.12 – Illustration des quartiles par les données de la modalité "plante étouffée" du trait "tolérance à la sédimentation".

	Taille potentielle				Forme de croissance			
	<0.08m	0.08-0.3 m	0.3 -1m	1-5 m	non ancrée	ancrée, feuilles flottantes	ancrée, tissus de soutien	ancrée, submerse
ALIP (%)	0	25	50	25	0	0	100	0
BERE (%)	20	40	40	0	0	0	33	66
CALO (%)	0	20	40	40	0	40	20	40
CALP (%)	0	50	50	0	0	33	0	66
HYDM (%)	0	0	100	0	0	0	100	0
min	0	0	40	0	0	0	0	0
max	20	50	100	40	0	40	100	66
max-min	20	50	60	40	0	40	100	66
$\sum(max - min)$ (écart permis par trait)	170				206			
moyenne	80	370	630	345	0	406	2172	1099
variance								
$\sum(variance)$ (écart permis par trait)	1425				3677			
Ecart-type	9	19	25	19	0	20	47	33
$\sum(ecart - type)$ (écart permis par trait)	72				100			
médiane	0	25	50	0	0	0	33	40
Q1 (quartile)	0	20	40	0	0	0	20	0
Q3 (quartile)	0	40	50	25	0	33	100	66
Q3-Q1 (quartile)	0	20	10	25	0	33	80	66

TABLE 4.13 – Résumé des différents seuillages par un exemple.

Cependant, avec des contextes conséquents, l'utilisation de seuils est nécessaire. Par ailleurs, les tests effectués sur plusieurs seuils montrent que le plus efficace et le plus pertinent est celui basé sur les quantiles. Nous avons donc implémenté la connexion min-max seuillée par les quantiles dans les deux algorithmes. Leur exécution sur le jeu de données précédent montrent que MinMaxC est plus efficace car MinMaxNC génère 1034 concepts en 0,42 seconde et MinMaxC en 0,06 seconde (pour un quantile = 1/6).

Nous présentons dans la table 4.14 des tests comparatifs des deux algorithmes menés sur le contexte intégral des macrophytes, puisque le seuillage permet d'aboutir ces calculs. Il s'agit des 47 espèces et de leur 15 traits de vie (10 biologiques et 5 écologiques) ce qui correspond à 181 histogrammes. Ces tests confirment que MinMaxC est le plus efficace pour traiter les seuils.

Ces résultats nous permettent de sélectionner MinMaxC et le seuillage par quantile pour le traitement des contextes complets des macrophytes et également des invertébrés. Le chapitre suivant développe cette mise en application.

Quantile	Nombre de concepts	MinMaxNC	MinMaxC
1/7	886	4,33 s	0,2 s
1/8	2301	11,49 s	0,56 s
1/9	2301	11,78 s	0,55 s

TABLE 4.14 – Comparaison des temps d'exécution des algorithmes MinMaxNC et MinMaxC seuillés par les quantiles.

4.6 Conclusion

Dans ce chapitre nous définissons les notions d'histogramme, d'échelonnage histogramme, de connexions min-max et leur seuillage qui s'appuient sur les définitions présentées dans le chapitre 2. Ces définitions nous permettent de prendre en compte un traitement flou de nos données comme cela a été mis en œuvre dans le chapitre suivant. Ces connexions ont été implémentées dans deux algorithmes : MinMaxNC et MinMaxC que nous décrivons et dont nous confrontons les performances. Leur efficacité est dépendante de l'utilisation des techniques de seuillage. En effet, si le contexte ne réclame pas l'utilisation du seuillage, alors MinMaxNC est le plus performant, en outre, si le seuillage est nécessaire, MinMaxC est plus efficace.

Chapitre 5

Applications à la biologie

Sommaire

5.1	Macrophytes	117
5.1.1	Processus de mise en évidence des traits de vies caractéristiques de groupes taxonomiques	117
5.1.2	Application	118
5.2	Invertébrés	122
5.2.1	Groupes biotypologiques d'invertébrés	122
5.2.2	Groupes d'invertébrés par guidage latticiel	123
5.2.3	Groupes d'invertébrés selon la classification de P. Usseglio-Polatera et al.	127
5.3	Conclusion	133

Dans cette partie, nous nous intéressons à appliquer l'algorithme MinMaxC avec un seuillage par quantiles sur des données biologiques : des macrophytes et des invertébrés. L'objectif est d'obtenir des corrélations entre les traits biologiques et écologiques communs à des ensembles de taxons.

5.1 Macrophytes

5.1.1 Processus de mise en évidence des traits de vies caractéristiques de groupes taxonomiques

Nous nous intéressons à des espèces différentes de macrophytes pouvant être révélatrices de la même information, à savoir l'état écologique de leur environnement. Ainsi, nous avons recherché parmi les espèces de macrophytes, celles qui possèdent des traits de vie semblables ou proches et qui de ce fait pourraient se remplacer l'une l'autre dans

des conditions écologiques données.

Nous débutons par la construction par les connexions min-max d'un treillis complet, comportant les 47 espèces et leurs 15 traits de vie et 53 modalités correspondant à 181 attributs histogrammes. Le seuillage choisi est celui par quantile avec un seuil égal à 1/8, générant 2433 concepts. Ce treillis met en évidence les espèces "isolées" (qui partagent des traits avec une ou deux autres) et des espèces "collectives" (qui appartiennent à beaucoup de concepts).

Avec Michèle Trémolières, nous extrayons de ce treillis des groupes de taille moyenne (en intension et en extension) de telle sorte que les espèces qui les composent forment un ensemble cohérent selon les connaissances de l'écologie des espèces.

La seconde étape du processus permet de valider ou d'invalidier ces choix. Il consiste à constituer un contexte pour chacun de ces groupes tel que l'ensemble des objets soit les espèces du groupe et que l'ensemble des attributs soient l'ensemble des traits de vie que ces espèces possèdent. Ensuite pour chacun de ces sous-contextes, un treillis est généré et nous relevons les concepts caractéristiques chaque groupe.

Par ailleurs, au cours des analyses de ces groupes, nous nous sommes aperçues que certains groupes que nous distinguions initialement, tels que {*Berula erecta*, *Mentha aquatica*, *Myosotis palustris*, *Nasturtium officinale*, *Phalaris arundinacea*, *Veronica beccabunga*} et {*Hottonia palustris*, *Juncus articulatus*, *Veronica beccabunga*, *Nasturtium officinale*, *Myosotis palustris*} devaient être fusionnés ; ou encore que d'autres tels que {*Nymphaea alba*, *Potamogeton crispus*, *Potamogeton natans*, *Veronica beccabunga*, *Myosotis palustris*} devaient être supprimés car ils avaient finalement peu de caractéristiques communes.

5.1.2 Application

Nous commençons par présenter les 8 groupes qui ont été formés à partir du treillis complet :

- Groupe 1 : {*Callitriche obtusangula*, *Ceratophyllum demersum*, *Elodea canadensis*, *Elodea nuttallii*, *Myriophyllum spicatum*, *Myriophyllum verticillatum*}
- Groupe 2 : {*Berula erecta*, *Hottonia palustris*, *Juncus articulatus*, *Mentha aquatica*, *Myosotis palustris*, *Nasturtium officinale*, *Phalaris arundinacea*, *Veronica beccabunga*}
- Groupe 3 : {*Alisma plantago-aquatica*, *Iris pseudacorus*, *Mentha aquatica*, *Myosotis palustris*, *Phalaris arundinacea*, *Sagittaria sagittifolia*, *Sparganium emersum*}
- Groupe 4 : {*Nuphar lutea*, *Potamogeton crispus*, *Potamogeton lucens*, *Potamogeton natans*, *Potamogeton nodosus*, *Ranunculus circinatus*, *Ranunculus fluitans*}
- Groupe 5 : {*Ceratophyllum demersum*, *Groenlandia densa*, *Oenanthe fluviatilis*, *Potamogeton pectinatus*, *Ranunculus fluitans*}

- Groupe 6 : {*Callitriche obtusangula*, *Callitriche platycarpa*, *Potamogeton pectinatus*, *Ranunculus trichophyllus*, *Veronica anagallis-aquatica*, *Zannichellia palustris*}
- Groupe 7 : {*Berula erecta*, *Myosotis palustris*, *Nymphaea alba*, *Potamogeton coloratus*, *Potamogeton crispus*, *Potamogeton natans*, *Veronica beccabunga*}
- Groupe 8 : {*Lemna gibba*, *Lemna minuta*, *Lemna minor*, *Lemna trisulca*}

Nous détaillons pour chacun de ces groupes (pour lequel un nouveau treillis est constitué) quelles sont les espèces et leurs traits de vie qui ont été retenus comme les plus caractéristiques du groupe. Enfin, une analyse des plantes retenues est faite pour chaque groupe, permettant de considérer leur pertinence et intérêt.

- Groupe 1 : *Callitriche obtusangula*, *Elodea canadensis*, *Elodea nuttallii*, *Myriophyllum spicatum* et *Myriophyllum verticillatum*.

Ces espèces principalement mésotrophes⁽⁶⁴⁾ à eutrophes se caractérisent par une forme de croissance ancrée submerse, une forte flexibilité ($> 300^\circ$), une phénologie⁽³⁹⁾ sur toute l'année, un index morphologique faible (1), une taille potentielle entre 0,08 et 1 mètre, des organes pérennes, une période de reproduction en juin-juillet, une reproduction végétative par fragments non spécialisés et elles fleurissent.

Ces plantes se retrouvent fréquemment ensemble dans le même habitat.

- Groupe 2 : *Berula erecta*, *Hottonia palustris*, *Juncus articulatus*, *Mentha aquatica*, *Nasturtium officinale*, *Phalaris arundinacea*, *Veronica beccabunga*.

Ces espèces mésotrophes⁽⁶⁴⁾ avec une tolérance à la sédimentation par un enracinement et une bonne tolérance aux variations d'humidité se caractérisent par une dispersion potentielle élevée, une flexibilité moyenne ($> 45^\circ$), une floraison, un indice morphologique intermédiaire (2) pour une taille entre 0,3 et 1 mètre, des organes pérennes et une reproduction végétative principalement par rhizomes ou stolons.

Ces plantes sont ubiquistes, ce qui signifie qu'on les trouve facilement un peu partout. De ce fait, elles ne sont donc pas de bons indicateurs.

- Groupe 3 : *Alisma plantago-aquatica*, *Iris pseudacorus*, *Sagittaria sagittifolia*, *Sparganium emersum*.

Ces espèces eutrophes⁽⁶⁴⁾, tolérantes à la sédimentation grâce à un enracinement fort et tolérantes aux matières organiques se caractérisent par une forme de croissance ancrée par tissu de soutien, une dispersion élevée, une faible flexibilité (entre 10 et 45°), une phénologie⁽³⁹⁾ pendant la période de reproduction uniquement, une floraison, des organes pérennes, une période de reproduction surtout en juin-juillet avec une reproduction végétative essentiellement par rhizomes ou stolons.

et une taille entre 0,3 et 1 mètre.

Ces espèces sont caractéristiques d'eaux calmes, peu profondes, quoique certaines d'entre elles (*Sagittaria sagittifolia* et *Sparganium emersum*) puissent vivre totalement immergées.

- Groupe 4 : *Potamogeton crispus*, *Potamogeton lucens*, *Potamogeton natans*, *Potamogeton nodosus*, *Ranunculus circinatus*, *Ranunculus fluitans*.

Ces espèces tolérantes à la sédimentation par un enracinement variable plutôt fort se caractérisent par une forte flexibilité ($> 300^\circ$), une phénologie⁽³⁹⁾ pendant la période de reproduction uniquement, une floraison, un indice morphologique conséquent (2-3) pour une taille entre 0,3 et 1 mètre, des organes pérennes et une reproduction végétative variée qui excepte cependant les bulbes et tubercules.

Ces espèces sont des hydrophytes⁽²²⁾ strictes à large amplitude écologique vis-à-vis de la trophie, cependant elles vivent dans des milieux plutôt eutrophes mais jamais oligotrophes⁽⁶⁴⁾.

- Groupe 5 : *Ceratophyllum demersum*, *Groenlandia densa*, *Oenanthe fluviatilis*, *Potamogeton pectinatus*, *Ranunculus fluitans*.

Ces espèces eutrophes tolérantes à la sédimentation par un enracinement moyen, voire aucun (*Ceratophyllum demersum*) se caractérisent par une dispersion importante, une forte flexibilité ($> 300^\circ$), une phénologie⁽³⁹⁾ pendant la période de reproduction uniquement, une floraison rare, des organes pérennes et une période de reproduction entre juin et septembre.

Ces espèces sont essentiellement reliées par leur tolérance à un niveau trophique élevé, et à une bonne tolérance à la sédimentation.

- Groupe 6 : *Callitriche obtusangula*, *Callitriche platycarpa*, *Potamogeton pectinatus*, *Ranunculus trichophyllus*, *Veronica anagallis-aquatica*, *Zannichellia palustris*.

Ces espèces surtout mésotrophes tolérantes à la sédimentation par un enracinement plutôt moyen ne résistent pas aux fortes variations d'humidité. Elles se caractérisent par une forme de croissance ancrée submerse, une dispersion élevée, une flexibilité moyenne ($> 45^\circ$), une floraison.

Ces espèces présentent une bonne tolérance à la trophie essentiellement mésotrophe, pouvant être occasionnellement eutrophe ou hypertrophe, mais jamais oligotrophe.

- Groupe 7 : *Berula erecta*, *Myosotis palustris*, *Nymphaea alba*, *Potamogeton coloratus*, *Potamogeton crispus*, *Potamogeton natans*, *Veronica beccabunga*.

Ces espèces essentiellement oligotrophes à eutrophes tolérantes à la sédimentation par un enracinement plutôt moyen se caractérisent par une dispersion élevée, une flexibilité moyenne ($> 45^\circ$), une floraison, des organes pérennes, une taille entre 0,3 et 1 mètre et une reproduction végétative surtout par rhizomes ou stolons.

Ce groupe est essentiellement oligotrophe, certaines espèces étant plutôt mésotrophes à eutrophes, mais jamais hypertrophes.

- Groupe 8 : *Lemna gibba*, *Lemna minuta*, *Lemna minor*, *Lemna trisulca*.

Ces espèces ont pour particularité d'appartenir au même genre *Lemna* (lentilles d'eau). Les espèces *Lemna gibba* et *Lemna minuta* ont été ajoutées au contexte afin de pouvoir constituer ce groupe de lentilles par intérêt de l'expert. Comme le groupe 6 elles ne sont absolument pas oligotrophes. Elles sont tolérantes à la sédimentation car elles flottent en surface ou entre deux eaux. Elles apprécient un environnement contenant plus de 10% de matières organiques. Elles se caractérisent par une forme de croissance non ancrée, une dispersion élevée, aucune floraison, un index morphologique minime (0) pour une taille inférieure à 8 cm, et une reproduction végétative plutôt par bulbilles, turions ou apex dormants.

Ce groupe est intéressant là encore pour sa tolérance à la trophie qui est mésotrophe ou eutrophe, mais jamais oligotrophe.

La constitution de ces groupes a permis de mettre en évidence que certains traits sont importants et récurrents tels que la flexibilité, la forme de croissance, les organes pérennes, et parmi les traits écologiques : la trophie et la sédimentation (cf les résultats de l'AFM) mais que d'autres ne sont jamais ou rarement caractérisés par les groupes constitués, tels que la stabilité de l'eau ou la tolérance à la matière organique pour lesquelles les plantes sont équitablement réparties et ne manifestent donc pas de préférence. On pourrait supposer que cela vient du nombre de modalités des traits, mais en l'occurrence, la stabilité de l'eau comporte 4 modalités et la tolérance à la matière organique en comprend 3. La différenciation entre les traits peut également provenir de la méthode puisque la composition des groupes inclut une participation humaine pour une évaluation des espèces afin justement de pouvoir constituer les groupes.

5.2 Invertébrés

Nous abordons, dans cette partie, trois axes de classification des invertébrés. L'analyse de ces classifications est toujours en cours et leur achèvement figure en tant que travaux futurs de cette thèse.

Par ailleurs, l'approche sur les invertébrés, quoique semblable à celle sur les macrophytes, diffère légèrement de par la taille du contexte, qui compte 122 taxons, 20 traits et 97 modalités ; et des classifications ont été déjà établies sur les invertébrés sur lesquelles nous nous appuyons.

Nous rappelons que les données concernent uniquement les taxons qui ont été relevés dans la plaine d'Alsace grâce au projet *Indices*. La sélection de ces données est présentée dans la partie 1.4.3, page 24.

Le premier axe de classification, cherche à classer les invertébrés selon 9 groupes biotypologiques définis par Verneaux [68]. Le second axe, reprend la méthodologie appliquée aux macrophytes pour une classification par les treillis. Enfin, le troisième axe, se réfère aux travaux de P. Usseglio-Polatera et al. qui ont défini à travers deux articles des groupes en fonction des traits biologiques et d'autres en fonction des traits écologiques [66], et également des groupes en fonctions de tous les traits de vie [67]. Nous abordons deux approches. La première consiste à extraire des concepts (selon la méthode appliquée aux macrophytes) du treillis des traits biologiques d'une part, et du treillis des traits écologiques d'autre part, puis à comparer ces concepts aux analyses de P. Usseglio-Polatera et al. Par la seconde approche, nous nous intéressons à sa classification des traits de vie. Nous commençons donc par construire les contextes avec les taxons correspondant aux niveaux systématiques supérieurs qui forment les groupes issus de cette classification. Puis nous relevons les concepts dont les traits sont particulièrement caractéristiques et confrontons les analyses.

5.2.1 Groupes biotypologiques d'invertébrés

L'approche utilisée pour les macrophytes consiste à générer les concepts sur l'ensemble du contexte permettant de guider l'expert sur la constitution de groupes d'espèces.

Constituer le treillis pour l'intégralité du contexte des invertébrés, qui est l'étape préliminaire à la constitution des groupes, produit une quantité de concepts trop importante et fastidieuse à analyser. Pour rendre le nombre de concepts exploitable, le seuil de l'algorithme doit être important, et les concepts générés ont des intensions très étroites, et donc des extensions peu fournies. Ce qui est pénalisant pour l'analyse que nous faisons.

Cependant, dans sa thèse [68], Verneaux s'intéresse à un classement des taxons selon leur présence au sein de 9 niveaux biotypologiques qu'il définit et qui sont répartis de l'amont à l'aval. Nous avons donc décidé d'œuvrer en nous ramenant à ces travaux et de traiter, de la même façon que pour les macrophytes, les 9 groupes correspondant

aux taxons de la plaine d'Alsace. Ces groupes sont présentés dans l'annexe A.3. Nous avons donc généré les concepts pour ces 9 nouveaux contextes. Pour chaque contexte, de l'amont à l'aval, nous avons sélectionné les taxons les plus caractéristiques et ceux qui apparaissent à partir d'un certain niveau en descendant vers l'aval. Par exemple *Baetis* n'a été retenu comme taxon représentatif pour aucun groupe car il est commun tout au fil de l'eau, alors que *Paraleptophlebia* n'apparaît qu'à partir du niveau 2 et a donc été retenu pour ce groupe, même si on peut le retrouver plus bas en aval.

Pour chacun des niveaux, nous avons ainsi identifié les taxons suivants :

- Niveau 1 : Capnia, Ephemerella, Rhithrogena, Planorbidae
- Niveau 2 : Brachyptera, Paraleptophlebia, Polycentropus, Protonemura, Taeniopteryx
- Niveau 3 : Ecdyonurus, Hydropsyche, Philopotamus, Rhithrogena
- Niveau 4 : Cheumatopsyche, Chloroperla, Heptagenia, Lepidostoma, Psychomyia
- Niveau 5 : Caenis, Goera, Hydropsyche, Micrasema
- Niveau 6 : Lepidostoma, Micrasema, Mystacides
- Niveau 7 : Aucun taxon n'est vraiment révélateur de ce niveau
- Niveau 8 : Bithynia, Ecnomus, Habrophlebia, Leptocerus, Lype, Physa, Planorbidae
- Niveau 9 : Calopteryx, Coenagrionidae, Rhyacophila

L'étape suivante de la démarche consiste à rechercher les traits correspondants aux taxons sélectionnés aux différents niveaux. Pour cela nous avons extrait les concepts du contexte intégral et avons recherché ceux contenant les taxons de chacun des niveaux. Malheureusement cette recherche est infructueuse : les taxons ne sont pas assez semblables pour former des concepts pertinents. Nous en déduisons donc que ce n'est pas parce que des individus vivent au même niveau d'un cours d'eau qu'ils partagent les mêmes traits de vie.

5.2.2 Groupes d'invertébrés par guidage latticiel

Nous nous intéressons donc à suivre la démarche effectuée sur les macrophytes en l'appliquant aux invertébrés. Le contexte intégral étant toujours très volumineux à explorer, nous avons recherché des groupes sur le contexte restreint aux traits écologiques et celui restreint aux traits biologiques. Nous avons ainsi distingué 7 groupes (3 groupes du contexte des traits écologiques et 4 groupes du contexte des traits biologiques) que nous listons ci-après :

- Contexte des traits écologiques (123 taxons, 10 traits, 48 modalités) :
 - Groupe 1 : Baetis, Calopteryx, Elmis, Esolus, Galba, Gammarus, Lepidostoma, Leuctra, Limnephilinae, Onychogomphus, Paraleptophlebia, Polycentropus, Sericostoma, Silo, Taeniopteryx
 - Groupe 2 : Asellidae, Caenis, Coenagrionidae, Corixidae, Cynus, Erpobdellidae, Glossiphoniidae, Hydroporinae, Leptocerus, Mystacides, Nepinae, Planorbidae, Tabanidae, Veliinae

- Groupe 3 : Bithynia, Caenis, Calopteryx, Ceratopogonidae, Ephemerella, Goera, Neureclipsis, Psychodidae, Sphaerium, Valvata
- Contexte des traits biologiques (123 taxons, 10 traits, 49 modalités) :
 - Groupe 4 : Acentrella, Drusinae, Ecdyonurus, Leptocerus, Micrasema, Mystacides, Planorbidae, Rhithrogena, Riolus, Sericostoma, Setodes, Silo
 - Groupe 5 : Goera, Habrophlebia, Heptagenia, Leuctra, Nemoura, Nemurella, Paraleptophlebia, Potamanthus, Protonemura, Taeniopteryx
 - Groupe 6 : Electrogena, Goera, Heptagenia, Leuctra, Paraleptophlebia, Rhithrogena, Torleya
 - Groupe 7 : Ecdyonurus, Lepidostoma, Leptocerus, Mystacides, Nemoura, Nemurella, Sericostoma, Setodes,

Nous savons que les individus des groupes 1 à 3 ont tendance à partager les mêmes traits écologiques et que ceux des groupes 4 à 7 ont tendance à partager les mêmes traits biologiques. Cependant nous nous intéressons à connaître tous les traits de vie similaires et donc caractéristiques de chacun de ces groupes. Nous relevons ainsi les traits principaux les plus marqués. Nous construisons donc les 7 contextes (issus du contexte des traits écologiques *ou* biologiques) avec la totalité de leurs traits de vie afin d'identifier quels taxons et quels traits sont les plus caractéristiques de ces groupes. Pour l'étape d'interprétation, nous avons indiqué entre parenthèses la valeur du groupe indicateur utilisé pour le calcul de l'IBGN (décrit dans la partie 1.1.3) de chaque taxon. Nous rappelons qu'il s'agit d'une note variant de 1 (taxons les plus polluo-résistants) à 9 (taxons les plus polluo-sensibles). L'IBGN n'attribue pas de valeur indicatrice à certains taxons à cause de la grande hétérogénéité de la polluo-sensibilité dans certains groupes, ou de l'absence de connaissance par exemple. Nous indiquons ces taxons par (-). Cette analyse peut être critiquable par le fait que les notes sont attribuées à une famille, car il est connu que dans certaines familles, certains genres sont plus polluo-sensibles ou plus polluo-résistants que les autres membres de la famille. Cependant, l'IBGN fonctionne selon ce procédé, et est un indicateur fortement utilisé. Nous suivons donc tout de même son système de notation.

Les concepts issus de ces contextes nous renseignent ainsi :

- Groupe 1 : Baetis(2), Elmis(2), Esolus(2), Lepidostoma(6), Limnephilinae(3), Taeniopteryx(9)
 - traits biologiques : ils sont monovoltins⁽³³⁾, se caractérisent par un cycle de vie inférieur à un an, une reproduction cimentée, aucune forme de résistance, une respiration par branchies, un déplacement rampant, une alimentation de plantes mortes de plus d'un millimètre ou de microphytes vivants et sont des broyeurs.
 - traits écologiques : ils vivent dans les chenaux, les rives ou bras parallèles connectés de la rivière, préfèrent les eaux non saumâtres eurythermes⁽¹⁶⁾.
- Ce groupe est très hétérogène en terme de polluo-sensibilité et est peu informatif.

- Groupe 2 : Coenagrionidae(-), Erpobdellidae(1), Glossiphoniidae(1), Nepinae(-)
 - traits biologiques : ils sont monovoltins, se caractérisent par une taille entre 1 et 4 cm, une respiration par tegument⁽⁶³⁾, un déplacement rampant, une alimentation de vertébrés ou d'invertébrés et sont des perceurs ou des prédateurs.
 - traits écologiques : ils vivent hors de l'hydrosystème fluvial, en plaine dans des eaux non saumâtres, mésotrophes⁽³¹⁾, sans courant et eurythermes, principalement au milieu de macrophytes.

Ce groupe comporte des individus qui se distinguent des autres groupes par une taille assez importante leur permettant de chasser, et le fait qu'ils vivent au milieu de macrophytes dans des zones lenticues. Ils sont par ailleurs très polluo-résistants.

- Groupe 3 : Caenis(2), Calopteryx(-), Ceratopogonidae(-), Neureclipsis(4), Psychodidae(-), Valvata(2)
 - traits biologiques : ils ont une durée de vie inférieure à un an, n'ont aucune forme de résistance, un déplacement rampant.
 - traits écologiques : ils vivent en plaine, dans des eaux non saumâtres eurythermes.

Ce groupe plutôt polluo-résistant ne témoigne pas d'éléments particulièrement intéressants pour nos travaux.

- Groupe 4 : Leptocerus(4), Micrasema(8), Mystacides(4), Silo(7)
 - traits biologiques : ils se caractérisent par une taille entre 0,5 et 1 cm, sont monovoltins avec un cycle d'une durée inférieure ou égale à un an, ont une reproduction par une ponte fixée ou cimentée, aucune forme de résistance, un déplacement rampant et ils sont des broyeurs ou des racleurs.
 - traits écologiques : ils vivent en plaine, dans des eaux non saumâtres eurythermes β -mésosaprobies⁽³²⁾.

Ces taxons sont plutôt polluo-sensibles avec cependant une large amplitude de polluo-sensibilité (de 4 à 8). Il est intéressant de noter que ce groupe est plutôt mésosaprobe, autrement dit assez résistant à la matière organique. Ainsi, la polluo-sensibilité (qui concerne tous les types de pollution) est nuancée sur la pollution spécifique par les matières organiques.

- Groupe 5 : Habrophlebia(7), Heptageni(5), Leuctra(7), Nemoura(6), Potaman-

thus(5)

- traits biologiques : ils mesurent entre 0,5 et 2 cm, sont monovoltins avec un cycle d'une durée inférieure ou égale à un an, ils pondent des œufs isolés et cimentés, n'ont pas de forme de résistance, ont une respiration par tégument⁽⁶³⁾, un déplacement rampant et sont des broyeurs.
- traits écologiques : ils vivent dans les chenaux, les rives ou bras parallèles connectés de la rivière dans la zone supérieure (épipotamon) des rivières de plaine de courant faible, dans des eaux non saumâtres eurythermes.

Ce groupe se comporte comme le groupe 4 en étant un peu plus polluo-sensible (de 5 à 7) et sans mention particulière concernant la saprobie. Ils vivent dans l'épipotamon.

- Groupe 6 : Electrogena(5), Goera(7), Heptagenia(5), Leuctra(7), Paraleptophlebia(7), Rhithrogena(5)

- traits biologiques : ils mesurent entre 1 et 2 cm, sont monovoltins avec un cycle d'une durée inférieure ou égale à un an, pondent des œufs isolés et cimentés, n'ont pas de forme de résistance, une respiration par tégument⁽⁶³⁾, un déplacement rampant et sont des racleurs.
- traits écologiques : ils vivent surtout dans les chenaux de rivières, en plaine, entre les pierres, galets et cailloux, ils vivent dans des eaux non saumâtres eurythermes. Ils sont oligosaprobies⁽³⁵⁾.

Ce groupe est autant polluo-sensible que le groupe 5, mais il est également peu résistant aux matières organiques. De plus, de part leur préférence au substrat, on peut supposer qu'ils vivent dans des milieux non colmatés.

- Groupe 7 : Ecdyonurus(5), Lepidostoma(6), Nemoura(6), Nemurella(6), Sericosoma(6)

- traits biologiques : ils mesurent entre 0,5 et 2 cm, ont une reproduction par ponte cimentée ou fixée, n'ont pas de forme de résistance, ont une respiration par tégument⁽⁶³⁾, un déplacement rampant et ils sont des broyeurs.
- traits écologiques : ils vivent dans les chenaux de rivières ou les lacs, dans des eaux non saumâtres, eurythermes, oligotrophes, d'un pH > 5.

Ce groupe est également polluo-sensible et caractérise plutôt les milieux oligotrophes et les taxons qu'il contient ne vivent pas dans les milieux très acides (< 5).

On constate que certaines caractéristiques se retrouvent à travers tous ces groupes, telles que le fait que ces taxons vivent plutôt en plaine dans des eaux non saumâtres et

eurhythmes. Cela s'explique par le fait que nous n'avons conservé dans notre échantillon que les taxons de la plaine d'Alsace. Il est donc cohérent que ces traits émergent puisque ce sont des caractéristiques de plaine.

Cependant nous trouvons plusieurs niveaux de saprobie et de polluo-sensibilité qui sont très intéressants car ils permettent de distinguer les différentes formes de polluo-sensibilité.

5.2.3 Groupes d'invertébrés selon la classification de P. Usseglio-Polatera et al.

Dans cette partie nous présentons deux approches basées sur la classification par les grands genres de la taxonomie effectuée par P. Usseglio-Polatera et al. [67, 66]. Dans l'article [66], les auteurs établissent deux classifications distinctes. La première basée sur les traits biologiques et la seconde sur les traits écologiques. Nous comparons donc notre précédente approche à leurs conclusions, dans une première partie.

L'article [67] présente une classification établie sur tous les traits de vie. Nous choisissons de constituer des groupes tels qu'ils sont présentés par la classification et de générer des treillis sur ces groupes afin d'en extraire les taxons et traits les plus caractéristiques, dans une seconde partie.

Des concepts vers les niveaux systématiques supérieurs (classes, ordres...)

Nous analysons les groupes formés à partir des treillis dans la partie précédente en fonction de leurs niveaux systématiques supérieurs au lieu du degré taxonomique utilisé dans le contexte original et nous les comparons aux analyses effectuées par P. Usseglio-Polatera et al. [66]. Nous faisons référence au nom des groupes constitués par les auteurs (nommés par une lettre), pour le lecteur qui voudrait approfondir cette comparaison.

– Contexte des traits écologiques :

- Groupe 1 : Baetis, Calopteryx, Elmis, Esolus, Galba, Gammarus, Lepidostoma, Leuctra, Limnephilinae, Onychogomphus, Paraleptophlebia, Polycentropus, Sericostoma, Silo, Taeniopteryx

Niveaux systématiques supérieurs : Ephemeroptera, Odonata, Coleoptera, Gasteropoda, Crustacea, Trichoptera, Plecoptera

Le seul groupe issu de l'analyse des traits écologiques de [66] comportant ces 7 niveaux systématiques supérieurs est le groupe C.

- Groupe 2 : Asellidae, Caenis, Coenagrionidae, Corixidae, Cynus, Erpobdellidae, Glossiphoniidae, Hydroporinae, Leptocerus, Mystacides, Nepinae, Planorbidae, Tabanidae, Veliinae

Niveaux systématiques supérieurs : Crustacea, Ephemeroptera, Odonata, Heteroptera, Trichoptera, Hirudinea, Coleoptera, Gasteropoda, Diptera

Le seul groupe issu de l'analyse des traits écologiques de [66] comportant ces 9 niveaux systématiques supérieurs est le groupe D.

- Groupe 3 : Bithynia, Caenis, Calopteryx, Ceratopogonidae, Ephemerella, Goera, Neureclipsis, Psychodidae, Sphaerium, Valvata

Niveaux systématiques supérieurs : Gasteropoda, Ephemeroptera, Odonata, Diptera, Trichoptera, Hirudinea, Bivalvia

Le seul groupe issu de l'analyse des traits écologiques de [66] comportant ces 7 niveaux systématiques supérieurs est également le groupe D.

- Contexte des traits biologiques :

- Groupe 4 : Acentrella, Drusinae, Ecdyonurus, Leptocerus, Micrasema, Mystacides, Planorbidae, Rhithrogena, Riolus, Sericostoma, Setodes, Silo

Niveaux systématiques supérieurs : Ephemeroptera, Trichoptera, Gasteropoda, Coleoptera

Le seul groupe issu de l'analyse des traits biologiques de [66] comportant ces 4 niveaux systématiques supérieurs est le groupe F.

- Groupe 5 : Goera, Habrophlebia, Heptagenia, Leuctra, Nemoura, Nemurella, Paraleptophlebia, Potamanthus, Protonemura, Taeniopteryx

Niveaux systématiques supérieurs : Trichoptera, Ephemeroptera, Plecoptera, Odonata

Le seul groupe issu de l'analyse des traits biologiques de [66] comportant ces 4 niveaux systématiques supérieurs est le groupe C.

- Groupe 6 : Electrogena, Goera, Heptagenia, Leuctra, Paraleptophlebia, Rhithrogena, Torleya

Niveaux systématiques supérieurs : Ephemeroptera, Trichoptera, Plecoptera

Dans ce cas, 3 groupes issus de l'analyse des traits biologiques de [66] comportent ces 3 niveaux systématiques supérieurs : les groupes C, E et F.

- Groupe 7 : Ecdyonurus, Lepidostoma, Leptocerus, Mystacides, Nemoura, Nemurella, Sericostoma, Setodes

Niveaux systématiques supérieurs : Ephemeroptera, Trichoptera, Plecoptera

Dans ce cas, 3 groupes issus de l'analyse des traits biologiques de [66] comportent ces 3 niveaux systématiques supérieurs : les groupes C, E et F.

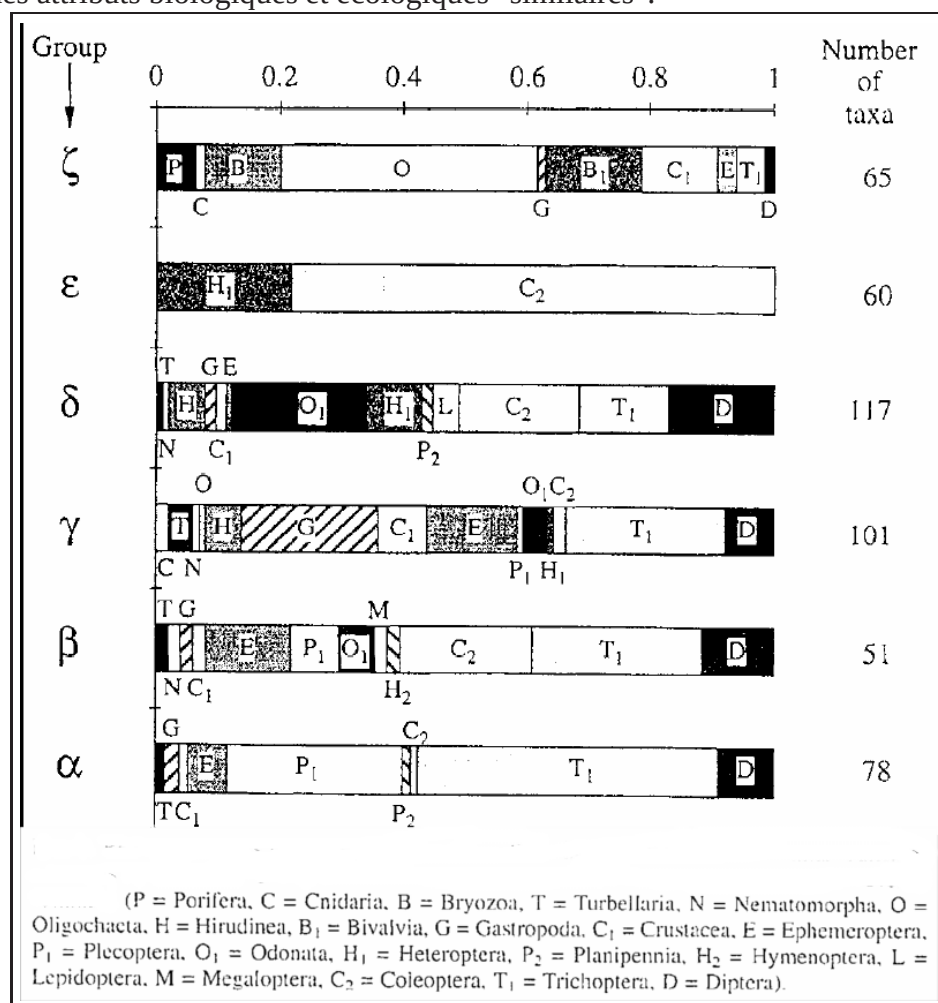
Les rapprochements entre les résultats des treillis et ceux établis dans l'article [66] ont été faits, mais leur analyse est encore à approfondir. Nous avons cependant pu remarquer que l'analyse issue des traits biologiques est plus riche et plus en adéquation avec les résultats de l'article [66] que celle menée à partir de traits écologiques. Malgré tout, nous gardons à l'esprit que les taxons que nous prenons en compte sont ceux de la plaine d'Alsace uniquement, et par ailleurs, les groupes sont établis sur leurs traits biologiques *ou* écologiques mais les treillis de chacun de ces groupes sont établis sur la totalité des traits de vie : biologiques *et* écologiques, ce qui peut donc perturber la

constitution des concepts, et faire que les taxons ne se retrouvent plus associés.

Des niveaux systématiques supérieurs vers les concepts

Dans son article [67], P. Usseglio-Polatera présente 6 groupes de taxons partageant des traits de vie similaires. Nous avons repris son schéma présentant ces groupes dans la figure 5.1. Pour chacun de ces groupes est représenté le pourcentage de chaque niveau systématique supérieur qui le compose, et le nombre de taxons qui constitue ce groupe.

FIGURE 5.1 – Apports relatifs de 21 unités taxonomiques à 6 groupes de taxons possédant des attributs biologiques et écologiques "similaires".



Dans les données que nous considérons, i.e. limitées aux taxons de la plaine d'Alsace, nous ne rencontrons pas les taxons des niveaux systématiques supérieurs Nematomorpha, Turbellaria, Megaloptera, Hymenoptera, Porifera et Bryozoa. Nous avons re-

constitué ces groupes à partir des taxons alsaciens et avons généré les concepts correspondants. Un seuillage a été nécessaire pour calculer ces concepts étant donné l'importance des contextes établis. Pour chacun de ces groupes, nous renseignons, ci-après, le nombre de taxons, le nombre d'histogrammes, le quantile du seuil et le nombre de concepts générés. Nous indiquons les traits de vie qui caractérisent les taxons qui émergent de chacun de ces groupes et les conclusions des auteurs de [67] concernant chacun de ces groupes originels. Nous indiquons en gras les points d'accord entre les traits venant des treillis et ceux venant de l'article, en italique les divergences. Les éléments ne figurant ni en gras, ni en italique sont ceux qui ne sont renseignés que dans l'une des deux approches.

- Groupe α : Elmis Esolus Limnius Macronychus Oulimnius Psychomyia Goera Philopotamus Silo
 - Données numériques : 94 taxons. 904 histogrammes. Quantile = 1/7. 29696 concepts.
 - Traits biologiques : **ces taxons mesurent moins de 2 cm, se reproduisent par des œufs isolés ou cimentés ou des pontes fixées ou cimentées, ont une forme de résistance nulle si elle n'est pas par diapause⁽⁸⁾ ou quiescence⁽⁴⁶⁾, respirent essentiellement par téguments et branchies, se déplacent en rampant ou par fixation temporaire et se nourrissent surtout de microphytes vivants en raclant.**
 - Traits écologiques : ils vivent dans les chenaux, les rives ou bras parallèle connectés de la rivière, *en plaine*, dans des courants plutôt forts d'eaux non saumâtres *eurythermes⁽¹⁶⁾*, **oligosaprobes⁽³⁵⁾** à β -*mésosaprobes⁽³²⁾* et d'un pH > 5,5.
Analyse du groupe originel : **de taille moyenne**, les taxons de ce groupe sont mono ou semivoltins⁽⁵³⁾, **rampants**, typiquement rhéophiles, *vivant dans des eaux froides des parties hautes des rivières de montagne (et alpines)* **xéno ou oligosaprobes**. Ce sont des individus qui vivent dans les sédiments minéraux grossiers, en milieux oligotrophes. **La respiration est aquatique. Ils ont des œufs isolés ou des pontes cimentées. Les œufs et les larves peuvent entrer en diapause⁽⁸⁾ et quiescence⁽⁴⁶⁾.** Les taxons sont essentiellement des racleurs et *broyeurs*, **se nourrissant principalement de microphytes**, de débris végétaux ou macro-invertébrés.
- Groupe β : Elmis Esolus Limnius Oulimnius Paraleptophlebia Lepidostoma Silo
 - Données numériques : 103 taxons. 951 histogrammes. Quantile = 1/6. 4051 concepts.
 - Traits biologiques : ces taxons mesurent moins de 2 cm, ont un cycle de vie de moins d'un an, **se reproduisent par des œufs isolés ou cimentés ou des pontes fixées ou cimentées**, ils n'ont aucune forme de résistance, **rampent**,

et se nourrissent de *débris végétaux* > 1 mm et de **microphytes vivants, par broyage ou raclage**.

- Traits écologiques : **ils vivent surtout dans les chenaux, les rives** ou *bras parallèles connectés* de la rivière (voire les lacs), **en plaine**, dans des courants faibles ou intermédiaires d'eaux non saumâtres **eurythermes, oligosaprobés à β –mésosaprobés** et d'un pH > 5,5.

Analyse du groupe originel : les taxons sont très représentatifs du rhithron⁽⁴⁹⁾, vivant dans un environnement oligo à mésotrophe, **dans les chenaux et rives de rivière**, en plaine et piémont. Ces individus sont de taille petite à moyenne, et vivent en général dans des milieux **oligo à β –mésosaprobe**, et **eurythermes**. Leur reproduction, leurs habitudes alimentaires et leur relation avec le substrat sont similaires au groupe α . Cependant ce groupe inclut un peu moins d'organismes rhéophiles⁽⁴⁸⁾, et si leur respiration reste aquatique, les branchies sont plus souvent utilisées que dans le groupe α .

- Groupe γ : Elmis Esolus Hydraena Hydrocyphon Limnius Macronychus Oulimnius
 - Données numériques : 113 taxons. 1046 histogrammes. Quantile = 1/6. 7356 concepts.
 - Traits biologiques : *ces taxons ne mesurent pas plus de 0,5 cm*, ils ont un cycle de vie qui dure plus d'un an, **une reproduction par pontes cimentées, fixées, endophytiques ou terrestres**. Ils n'ont aucune forme de résistance, **rampent, et se nourrissent de micro-organismes, débris < 1 mm et surtout de microphytes vivants par broyage ou raclage**.
 - Traits écologiques : ils vivent dans *les chenaux, les rives ou bras parallèles connectés de la rivière*, **en plaine**, dans des eaux non saumâtres **eurythermes, oligotrophes à mésotrophes, plutôt β –mésosaprobés** et d'un pH > 6.
Analyse du groupe originel : *ils sont de taille moyenne, vivent dans les étangs, mares et méandres abandonnés de rivières* de plaine pouvant être en rhithron ou potamon⁽⁴²⁾, **eurythermes, β –mésosaprobés**. Ils s'établissent dans des habitats variés (dalles, blocs, pierres, galets et macrophytes). Tous les autres traits biologiques sont similaires à ceux du groupe β .
- Groupe δ : Elmis Esolus Limnius Macronychus Oulimnius Caenis Lepidostoma Silo Valvata
 - Données numériques : 107 taxons. 1011 histogrammes. Quantile = 1/6. 5788 concepts.
 - Traits biologiques : ces taxons mesurent moins d'un centimètre, **se reproduisent par des œufs isolés ou cimentés ou des pontes fixées ou cimentées, n'ont quasiment aucune forme de résistance et rampent**.
 - Traits écologiques : *ils vivent dans les chenaux, les rives ou bras parallèles*

connectés de la rivière, en plaine, dans des eaux non saumâtres **eurythermes**, oligosaprobies à β -**mésosaprobies** et d'un pH > 5,5.

Analyse du groupe originel : ce groupe vit dans un environnement **eurytherme**, β -**mésosaprobe**. Les individus **rampent** et *nagent* et peuvent coloniser une vaste variété de substrats mais préfèrent les substrats organiques. Ils vivent dans des *étangs, mares, méandres abandonnés, marais, tourbières et lacs de plaine*. Les genres sont polyvoltins⁽⁴³⁾ et **utilisent un large panel de techniques de reproduction** et de tactiques de dispersion. La *diapause larvaire peut être utilisée comme un mode de résistance*. La respiration est à la fois aquatique (téguments ou branchies) et aérienne (stigmates). En tant que larves, ils sont surtout des perceurs et broyeurs, et ils se nourrissent essentiellement d'invertébrés.

- Groupe ϵ : Elmis Esolus Limnius Macronychus Orectochilus Oulimnius Riolus Aphelocheiridae
 - Données numériques : 25 taxons. 255 histogrammes. Quantile = 1/10. 4887 concepts.
 - Traits biologiques : ces taxons sont tout petits (généralement moins de 0,5 cm), ils n'ont qu'un seul cycle de vie durant plus d'un an, **se reproduisent par des œufs isolés ou cimentés ou des pontes fixées ou cimentées**, une forme de résistance nulle si elle n'est pas par diapause ou quiescence, ils **rampent et nagent**, se nourrissent de microphytes vivants ou de vertébrés et macroinvertébrés vivants.
 - Traits écologiques : ils vivent dans les chenaux de rivière, en plaine, dans des eaux non saumâtres, **eurythermes**, oligosaprobies à β -**mésosaprobies**, d'un pH > 6, et *ne supportent pas les milieux eutrophes*.
- Analyse du groupe originel : Ils vivent dans tout type d'habitat excepté dans la partie la plus rapide du chenal principal. **Ils sont surtout des nageurs et rampeurs**, vivant dans des habitats lenticques ou stagnants et *eutrophes* (spécialement dans les macrophytes, litières et vases). Ils vivent dans des eaux **eurythermes** α ou β -**mésosaprobies**. **Les pontes sont cimentées sur le substrat minéral, ou les œufs isolés** sont déposés sur ou dans les macrophytes. La dispersion et la respiration sont essentiellement aériennes.
- Groupe ζ : Baetis Ecdyonurus Ephemerella Heptagenia Paraleptophlebia Goera Lepidostoma Silo Leptocerus Mystacides.
 - Données numériques : 75 taxons. 796 histogrammes. Quantile = 1/7. 6981 concepts.
 - Traits biologiques : ces taxons partagent *une taille entre 0,5 et 2 cm*, un cycle de vie de moins d'un an, une reproduction par des œufs isolés ou cimentés ou des pontes fixées ou cimentées, *aucune forme de résistance* et **une respiration**

par téguments et branchies.

- Traits écologiques : ils vivent en plaine dans des eaux non saumâtres plutôt mésotrophes, eurythermes, plutôt β –mésosaprophes et d'un pH > 5,5.

Analyse du groupe originel : les organismes de ce groupe sont uni ou plurivoltins et utilisent les œufs ou les cocons comme mode de résistance. Ils se dispersent par dérive et se reproduisent par ovoviviparité ou par reproduction asexuée. **Ils ont une respiration branchiale ou tégumentaire.** Ils sont filtreurs ou mangeurs de sédiments fins et se nourrissent de microphytes ou de fins débris. Dans ce groupe il y a une *importante variation de taille* et de préférence de substrat.

Cette analyse reste à interpréter. On peut noter cependant que même s'il y a un bon nombre de points communs, une grande partie des traits ne trouve pas son pendant d'une analyse à l'autre. Nous attribuons ce fait à plusieurs possibilités. D'une part, nous ne considérons que les taxons de la plaine d'Alsace. D'autre part, la reconstitution des groupes prend en compte tous les taxons correspondant aux niveaux systématiques supérieurs listés dans l'article [67], alors que dans cet article, un taxon Diptera présent dans le groupe α ne sera pas présent dans le groupe β par exemple. Par ailleurs, les traits dans l'article sont décrits grâce à la connaissance de l'expert (comme le fait de distinguer un comportement à l'état larvaire de l'état d'œuf, distinction qui n'apparaît pas dans les traits et modalités du contexte et donc que les treillis (sans interprétation) ne peuvent faire ressortir.

5.3 Conclusion

Dans cette partie nous avons appliqué l'algorithme MinMaxC au jeu de données des macrophytes, ceci nous permettant d'associer des traits écologiques aux traits biologiques par le regroupement d'espèces suggéré par les treillis et effectué avec l'aide de Michèle Trémolières.

Avec Corinne Grac, nous avons également appliqué cet algorithme aux données des invertébrés qui représentent un contexte global 16 fois plus important que celui des macrophytes et dont la constitution des groupes à partir des treillis est peu aisée. Par ailleurs, des classifications existent, nous les avons donc suivies pour établir nos groupes. Les groupes biotypologiques de Verneaux ne nous ont pas permis d'aboutir à une association entre les traits écologiques et les traits biologiques. Une classification par les treillis a donc été menée. Nous avons dû scinder la constitution des groupes en deux pour alléger l'interprétation : en regroupant les taxons en fonction de leurs traits biologiques d'une part et de leurs traits écologiques d'autres part. Ces groupes constitués ont permis d'établir de nouveaux contextes plus réduits pour lesquels tous les traits de vie pouvaient être pris en compte. Une analyse a été établie sur ces groupes afin de

faire émerger les taxons et les traits de vie les plus représentatifs de chaque groupe. Par ailleurs, Usseglio-Polatera et al. ont publié deux articles [66, 67] établissant et interprétant des groupes avec les niveaux systématiques supérieurs des taxons à partir des traits biologiques, puis des traits écologiques et enfin de tous les traits de vie. Nous avons donc comparé les résultats issus de nos groupes et de ceux des articles. La comparaison établit des points de similitude qui confortent notre analyse, mais également des divergences dont la source peut provenir de plusieurs éléments. Tout d'abord, notre jeu de données se limite aux taxons échantillonnés dans la plaine d'Alsace. Ainsi les traits écologiques qui indiquent que les taxons vivent en plaine dans des eaux non saumâtres eurythermes sont parfaitement cohérents et peuvent différer de l'analyse des articles qui regroupent des taxons vivant dans des milieux beaucoup plus divers. Par ailleurs, par les treillis, nous relevons les traits communs les plus représentatifs du groupe de taxons. Ainsi, si des taxons respirent essentiellement par les téguments, mais en parallèle utilisent les branchies selon un pourcentage bien plus faible, nous n'indiqueront que les téguments, quand l'article évoquera une "respiration aquatique" qui est à la fois par les téguments et par les branchies. De plus les auteurs sont des experts reconnus qui connaissent les particularités des différents taxons et peuvent ainsi fournir des distinctions que les taxons peuvent avoir pour différents traits en fonction de leur état, comme de distinguer la forme de résistance à l'état d'œuf ou larvaire. Ces informations n'étant pas rendues dans les contextes, les treillis sont donc incapables de les mettre en évidence. Cependant ils sont très efficaces pour faire ressortir les relations entre les taxons, les traits biologiques et écologiques renseignés dans le contexte.

La dernière approche suivie constitue les groupes en partant de la classification établie dans l'article [67]. Cette partie est en cours d'analyse et d'interprétation, mais nous mettons les premiers éléments afin qu'ils puissent être poursuivis ultérieurement.

Chapitre 6

Conclusion et perspectives

Sommaire

6.1	Traitement binaire	135
6.2	Traitement flou	137
6.3	Application	138
6.4	Discussion et perspectives	139

Comme nous l'avons vu, l'objet principal de cette thèse est de développer une approche pour la fouille de données complexes et porte plus spécifiquement sur l'analyse de concepts formels.

Nous sommes intéressés par l'étude de données particulières représentées par des *contextes multi-valués flous* qui sont des contextes multi-valués (dont les attributs sont divisés en modalités) pour lesquels la relation entre objets et modalités est floue, c'est à dire qu'un objet possède une modalité selon un certain pourcentage appelé *affinité*.

Nous présentons deux démarches pour traiter ces contextes. La première consiste à adapter les données aux traitements classiques par une binarisation. La seconde consiste à adapter les méthodes aux données.

Enfin nous appliquons les méthodes au domaine hydrobiologique pour la sélection d'attributs écologiques et leur mise en correspondance avec des taxons et leurs attributs biologiques.

6.1 Traitement binaire

Nous présentons deux approches de binarisation.

La première est une disjonction totale par un *échelonnage nominal*, nous fournissant un *contexte nominal* (dit "Lxy"). Ce contexte binaire permet alors d'utiliser des *règles d'implications*. Leur intérêt porte, d'une part, sur les relations qu'elles mettent en ex-

ergue entre les attributs, à travers les implications fréquentes et à fort support, mais également par les implications négatives ou de faible support. D'autre part, elles nous permettent d'ajuster le jeu de données, corriger les éventuelles erreurs, ajuster la plage des affinités, limiter la dispersion d'information au sein des modalités et les redondances d'informations entre les attributs et les modalités et ôter les attributs non pertinents. Par ailleurs, pour nous rapporter à des approches plus habituelles en biologie, nous utilisons des méthodes statistiques puisque le contexte est binaire et donc le permet. Nous comparons et combinons l'approche latticielle et l'approche statistique, notamment par l'Analyse Factorielle Multiple (AFM).

La comparaison permet d'établir que les treillis sont une approche pertinente pour traiter ce type de problématique car ses résultats se complètent avec ceux issus des statistiques.

La combinaison des deux techniques se fait par une première étape de pré-sélection des données par l'AFM sur lesquelles est ensuite construit le treillis. Cela permet d'évaluer les apports réciproques de chaque méthode de laquelle se dégage un double intérêt. D'une part les treillis permettent de préciser les résultats de l'AFM. D'autre part l'AFM permet une pré-segmentation des données qui permet de construire des treillis plus petits et mieux focalisés. Différents niveaux de complexité peuvent être explorés, en fonction du nombre d'individus/propriétés retenus sur les axes de l'AFM ou en utilisant les informations de plusieurs axes. Ceci permet alors de sélectionner les données les plus pertinentes à prendre en compte et ainsi d'alléger les résultats à interpréter, ce qui est utile lorsque les contextes sont importants, et que les treillis deviennent rapidement volumineux et illisibles donc inexploitable. Pour palier ce type de contrainte, des travaux ont été menés, notamment par l'utilisation de calcul de χ^2 [42].

La seconde méthode de binarisation est un échelonnage histogramme. Nous définissons formellement les notions d'*histogrammes* et d'*échelonnage histogramme*. Les histogrammes permettent de prendre en considération la répartition des affinités entre les différentes modalités de chaque attribut et qui forment ainsi la partie floue des données.

La comparaison des treillis nominal et histogramme pointe une lecture du treillis histogramme plus aisée, non seulement parce qu'il est moins volumineux, mais aussi parce qu'un concept histogramme regroupe l'équivalent de 3 ou 4 concepts L_{xy} . Pour un objet (ou groupe d'objets), un seul concept histogramme résume directement l'information sans avoir besoin de rechercher dans le treillis. L'inconvénient est qu'il faut une égalité des distributions d'affinités pour pouvoir construire les concepts.

L'application de cette thèse à la biologie, présente la difficulté de traiter des données réelles d'un domaine avec lequel il faut se familiariser et qui est très complexe. Cependant, il nous offre également l'avantage de travailler sur des données et d'avoir des experts présents pour interpréter nos résultats.

Par ailleurs, l'approche des treillis est novatrice dans ce domaine, où les biologistes

se réfèrent préférentiellement à des méthodes statistiques. Etablir une comparaison et une combinaison des deux méthodes a permis de convaincre les biologistes de l'intérêt de l'approche latticielle. Les implications sont aussi très appréciées par les utilisateurs car elles se lisent facilement. Les informations qu'elles font ressortir sont intéressantes, cependant c'est lorsqu'on les combine qu'on découvre des relations particulièrement digne d'intérêt, même si ce travail est rapidement complexe à cause des nombreuses combinaisons possibles et on peut facilement manquer certaines informations.

Concernant les conversions binaires, que l'on soit intéressé par le diagramme de Hasse ou uniquement par la liste des concepts, dans le cas de contextes multi-valués flous, l'échelonnage histogramme est à préférer. D'une part parce que les attributs histogrammes sont moins nombreux que les attributs Lxy, et que le nombre de concepts générés par des connexions de Galois binaires et donc leur temps de calcul, sont fonction du nombre d'attributs. D'autre part, car ils permettent de conserver l'information de distribution des valeurs entre les modalités d'un trait.

6.2 Traitement flou

La seconde approche pour le traitement des données multi-valuées floues consiste à adapter les treillis aux données et passe par la définition de nouvelles connexions de Galois que nous appelons *connexions min-max*. Ces connexions manipulent des histogrammes et permettent de mettre en évidence les propriétés minimales et maximales communes à un groupe d'objets. Ainsi il est possible de comparer non seulement des distributions d'affinités sur les modalités mais également des "formes" de distribution (représentées par les histogrammes) comme par exemple $B_{110} \leq B_{210} \leq B_{221}$, ce qui est particulièrement significatif pour des modalités ordonnées ou temporelles. Par ailleurs, les fonctions sur lesquelles les connexions min-max sont établies sont croissantes monotones, et donc le nombre de concepts générés et les temps de calculs augmentent très rapidement avec le nombre d'objets et d'attributs. Pour contrer cet inconvénient, nous introduisons une notion de seuillage dans la fermeture de Galois qui ne conserve que les concepts dont l'écart entre les propriétés minimales et maximales est inférieur au seuil. Nous évaluons diverses options de seuillage et la plus adaptée apparaît être l'option d'un seuillage par trait et par quantile. Cette option permet non seulement de réduire le temps de calcul et le nombre de concepts générés, mais également de générer les concepts les plus pertinents.

Ces connexions min-max ont été implémentées dans deux algorithmes : MinMaxNC et MinMaxC qui sont deux adaptations des algorithmes NextClosure et Charm. Les tests montrent que sans seuillage MinMaxNC est plus efficace que MinMaxC, cependant avec seuillage, MinMaxC est à préférer à MinMaxNC.

Les histogrammes peuvent avoir des sens très différents en fonction des attributs

et des modalités considérés. Ainsi avec des modalités qualitatives nous conseillons de conserver l'échelonnage histogramme actuel. En outre, pour des modalités quantitatives, même si on peut tout à fait conserver l'échelonnage histogramme proposé, il peut aussi être abordé différemment. Par exemple, considérons un attribut "distance" divisé en 3 modalités : courte ($< 1\text{km}$), moyenne ($< 2\text{km}$) et longue ($> 2\text{km}$). L'échelonnage considéré dans nos travaux estime qu'un individu parcourant 500 m parcourt une petite distance. Cependant, un autre échelonnage pourra considérer que $0,5 < 1$ mais également que $0,5 < 2$. Auquel cas, 500 m représentent une petite et une moyenne distance.

6.3 Application

Ces travaux ont été appliqués à l'hydrobiologie dans le cadre du projet *Indices*, mis en œuvre suite à la directive cadre européenne sur l'eau qui préconise l'évaluation de l'état écologique de toutes les masses d'eau d'ici 2015. Actuellement cette évaluation s'opère par des analyses physico-chimiques ou des prélèvements au sein de la faune et la flore des cours d'eau. La physico-chimie est précise mais ne conserve pas les traces des pressions antérieures aux prélèvements. Les paramètres biologiques quant à eux sont très compartimentés et difficilement comparables. De plus, les 5 indices normalisés qui en dépendent ne fonctionnent presque qu'en France, et cependant, pas de manière uniforme d'une hydro-éco-région à une autre. Ainsi un nouvel outil est nécessaire. Dans cette thèse, nous avons suivi l'idée de rechercher un outil basé sur les traits écologiques qui caractérisent le comportement des taxons au sein de leur environnement. Nous avons étudié les macrophytes et invertébrés de la plaine d'Alsace. Le chapitre 3 décrit une méthode de sélection des traits écologiques basée sur le traitement binaire du contexte histogramme des traits biologiques des macrophytes. En effet, il existe de nombreux traits écologiques mais nous recherchons les plus pertinents pour qualifier l'état écologique des cours d'eau en fonction des espèces et de leurs traits biologiques (morphologiques et physiologiques). Pour cela, les concepts sont analysés par un expert pour relever les relations entre les espèces des concepts sélectionnés et les traits écologiques (par rapport à leur environnement). Ensuite ces traits écologiques sont ajoutés au jeu de données et leurs affinités sont renseignées en accord avec la littérature et les connaissances de l'expert. Les concepts du treillis étendu sont ensuite examinés pour valider le processus et mettre en évidence les ensembles des traits biologiques et écologiques. Conformément à nos attentes, nous retrouvons les informations fournies par l'expert, et plus encore, le treillis nous fournit davantage d'informations et des informations plus précises.

Cette méthode a permis la sélection de 5 traits écologiques : stabilité de l'eau, tolérance à la sédimentation, tolérance aux variations d'humidité, tolérance aux matières organiques et tolérance à la trophie. Ces traits ont pu ensuite être associés à des groupes de plantes et de traits biologiques grâce aux connexions min-max. Ceci permet d'une

part de pouvoir déterminer un profil écologique en rencontrant une espèce ou une autre sur un cours d'eau, car toutes les espèces d'un même groupe partagent les mêmes traits écologiques. D'autre part, cette analyse est utilisable dans d'autres hydro-éco-régions car si les plantes de ces groupes n'y vivent pas, d'autres espèces partageant les mêmes attributs biologiques peuvent les remplacer.

Enfin, les algorithmes ont été éprouvés également sur des données concernant des invertébrés de la plaine d'Alsace qui forment un contexte 16 fois plus volumineux que celui des macrophytes. Grâce au seuillage, les algorithmes aboutissent et nous fournissent des concepts plus intéressants pour l'analyse car en nombre plus réduit et plus pertinents.

Lors de la seconde phase de l'analyse des données qui associe les traits écologiques aux traits biologiques, nous nous sommes aperçues que certains traits écologiques qui avaient été sélectionnés par l'approche binaire, sont rarement mis en évidence. En effet, certains traits sont importants et récurrents tels que la tolérance à la trophie et à la sédimentation mais d'autres ne caractérisent jamais ou rarement un concept tels que la stabilité de l'eau ou la tolérance à la matière organique pour lesquelles les plantes sont équitablement réparties et ne manifestent donc pas de préférence. Cela implique que ces traits sont informatifs à propos du milieu mais ne permettent pas de segmenter les espèces ou pointer un biais dans la méthode.

6.4 Discussion et perspectives

Afin de compléter les démarches entamées durant la thèse, diverses pistes sont à envisager tant du côté biologique qu'informatique.

D'un point de vue hydrobiologique, des analyses ont été débutées sur les invertébrés, cependant l'interprétation reste à terminer. Par ailleurs, on pourrait reprendre la méthode, qui part des treillis pour obtenir les concepts à analyser en fonction de leurs niveaux systématiques supérieurs, à partir du contexte de tous les traits de vie, au lieu des deux contextes séparés des traits biologiques et écologiques, pour extraire les concepts et de les comparer à l'article [67].

Il serait également intéressant de reprendre la dernière analyse sur l'ensemble de tous les taxons (pas seulement les taxons alsaciens) pour comparer l'approche et les résultats à ceux de l'article [67].

Les données qui ont été traitées sont les macrophytes servant de base à l'indice IBMR et les invertébrés servant pour l'IBGN. Dans le cadre du projet *Indices* qui vise à comparer les 5 indices normalisés, les analyses pourraient être appliquées aux autres compartiments biologiques : les oligochètes, diatomés et poissons.

Par ailleurs, il faudrait intégrer les paramètres physico-chimiques dans les concepts au même titre que les traits de vie (biologiques et écologiques). Les données physico-

chimiques existent et sont, comme pour les espèces, associées aux stations où elles sont mesurées. Des travaux statistiques par des méthodes RLQ sont en cours à ce propos, et une étude par une Analyse Relationnelle de Concepts (ARC) a été esquissée également pour commencer à relier la physico-chimie aux traits biologiques. Précisons rapidement ce qu'est l'ARC et comment elle peut concerner nos données. L'ARC se base sur l'analyse de concepts formels et permet de prendre en compte plusieurs contextes formels et de regrouper des objets partageant les mêmes attributs qui peuvent être relationnels.

Le modèle de données de l'ARC est une famille de contextes relationnels (FCR) [27, 43, 59] définie ainsi :

Définition 38 Famille de Contextes Relationnels. Une famille de contextes relationnels est un couple $(K; R)$. K est un ensemble de contextes formels $K_i = (O_i; A_i; I_i)$, R est un ensemble de contextes relationnels $R_j = (O_k; O_l; I_j)$ (O_k et O_l sont les ensembles d'objets des contextes K_k et K_l de K).

Dans notre cas, nous disposons des contextes $K_k = (\{\text{stations de prélèvement}\}; \{\text{paramètres physico-chimiques}\}; I_k)$, $K_l = (\{\text{taxons}\}; \{\text{traits de vie}\}; I_l)$, et du contexte relationnel $R_j = (\{\text{stations de prélèvement}\}; \{\text{taxons}\}; I_j)$.

De plus, l'intégration des connexions min-max dans l'approche de l'ARC pour le traitement des histogrammes est également une piste très intéressante tant du point de vue biologique qu'informatique.

Par ailleurs, les connexions min-max actuelles prennent en considération un nombre d'histogrammes fixe et identique pour chaque concept : un par attribut. Il serait pertinent de redéfinir la notion d'ordre de telle sorte que les intensions des concepts s'établissent sur un nombre variable d'attributs. Les concepts seraient alors porteurs d'un autre sens qui peut intéresser les biologistes et ils permettrait également d'alléger certaines analyses. En effet, nous générons des listes de concepts dans lesquelles nous recherchons quels sont les concepts dont les attributs et modalités sont les plus marqués et mis en avant. Dans le cas d'un nombre d'attributs variable, ces concepts émergeraient d'eux mêmes sans être perturbés par la présence des autres attributs, et faciliteraient ainsi leur recherche, parmi un nombre de concepts généré qui peut être rapidement faramineux.

Les travaux menés sur les connexions min-max et les histogrammes correspondent à une problématique de manipulation de contextes complexes à laquelle de nombreux travaux cherchent à répondre, tel qu'exposé dans la partie 2.3. Certains de ces travaux, comme la manipulation d'intervalles [47] pourraient être exploités dans le cadre de notre problématique biologique. En effet, on peut considérer le seuillage des connexions min-max comme un intervalle entre deux histogrammes min et max.

Bibliographie

- [1] Qualité de l'eau : détermination de l'indice biologique global normalisé (ibgn). AFNOR NF T90-350, 1992.
- [2] *Invertébrés d'eau douce : systématique, biologie, écologie*. 2000.
- [3] Qualité de l'eau : détermination de l'indice biologique diatomées (ibd). AFNOR NF T90-354, 2000.
- [4] Qualité de l'eau : détermination de l'indice oligochètes de bioindication des sédiments (iobs). AFNOR NF T90-390, 2002.
- [5] Qualité de l'eau : détermination de l'indice biologique macrophytique en rivière (ibmr). AFNOR NF T90-395, 2003.
- [6] Qualité de l'eau : détermination de l'indice poissons rivière (ipr). AFNOR NF T90-344, 2004.
- [7] V.J. ABERNETHY : Functional ecology of euhydrophyte communities of european riverine wetland ecosystems. Ph.D. Thesis, University of Glasgow, Glasgow, 1994.
- [8] R. AGRAWAL, H. MANNILA, R. SRIKANT, H. TOIVONEN et A. I. VERKAMO : Fast discovery of association rules. *In Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press, 1996.
- [9] M. BARBUT et B. MONJARDET : *Ordre et classification – Algèbre et combinatoire*. Hachette, Paris, 1970.
- [10] M.-F. BAZERQUES : Directive-cadre sur l'eau : le bon état écologique des eaux douces de surface : sa définition, son évaluation. Communication au Ministère de l'Écologie et du Développement Durable, Paris, 2004.
- [11] A. BERTAUX, A. BRAUD et F. LE BER : Mining complex hydrobiological data with galois lattices. *In DEXA '07 : Proceedings of the 18th International Conference on Database and Expert Systems Applications*, pages 519–523, Washington, DC, USA, 2007. IEEE Computer Society.
- [12] A. BERTAUX, F. LE BER, A. BRAUD et M. TRÉMOLIÈRES : Mining complex hydrobiological data with galois lattices. *International Journal of Computing and Information Sciences*, 2009. à paraître.

- [13] G. BIRKHOFF : *Lattice Theory*, volume XXV de *Colloquium Publications*. American Mathematical Society, New York, 1948.
- [14] G. BORNETTE, C. HENRY, M. H. BARRAT et C. AMOROS : Theoretical habitat templates, species traits and species richness : aquatic macrophytes in the upper rhône river and its floodplain. *In Freshwater Biology* 31, pages 487–505, 1994.
- [15] R. BĚLOHLAVEK : Fuzzy galois connections. *Math. Logic Quaterly*, 1999.
- [16] R. BĚLOHLAVEK : *Introduction to formal concept analysis*, volume XXV. department of computer science, faculty of science, Palack'y University, Olomouc, 2008.
- [17] R. BĚLOHLAVEK et V. VYCHODIL : What is a fuzzy concept lattice ? *In 3rd Int. Conference on Concept Lattices and Their Applications (CLA 2005)*, pages 34–45, 2005.
- [18] C. CARPINETO et G. ROMANO : A Lattice Conceptual Clustering System and Its Application to Browsing Retrieval. *ml*, 24(2):95–122, 1996.
- [19] P. CLARK et R. BOSWELL : Rule induction with CN2 : some recents improvements. *In ECML'91*, pages 151–163, 1991.
- [20] J.-F. CONDOTTA : Raisonnement sur l'espace et le temps. IGAT, chapitre 7, pages 181–223. Lavoisier, Hermes, 2007.
- [21] Le conseil des communautés EUROPÉENNES : Directive 92/43/cee du conseil, du 21 mai 1992, concernant la conservation des habitats naturels ainsi que de la faune et de la flore sauvages., mai 1992.
- [22] C.D.K. COOK : *Aquatic plant book*. SPB Academic Publishing, the Hague, 1990.
- [23] Secrétariat d'Administration National des Données Relatives sur L'EAU : Processus d'acquisition des données biologiques. Document de travail V2, 2008.
- [24] B. A. DAVEY et H. A. PRIESTLEY : *Introduction to Lattices and Order*. Cambridge University Press, Cambridge, UK, 1990.
- [25] Ministère de l'écologie et du développement DURABLE : Circulaire dce 2007/22, du 11 avril 2007, relative au protocole de prélèvement et de traitement des échantillons des invertébrés pour la mise en œuvre du programme de surveillance sur cours d'eau., avril 2007.
- [26] E. DIDAY : Introduction à l'approche symbolique en analyse de données. *In Actes des journées "Symboliques-Numériques" pour l'apprentissage de connaissances à partir d'observation*, pages 21–56. Diday, E. and Kodratoff, Y. eds, 1987.
- [27] X. DOLQUES, M. HUCHARD et C. NEBUT : Génération de transformation de modèles par application de l'ARC sur des exemples. *In LMO'09 : Langages et Modèles à Objets*, pages 61–75, 2009.

- [28] V. DUQUENNE : Contextual implications between attributes and some representational properties for finite lattices. In *Beiträge zur Begriffsanalyse*, pages 213–239. B.I. Wissenschaftsverlag, Mannheim, 1987.
- [29] V. DUQUENNE et J.-L. GUIGUES : Famille minimale d'implications informatives résultant d'un tableau de données binaires. *Math. et Sci. Hum.*, 24(95):5–18, 1986.
- [30] S. DÍAZ et M. CABIDO : Plant functional types and ecosystem function in relation to global change. In *J. Veg. Sci.*, volume 8, pages 463–474, 1997.
- [31] B. ESCOPIER et J. PAGÈS : *Analyses factorielles simples et multiples*. Dunod, Paris, 1988.
- [32] A. H. FITTER et H. J. PEAT : The ecological flora database. In *Journal of Ecology*, pages 415–425, 1994.
- [33] H. FU et E. MEPHU NGUIFO : A lattice algorithm for data mining, 2008.
- [34] B. GANTER : Two basic algorithms in concept analysis. In *Technische Hochschule Darmstadt*, page 28 p. Preprint, 1984.
- [35] B. GANTER et S. KUZNETSOV : Pattern structures and their projections. In *Proceedings of the 9th International Conference on Conceptual Structures (ICCS 2001)*, LNCS 2120, pages 129–142. Springer, 2001.
- [36] B. GANTER et R. WILLE : Applied lattice theory : Formal concept analysis. In *General Lattice Theory*, G. Grätzer editor, Birkhäuser. Preprints, 1997.
- [37] B. GANTER et R. WILLE : *Formal Concept Analysis : Mathematical Foundations*. Springer Verlag, 1999.
- [38] C. GRAC, F. LE BER, A. BRAUD, A. HANDJA, A. HERMANN, N. LACHICHE et M. TRÉMOLIÈRES : Mining a database on alsatian rivers. In *Proceedings of the seventh International Conference on Hydroinformatics (HIC 2006)*, 2006.
- [39] J. P. GRIME, J. G. HODGSON et R. HUNT : Comparative plant ecology. Unwin Hyman, London, 1988.
- [40] J.P. GRIME, K. THOMPSON, R. HUNT, J.G. HODGSON, J.H.C. CORNELISSEN, I.H. RORISON, G.A.F. HENDRY, T.W. ASHENDEN, A.P. ASKEW, S.R. BAND, R.E. BOOTH, C.C. BOSSARD, B.D. CAMPBELL, J.E.L. COOPER, A.W. DAVISON, P.L. GUPTA, W. HALL, D.W. HAND, M.A. HANNAH, S.H. HILLIER, D.J. HODKINSON, A. JALILI, Z. LIU, J.M.L. MACKEY, N. MATTHEWS, M.A. MOWFORTH, A.M. NEAL, R.J. READER, K. REILING, W. ROSSFRASER, R.E. SPENCER, F. SUTTON, D.E. TASKER, P.C. THORPE et J. WHITEHOUSE : Integrated screening validates primary axes of specialisation in plants. In *Oikos*, volume 79, pages 259–281, 1997.
- [41] A. GUÉNOCHE : Construction du treillis de galois d'une relation binaire. *Mathématiques, Informatique et Sciences Humaines*, 109:41–53, 1990.

- [42] J. HERETH, G. STUMME, R. WILLE et U. PRISS : Conceptual knowledge discovery and data analysis. In *ICCS '00 : Proceedings of the Linguistic on Conceptual Structures*, pages 421–437, London, UK, 2000. Springer-Verlag.
- [43] Marianne HUCHARD, Mohamed Rouane HACENE, Cyril ROUME et Petko VALTCHEV : Relational concept discovery in structured datasets. *Ann. Math. Artif. Intell.*, 49(1-4):39–76, 2007.
- [44] B. HÉRAULT et O. HONNAY : Using life-history traits to achieve a functional classification of habitats. *Applied Vegetation Science*, 10:73–80, 2007.
- [45] P. JONES et P. MULHOLLAND, éditeurs. *Streams and Ground Waters*. academic press, 1999.
- [46] L. KAUTSKY : Life strategies of aquatic soft-bottom macrophytes. In *Oikos*, pages 126–135, 1988.
- [47] M. KAYTOUE-UBERALL, Z. ASSAGHIR, Messai N. et A. NAPOLI : Two complementary classification methods for designing a concept lattice from interval data. In *FoIKS*, pages 345–362, 2010.
- [48] M. LAFONT : A conceptual approach to the biomonitoring of freshwater : the ecological ambience system. *Journal of Limnology*, 6:17–24, 2001.
- [49] M. LAFONT, P. BREIL, P. NAMOUR, J.-C. CAMUS, F. MALARD et P. LE PIMPEC : Concept d’ambiance écologique dans les systèmes aquatiques continentaux. In *Actes du séminaire "État écologique des milieux aquatiques continentaux"*, pages 136–153, 2001.
- [50] M. LAFONT, A. VIVIER, S. NOGUEIRA, P. NAMOUR et P. BREIL : Surface and hyporheic oligochaete assemblages in a french suburban stream. In *Aquatic Oligochaete Biology IX*, pages 183–193, 2006.
- [51] Josse J. LÊ, S. et F. HUSSON : Factominer : An r package for multivariate analysis.
- [52] N. MESSAI : Analyse de concepts formels guidée par des connaissances de domaine : Application à la découverte de ressources génomiques sur le web. Thèse de doctorat, Université Henri Poincaré. Nancy 1, 2009.
- [53] N. MESSAI, M. D. DEVIGNES, A. NAPOLI et M. SMAÏL-TABBONE.
- [54] MINISTÈRE DE L’ÉCOLOGIE, DE L’ÉNERGIE, DU DÉVELOPPEMENT DURABLE ET DE L’AMÉNAGEMENT DU TERRITOIRE : Evaluation de l’état des eaux douces de surface de métropole, 2009.
- [55] A. NAPOLI : A smooth introduction to symbolic methods in knowledge discovery. In *Categorization in Cognitive Science*. Elsevier, 2006.
- [56] G. POLAILLON : Organisation et interprétation par les treillis de galois de données de type multivalué, intervalle ou histogramme. Thèse de doctorat, Université Paris IX Dauphine, 1998.

- [57] C. D. PRESTON : Pondweeds of great britain and ireland. Botanical Society of the British Isles, London, 1995.
- [58] C. D. PRESTON et J. M. CROFT : Aquatic plants in britain and ireland. Harley Books, Colchester., 1997.
- [59] M. H. ROUANE, M. DAO, M. HUCHARD et P. VALTCHEV : Analyse formelle de données relationnelles pour la réingénierie des modèles UML. In *LMO'07 : Langages et Modèles à Objets*, pages 151–166. Hermès Lavoisier, 03 2007.
- [60] C. STACE : New flora of the british isles. Cambridge University Press, Cambridge, 1991.
- [61] J.-F. STAERCK : Analyse des traits biologiques de macrophytes aquatiques en relation avec des perturbations types. Mémoire professionnel, 2005.
- [62] G. STUMME : Hierarchies of conceptual scales. In T. B. GAINES, R. KREMER et M. MUSEN, éditeurs : *Proceedings of Workshop on Knowledge Acquisition, Modeling and Management (KAW'99), Banff*, volume 2, pages 78–95, 1999.
- [63] K. THOMPSON, J. BAKKER et R. BEKKER : The soil seed banks of north western europe. Cambridge University Press, Cambridge, 1997.
- [64] P.-L. TISSERANT : Evaluation des pressions physiques dans les cours d'eau de la plaine d'alsace et comparaison d'indices biologiques pour l'évaluation de la qualité des milieux aquatiques. Ecole Nationale du Génie de l'Eau et de l'Environnement de Strasbourg, 2008.
- [65] T. G. TUTIN, V. H. HEYWOOD, N. A. BURGESS, D. M. MOORE, D. H. VALENTINE, S. M. WALTERS et D. A. WEBB : Flora europeaea. Cambridge University Press, Cambridge, 1964-1980.
- [66] P. USSEGLIO-POLATERA, M. BOURNAUD, P. RICHOUX et H. TACHET : Biological and ecological traits of benthic freshwater macroinvertebrates : relationships and definition of groups with similar traits. *Freshwater Biology*, 43, 2000.
- [67] P. USSEGLIO-POLATERA, P. RICHOUX, M. BOURNAUD et H. TACHET : A functional classification of benthic macroinvertebrates based on biological and ecological traits : application to river condition assessment and stream management. *Archiv für Hydrobiologie. Supplementband. Monographische Beiträge ISSN 0341-2881*, 139, 2001.
- [68] J. VERNEAUX : Cours d'eau de franche-comté (massif du jura). recherches écologiques sur le réseau hydrographique du doubs-essai de biotypologie. *Annales de l'Université de Franche Comté, Biologie Animale* 3 :1-260, 1973.
- [69] N. J. WILLBY, V. J. ABERNETHY et B. O. L. DEMARS : Attribute-based classification of european hydrophytes and its relationship to habitat utilization. *Freshwater Biology*, 43, January 2000.

- [70] S. YEVTUSHENKO et CONTRIBUTEURS : [http ://conexp.sourceforge.net/](http://conexp.sourceforge.net/), 2000-2006.
- [71] L. A. ZADEH : Fuzzy sets. *Information Control*, 8:338–353, 1965.
- [72] M. J. ZAKI et C.-J. HSIAO : Charm : An efficient algorithm for closed association rule mining. Rapport technique, Computer Science, Rensselaer Polytechnic Institute, 1999.

Glossaire biologique

(1) **Benthos ou benthon.** Ensemble des espèces aquatiques fixes ou mobiles vivant sur les fonds ou à leur proximité, et qui en dépendent pour leur nourriture [*Grand Dictionnaire Terminologique*].

(2) **Biocénose.** Ensemble des organismes vivants qui vivent dans les mêmes conditions de milieu et dans un espace donné [*Grand Dictionnaire Terminologique*].

NB : Une biocénose vit sur un biotope donné.

(3) **Biotope.** Milieu physique relativement bien délimité dans l'espace et de composition strictement définie, servant de support aux espèces constituant la biocénose [*Grand Dictionnaire Terminologique*].

NB : le biotope héberge une biocénose tandis que l'habitat héberge une espèce. Dans un biotope, l'ensemble des facteurs physiques et chimiques de l'environnement reste sensiblement constant ou subit des variations périodiques.

(4) **Bivoltin.** adjectif qualifiant un organisme réalisant deux générations par an (voir monovoltin, plurivoltin, semivoltin)

(5) **Bryophytes.** Embranchement du règne végétal : plante de petite taille feuillée ou à lame foliacée vivant dans les milieux humides terrestres et aquatiques. Elles ont la propriété d'accumuler les métaux et de garder en mémoire le passage des pollutions [*Atlas 1997*]

(6) **Characées ou Charophytes.** Végétaux aquatiques des eaux douces (milieu lacustre) et saumâtres (milieu lagunaire), proches des Algues et qui affectionnent les eaux calcaires. Les rameaux et les organes reproducteurs femelles (oogones) s'incrudent de carbonate de calcium permettant ainsi une excellente conservation [*TICE de l'académie de Toulouse*].

(7) **Crénon.** Section d'extrême amont des rivières, sous l'influence directe des sources (notamment en ce qui concerne la température et les substances dissoutes) (voir aussi rhithron, potamon)

(8) Diapause. Arrêt temporaire de l'activité ou du développement chez les insectes en hiver ou en cas de carence alimentaire.

(9) Eaux côtières. les eaux de surface situées en-deçà d'une ligne dont tout point est situé à une distance d'un mille marin au-delà du point le plus proche de la ligne de base servant pour la mesure de la largeur des eaux territoriales et qui s'étendent, le cas échéant, jusqu'à la limite extérieure d'une eau de transition *Journal officiel des Communautés européennes*.

(10) Eaux intérieures. toutes les eaux stagnantes et les eaux courantes à la surface du sol et toutes les eaux souterraines en amont de la ligne de base servant pour la mesure de la largeur des eaux territoriales *Journal officiel des Communautés européennes*.

(11) Eaux souterraines. Toutes les eaux se trouvant sous la surface du sol dans la zone de saturation et en contact direct avec le sol ou le sous-sol *Journal officiel des Communautés européennes*.

(12) Eaux de surface. Les eaux intérieures, à l'exception des eaux souterraines, les eaux de transition et les eaux côtières, sauf en ce qui concerne leur état chimique, pour lequel les eaux territoriales sont également incluses *Journal officiel des Communautés européennes*.

(13) Eaux de transition. Des masses d'eaux de surface à proximité des embouchures de rivières, qui sont partiellement salines en raison de leur proximité d'eaux côtières, mais qui sont fondamentalement influencées par des courants d'eau douce *Journal officiel des Communautés européennes*.

(14) Ecosystème. Ensemble des organismes vivants en interaction avec le biotope dans lequel ils vivent [*Grand Dictionnaire Terminologique*].

NB : L'écosystème est l'unité fonctionnelle de base en écologie, puisqu'elle inclut à la fois les êtres vivants : la biocénose, et le milieu : le biotope, dans lequel ils vivent, avec toutes les interactions entre le milieu et les organismes.

(15) Etiage. Le plus bas niveau d'un cours d'eau.

(16) Eurytherme. Adjectif qualifiant un organisme acceptant de vivre dans une gamme étendue de variations de température.

(17) Eutrophe. Se dit d'un milieu de capacité nutritive riche à très riche en nutriments [*Grand Dictionnaire Terminologique*].

(18) Eutrophisation. : Enrichissement de l'eau, qu'elle soit douce ou saline, par des nutriments, en particulier par des composés d'azote et de phosphore, qui accélèrent la croissance d'algues et des formes plus développées de la vie végétale [*Grand Dictionnaire Terminologique*].

(19) Habitat. Au sens de la DCE , c'est le milieu dans lequel vit une espèce ou un groupe d'espèces animales ou végétales (ex. : tourbières, roselières d'estuaire, chênaies, ...). Ce sont des zones terrestres ou aquatiques possédant des caractéristiques biogéographiques et géologiques particulières et uniques [*Directive 92-43-CEE du 21/05/92*].

(20) Hélophyte. Plante aquatique dont la base est enracinée sous l'eau et la majeure partie des tiges hors de l'eau (type roseau) (voir aussi hydrophyte).

(21) Hydrobiologie. Science ayant pour objet l'étude de la biologie des organismes aquatiques.

(22) Hydrophyte. Plante aquatique, enracinée ou non, flottant sous ou à la surface de l'eau (types nénuphar et potamot) (voir aussi hélophyte).

(23) (Zone) hyporhéique. Zone de sédiments saturés où s'effectuent les échanges d'eau et de matière se trouvant dans les eaux interstitielles (interface entre les eaux de surface et les eaux souterraines).

(24) Lentique (ou lénitique). Qualifie un milieu ou un habitat d'eau douce à vitesse nulle à très lente (les habitats lenticques prédominent dans le potamon) (voir aussi lotique).

(25) Lit. Chenal d'écoulement (d'un cours d'eau, d'un glacier par exemple).

– **Lit mineur.** Lit en eau en permanence, délimité par des berges.

– **Lit majeur.** Partie adjacente au chenal, inondé en cas de crue.

[*Trésor de la Langue Française*].

(26) Lithophile. Qui croît sur du rocher.

(27) Lotique. Qualifie un milieu ou un habitat d'eau douce à écoulement turbulent, à débit rapide formant des tourbillons. Leur lit est le plus souvent pierreux (les habitats lotiques prédominent dans le rhithron) (voir aussi lentique).

(28) Macroinvertébrés. Invertébrés aquatiques dont la taille est supérieure à 500 μm , c'est à dire visibles à l'œil nu. On y trouve notamment des vers plats et annelés, des mollusques, des crustacés et surtout des insectes.

(29) Macrophytes. Végétaux aquatiques visibles à l'œil nu.

(30) Masse d'eau. Portion de cours d'eau, canal, Aquifère, plan d'eau ou zone côtière homogène. Il s'agit d'un découpage élémentaire des milieux aquatiques destiné à être l'unité d'évaluation de la DCE. Une masse d'eau de surface est une partie distincte et significative des Eaux de surface, telles qu'un lac, un réservoir, une rivière, un fleuve ou un canal, une partie de rivière, de fleuve ou de canal, une eau de transition ou une portion d'eaux côtières. Pour les cours d'eau la délimitation des masses d'eau est basée principalement sur la taille du cours d'eau et la notion d'hydro-écorégion. Les masses d'eau sont regroupées en types homogènes qui servent de base à la définition de la notion de Bon état. Une masse d'eau souterraine est un volume distinct d'eau souterraine à l'intérieur d'un ou de plusieurs aquifères [*Eaufrance / Agence de l'eau Adour-Garonne*].

(31) Mésotrophe. Milieu aquatique où la teneur en matières nutritives est moyenne.

(32) Mésosaprobe. Adjectif qualifiant un milieu moyennement riche en matières organiques (ou des organismes moyennement polluo-sensibles) (voir aussi pollution organique). On distingue deux sous-classes : α et β mésosaprobe, avec α correspondant à un milieu plus riche en matière organique que β (ou avec β pour des organismes plus polluo-sensibles que α).

(33) Monovoltin (ou univoltin). Adjectif qualifiant un organisme réalisant une génération par an (voir aussi bivoltin, plurivoltin, semivoltin).

(34) Niche écologique. Place occupée par une espèce végétale ou animale dans un écosystème en fonction de ses relations avec son environnement. Situation qu'occupe un organisme vivant dans le système écologique, considéré du point de vue fonctionnel. Ensemble à la fois spatial et fonctionnel occupé par une des populations⁽⁴¹⁾ constituant une biocénose [*Grand Dictionnaire Terminologique*].

(35) Oligosaprobe. Adjectif qualifiant un milieu pauvre en matières organiques (ou des organismes polluosensibles).

(36) Oligotrophe. Se dit d'un milieu de faible capacité nutritive, c'est à dire pauvre en nutriments [*Grand Dictionnaire Terminologique*].

(37) (Zone) pélagique. Zone de pleine eau d'un lac, par opposition à la zone littorale.

(38) Peuplement. Ensemble des espèces animales ou végétales ou microbiennes vivant dans un endroit donné à un instant donné [*Grand Dictionnaire Terminologique*].

(39) Phénologie. Etude des phénomènes récurrents caractéristiques (d'une espèce vivante ou d'un biotope) et de leur évolution selon les saisons et les climats [*Encarta*].

(40) Phylogénie. Formation et développement des espèces vivantes au cours des temps ; étude de ce processus [*Trésor de la langue française*].

(41) Population. Ensemble d'individus de même espèce, occupant un territoire déterminé et présentant des caractéristiques propres [*Grand Dictionnaire Terminologique*].

(42) Potamon. Section d'aval des rivières, où la vitesse du courant $< 0,75$ m/s. Le substrat est majoritairement "déposé" (graviers, sable, vase) (voir aussi crénon, rhithron).

(43) Plurivoltin. Adjectif qualifiant un organisme réalisant plusieurs (plus de deux) générations par an (voir bivoltin, monovoltin, semivoltin).

(44) Pression. Exercice d'une activité humaine qui peut avoir une incidence sur les milieux aquatiques. Il peut s'agir de rejets, prélèvements d'eau, artificialisation des milieux aquatiques, capture de pêche, ... [*Glossaire DCE 2004*].

(45) Ptéridophyte. Les ptéridophytes (du grec pteros, aile et phytos, végétal), ou fougères, sont un embranchement du règne des végétaux. Les plantes de cet embranchement sont des végétaux vasculaires : ils disposent d'un système de conduction de la sève et présente un cycle de reproduction original [<http://www.futura-sciences.com>].

(46) Quiescence. Pause dans le processus de développement d'un insecte, résultante de conditions climatiques.

(47) Résilience. D'une façon générale, c'est la capacité d'un système à revenir à un état antérieur.

- Dans un cadre géologique, il s'agit de la faculté qu'a un substrat de revenir à son équilibre après avoir subi des perturbations ; mais c'est aussi la résistance qu'un substrat (ou sol) oppose à l'évolution de ses caractéristiques à la suite d'une perturbation extérieure, qu'elle soit physique, chimique ou biologique.
- Dans le cadre biologique, la résilience représente l'aptitude des communautés à retourner, après une perturbation, à une situation stable différente ou non de celle observée initialement.

(48) Rhéophile. Dit d'une espèce vivante, qui vit dans des eaux à courant vif : elle possède des caractères adaptatifs pour résister au courant, en milieu torrentiel par exemple [*Grand Dictionnaire Terminologique*].

NB : S'oppose à **limnophile**.

(49) Rhithron. Section d'amont des rivières, où la vitesse du courant $> 0,75$ m/s. Le substrat y est majoritairement "érodé" (galets, pierres, roches) (voir aussi crénon, potamon).

(50) Rhizome. Tige souterraine ou affleurante généralement horizontale émettant des racines et des tiges [*Dictionnaire botanique gardenbreizh*].

(51) SANDRE. Le Service d'Administration Nationale des Données et Référentiels sur l'Eau (SANDRE) élabore le langage commun des données sur l'eau. A ce titre, il est chargé au sein du Système National des Informations sur l'Eau (SIE) d'établir la normalisation des données afin de rendre compatible et homogène la définition et l'échange des données entre les producteurs, les utilisateurs et les banques de données. Il propose pour cela des modèles et dictionnaires de données, des formats d'échange et des listes de référence [*Glossaire DCE 2004*].

(52) Saprobie. Teneur d'une eau en matière organique.

(53) Semivoltin. Adjectif qualifiant un organisme réalisant une demi-génération par an (ou une génération par deux ans) (voir bivoltin, plurivoltin, bivoltin).

(54) Site de mesure. Sous-espace caractéristique et représentatif de la Station de mesure, clairement identifié et localisé afin d'y effectuer de façon répétitive des mesures pour une connaissance approfondie du milieu à l'endroit de la station. Les sites de mesure sont aussi les lieux sur la station où le préleveur devra effectuer, dans la mesure du possible, ses prélèvements ou ses mesures in situ. En règle générale, un site est consacré à un support : eau, sédiments, bryophytes ... Un support peut être prélevé en plusieurs sites. Chaque site de mesure peut appartenir à un ou plusieurs réseaux de mesure et faire l'objet d'une ou plusieurs utilisations pour chacun desquels la ou les périodes d'appartenance seront précisées. On dit aussi point de mesure [*SANDRE v1.1 : Eaux superficielles 1997-1*].

(55) Site de référence. Site sélectionné, sur un lac ou un cours d'eau, pour la bonne qualité de l'écosystème aquatique constituant une référence pour caractériser la qualité des autres lacs ou cours d'eau [*Portail des données sur l'eau du bassin Adour-Garonne*].

(56) Spermatophyte. Plante se reproduisant par l'intermédiaire de graines.

(57) Sténoèce. Désigne les êtres vivants qui présentent une niche écologique étroite et ayant une faible capacité d'adaptation lors de variations de facteurs écologiques propres à son habitat. ce sont des espèces à basse valence ou amplitude écologique, cantonnées à un petit nombre de milieux [*Dictionnaire environnement et développement durable*].

(58) Stolon. Ramification rampante de la base de la tige d'une plante [*Dictionnaire botanique gardenbreizh*].

(59) Substrat. Tout matériau servant de support physique à des organismes (exemple : galets d'un cours d'eau) [*Glossaire DCE 2004*].

(60) Surber. Instrument de prélèvement des macroinvertébrés, dont l'usage est préconisé dans la norme AFNOR de l'IBGN (Indice Biologique Global Normalisé). Il est constitué d'un cadre de surface normée : $1/20\text{ m}^2$ et d'un filet dérivant de maille normée : $500\mu\text{m}$.

(61) Taxon. Élément d'une classification qui regroupe tous les organismes vivants ayant en commun certains caractères, de quelque rang qu'ils soient. Du plus petit au plus grand ensemble, on distingue généralement les taxons suivants : variété, sous-espèce, espèce, sous-genre, genre, tribu, sous-famille, famille, super-famille, sous-ordre, ordre, sous-classe, sous-embranchement et embranchement (ou phylum) [*Grand Dictionnaire Terminologique*].

(62) Taxonomie (ou taxinomie). Branche des sciences biologiques qui vise à établir les bases scientifiques et systématiques d'un système de classification qui regroupe tous les organismes vivants en catégories appelées taxons, suivant les caractères qu'ils ont en commun, des plus généraux aux plus particuliers [*Grand Dictionnaire Terminologique*].

(63) Tégument. Ensemble des tissus et des formations organiques (peau, poils, plumes, carapace, écailles, etc.) qui constituent le revêtement externe du corps de l'homme et des animaux [*Trésor de la Langue Française*].

(64) Trophie. Etat de nutrition ou de développement [*Trésor de la Langue Française*].

(65) Tubificidae. Vers annélides aquatiques qui se reproduisent en abondance dans les eaux polluées par des matières organiques. La partie antérieure de son corps est enterré dans les sédiments, tandis que l'arrière se meut dans l'eau.

(66) Turion. Jeune pousse issue de la souche d'une plante vivace [*Dictionnaire botanique gardenbreizh*].

Annexe A

Sommaire

A.1	Stations du projet Indices	156
A.2	Indice IBGN	158
A.3	Groupes biotypologiques de Verneaux	161
A.4	Implications entre les propriétés Lxy	164
A.5	Algorithme NextClosure	181
A.6	Concepts issus du contexte histogramme pour la validation des algorithmes	182

A.1 Stations du projet Indices

FIGURE A.1 – Carte de localisation des stations dans la plaine d'Alsace [64].



Typologie (d'après DCE, 2005)	Station (nom par masse d'eau)	Rivière	Localisation	Code station
Ruisseaux phréatiques	Zembs	Zembs	Hilsenheim	ZEMB001
	Brunnwasser amont	Brunnwasser	Rhinau	BRUN001
	Brunnwasser aval	Brunnwasser	Daubensand	BRUN002
	Muhlbach de Schoenau amont	Rhin de Biesheim	Kunheim	MSCH001
	Muhlbach de Schoenau aval	Muhlbach	Schoenau	MSCH002
	Muhlbach de Gerstheim amont	Heulachgraben	Diebolsheim	MGER001
	Muhlbach de Gerstheim aval	Muhlbach	Gerstheim	MGER002
Petits CE eaux calmes et tempérées	Rothbach	Rothgraben	Forêt de Weitbruch	ROTH001
	Lomdgraben amont	Lomdgraben	Morschwiller	LOMD001
	Lomdgraben aval	Lomdgraben	Schweighouse sur Moder	
	Rhin tortu amont	Rhin tortu	Plobsheim	RHTO001
	Rhin tortu aval	Rhin tortu	Strasbourg	RHTO002
	Souffel centre	Souffel	Stutzheim-Offenheim	SOUF001
	Souffel aval	Souffel	Bieshheim	SOUF002
Petits CE eaux calmes et fraîches	Brumbach	Brumbach	Forêt de Haguenau	BRUMB001
	Sauer 2	Sauer	Woerth	SAUE001
	Rosenmeer amont	Rosenmeer	Rosheim	ROSE001
	Rosenmeer aval	Rosenmeer	Rosheim	
	Horgiessen amont	Horgiessen	Près de l'Illwald	HORG001
	Horgiessen aval	Muhlbach	Sélestat	HORG002
	Seltzbach amont	Seltzbach	Preuschdorf	SELT001
	Seltzbach centre	Seltzbach	Kutzenhausen	SELT002
	Seltzbach aval	Seltzbach	Seltz	SELT003
Petits CE eaux vives et fraîches	Thalbach	Thalbach	Wittersdorf	THAL001
	Zinsel du sud 2 aval	Zinsel du sud	Steinbourg	ZISU001
	Breitbrunnwasser	Breitbrunnwasser	Illhausern	BREI001
	Steinbach amont	Steinbach	Obersteinbach	
	Steinbach aval	Steinbach	Niedersteinbach	STEI002
	Moder 2 amont	Moder	Ingwiller	MODE002
	Moder 2 aval	Moder	Obermodern	MODE002
	Dollerbaechlein aval	Dollerbaechlein	Ruelisheim	DOLL001
	Ill 2 amont	Ill	Waldighoffen	ILL_ 001
	Ill 2 aval	Ill	Tagolsheim	ILL_ 002
Grand CE eaux calmes et tempérées	Ill 7 aval	Ill	Moulin de la Wantzenau	ILL_ 006
Grands CE eaux calmes et fraîches	Zinsel du Nord 2 amont	Zinsel du Nord	Zinswiller	ZINO001
	Zinsel du Nord 2 centre	Zinsel du Nord	Gumbrechtshoffen	
	Zinsel du Nord 2 aval	Zinsel du Nord	Uttenhoffen	ZINO003
	Ill 5 aval	Ill	Sélestat	ILL_ 005
	Ill 4 amont	Ill	Ruelisheim	ILL_ 003
	Ill 4 aval	Ill	Houssen	ILL_ 004
Grands CE eaux vives et fraîches	Fecht 4	Fecht	Guémar	FECH001
?	Riedbrunnen (nvlle réf)	Riedrunnen	Illhausern	RIED001
?	Etzgerbach (nvlle réf)	Etzgerbach	Mettau (Suisse)	ETZG001

TABLE A.1 – Liste des 43 stations sélectionnées pour le projet Indices.

A.2 Indice IBGN

FIGURE A.2 – Tableau d'échantillonnage vierge pour l'indice IBGN

MESURE HYDROBIOLOGIQUE							
Cours d' eau :		Station :		Date du prélèvement :			
Hydrologie		Situation :					
☐	Etiage	Coordonnées Lambert	X :	km	Altitude :	m	
☐	Moyennes eaux		Y :	km			
☐	Autres situations (à préciser) :						
	Hydrologie des jours précédents :						
Conditions de prélèvement							
☐	Facile						
☐	Difficile (préciser pourquoi) :						
		Largeur du lit mouillé :				m	
		longueur de la station				m	

VITESSES SUPERFICIELLES	V	cm/s	V > 150	150 > V > 75	75 > V > 25	25 > V > 5	V < 5
SUPPORTS	S	V	2	4	5	3	1
Bryophytes	9	N° (R) h S					
Spermaphytes immergés	8	N° (R) h S					
Eléments organiques grossiers (litières, branchages, racines)	7	N° (R) h S					
Sédiments minéraux de grande taille (pierres, cailloux, galets) 250 mm > Ø > 25 mm	6	N° (R) h S					
Granulats grossiers 25 mm > Ø > 2.5 mm	5	N° (R) h S					
Spermaphytes émergents de la strate basse	4	N° (R) h S					
Sédiments fins ± organiques ("vases") Ø < 0.1 mm	3	N° (R) h S					
Sables et limons Ø < 2.5 mm	2	N° (R) h S					
Surfaces naturelles et artificielles (roches, dalles, sols, parois) blocs > 250 mm	1	N° (R) h S					
Algues, ou à défaut, mame et argile	0	N° (R) h S					

N° = Numéro de l'échantillon 1 à 8
R = Recouvrement du couple S - V
h = Hauteur d'eau : en cm
S = Support prélevé: par ex. Galets
Toutes les mesures sont estimées

(1) accessoire (≤ 1 %)
(2) peu abondant (< 10 %)
R = (3) abondant (10 - 50 %)
(4) très abondant (> 50 %)

Habitat dominant	N°(R)	
	h	
	S	

Classe de variété	VT	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Taxons ⁽⁶¹⁾ indicateurs	GI	>50	49	44	40	36	32	28	24	20	16	12	9	6	3
		45	41	37		33	29	25	21	17	13	10	7	4	1
Chloroperlidae															
Perlidae															
Perlodidae	9	20	20	20	19	18	17	16	15	14	13	12	11	10	9
Taeniopterygidae															
Capniidae															
Brachycentridae															
Odontocéridae	8	20	20	19	18	17	16	15	14	13	12	11	10	9	8
Philopotamidae															
Leuctridae															
Glossosomatidae															
Beraeidae	7	20	19	18	17	16	15	14	13	12	11	10	9	8	7
Goeridae															
Leptophlébiidae															
Nemouridae															
Lepidostomatidae															
Sericostomatidae	6	19	18	17	16	15	14	13	12	10	9	8	7	6	5
Ephemeridae															
Hydroptilidae															
Heptageniidae															
Polymitarcidae	5	18	17	16	15	14	13	12	11	10	9	8	7	6	5
Potamanthidae															
Leptoceridae															
Polycentropodidae															
Psychomyidae	4	17	16	15	14	13	12	11	10	9	8	7	6	5	4
Rhyacophilidae															
Limnephilidae (1)															
Ephemerellidae (1)															
Hydropsychidae	3	16	15	14	13	12	11	10	9	8	7	6	5	4	3
Aphelocheiridae															
Baetidae (1)															
Caenidae (1)															
Elmidae (1)	2	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Gammaridae (1)															
Mollusques															
Chironomidae (1)															
Asellidae (1)															
Achètes	1	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Oligochètes (1)															

TABLE A.2 – Tableau de détermination des classes de variétés pour l'indice IBGN. Les taxons marqués d'un (1) nécessitent d'être représentés par au moins 10 individus. Les autres par au moins 3.

A.3 Groupes biotypologiques de Verneaux

Niveau 1	niveau2	niveau 3	niveau4
Capnia	Baetis	Baetis	Baetis
Chloroperla	Brachyptera	Brachyptera	Bithynia
Drusinae	Capnia	Caenis	Brachyptera
Ecdyonurus	Colymbetinae	Capnia	Calopteryx
Ephemerella	Drusinae	Chloroperla	Capnia
Erpobdellidae	Ecdyonurus	Ecdyonurus	Cheumatopsyche
Glossiphoniidae	Ephemerella	Ephemerella	Chloroperla
Hydroptila	Ephemerella	Ephemerella	Colymbetinae
Leuctra	Gammarus	Erpobdellidae	Cordulegaster
Limnephilinae	Glossiphoniidae	Gammarus	Drusinae
Lymnaea	Hydropsyche	Glossiphoniidae	Ecdyonurus
Nemoura	Leuctra	Habrophlebia	Ephemerella
Neureclipsis	Libellula	Hydropsyche	Ephemerella
Odontocerum	Limnephilinae	Hydroptila	Erpobdellidae
Philopotamus	Lymnaea	Leuctra	Gammarus
Phryganea	Nemoura	Limnephilinae	Glossiphoniidae
Platycnemis	Nemurella	Lymnaea	Habrophlebia
Polycentropus	Neureclipsis	Mystacides	Heptagenia
Potamanthus	Odontocerum	Nemoura	Hydropsyche
Protonemura	Paraleptophlebia	Nemurella	Hydroptila
Radix	Planorbidae	Neureclipsis	Lepidostoma
Rhithrogena	Plectrocnemia	Odontocerum	Leuctra
Rhyacophila	Polycentropus	Paraleptophlebia	Limnephilinae
Sericostoma	Protonemura	Philopotamus	Lymnaea
Silo	Psychomyia	Physa	Lype
Valvata	Rhithrogena	Plectrocnemia	Mystacides
Planorbidae	Rhyacophila	Polycentropus	Nemoura
	Taeniopteryx	Protonemura	Nemurella
	Valvata	Radix	Neureclipsis
		Rhithrogena	Odontocerum
		Rhyacophila	Paraleptophlebia
		Sericostoma	Phryganea
		Silo	Plectrocnemia
		Taeniopteryx	Polycentropus
			Protonemura
			Psychomyia
			Radix
			Rhithrogena
			Rhyacophila
			Sericostoma
			Silo
			Taeniopteryx
			Torleya

TABLE A.3 – Taxons alsaciens des 4 premiers niveaux de Verneaux.

niveau5	Niveau 6	Niveau 7	niveau8	Niveau 9
Baetis	Baetis	Baetis	Baetis	Baetis
Brachyptera	Bithynia	Bithynia	Bithynia	Bithynia
Caenis	Brachyptera	Brachyptera	Brachyptera	Caenis
Calopteryx	Caenis	Caenis	Caenis	Calopteryx
Capnia	Calopteryx	Cheumatopsyche	Calopteryx	Cheumatopsyche
Cheumatopsyche	Chloroperla	Chloroperla	Cheumatopsyche	Coenagrionidae
Chloroperla	Cyrnus	Cyrnus	Coenagrionidae	Colymbetinae
Cyrnus	Ecdyonurus	Ecdyonurus	Colymbetinae	Cordulegaster
Ecdyonurus	Ephemera	Ecnomus	Cordulegaster	Cyrnus
Ecnomus	Ephemerella	Ephemera	Cyrnus	Ecdyonurus
Ephemera	Gammarus	Ephemerella	Ecdyonurus	Ecnomus
Erpobdellidae	Glossiphoniidae	Gammarus	Ecnomus	Ephemera
Gammarus	Goera	Glossiphoniidae	Ephemera	Ephemerella
Glossiphoniidae	Gomphus	Goera	Ephemerella	Erpobdellidae
Goera	Habrophlebia	Habrophlebia	Erpobdellidae	Gammarus
Gomphus	Heptagenia	Heptagenia	Gammarus	Glossiphoniidae
Habrophlebia	Hydropsyche	Hydropsyche	Glossiphoniidae	Gomphus
Heptagenia	Hydroptila	Hydroptila	Gomphus	Heptagenia
Hydropsyche	Lepidostoma	Lepidostoma	Habrophlebia	Hydropsyche
Hydroptila	Leuctra	Leuctra	Heptagenia	Hydroptila
Lepidostoma	Limnephilinae	Limnephilinae	Hydropsyche	Leptocerus
Leuctra	Lype	Lymnaea	Hydroptila	Limnephilinae
Libellula	Micrasema	Lype	Leptocerus	Lymnaea
Limnephilinae	Mystacides	Micrasema	Leuctra	Mystacides
Lype	Nemoura	Mystacides	Limnephilinae	Odontocerum
Micrasema	Odontocerum	Nemoura	Lymnaea	Physa
Nemoura	Paraleptophlebia	Odontocerum	Lype	Planorbidae
Odontocerum	Planorbidae	Paraleptophlebia	Micrasema	Platycnemis
Paraleptophlebia	Platycnemis	Piscicolidae	Mystacides	Polycentropus
Phryganea	Polycentropus	Planorbidae	Odontocerum	Potamanthus
Polycentropus	Protonemura	Polycentropus	Paraleptophlebia	Potamopyrgus
Protonemura	Psychomyia	Potamanthus	Physa	Psychomyia
Psychomyia	Radix	Protonemura	Piscicolidae	Radix
Rhyacophila	Rhyacophila	Psychomyia	Planorbidae	Rhyacophila
Sericostoma	Sericostoma	Radix	Platycnemis	Taeniopteryx
Silo	Silo	Rhithrogena	Polycentropus	Torleya
Taeniopteryx	Taeniopteryx	Rhyacophila	Potamanthus	
Torleya	Torleya	Silo	Potamopyrgus	
	Valvata	Taeniopteryx	Protonemura	
		Torleya	Psychomyia	
			Radix	
			Rhithrogena	
			Rhyacophila	
			Sphaerium	
			Taeniopteryx	
			Torleya	

TABLE A.4 – Taxons alsaciens des 5 derniers niveaux de Verneaux.

A.4 Implications entre les propriétés Lxy

Ces implications sont numérotées de 1 à 545 dans l'ordre décroissant de leur support qui est indiqué entre chevrons.

- 1 < 19 > C10 ==> M10 P30
- 2 < 19 > D30 ==> F10
- 3 < 19 > F10 ==> D30
- 4 < 19 > M10 ==> C10 P30
- 5 < 19 > P30 ==> C10 M10
- 6 < 18 > M50 ==> D30 F10
- 7 < 18 > T10 ==> C10 M10 P30
- 8 < 17 > R10 ==> D30 F10 M50
- 9 < 16 > C20 ==> D30 F10
- 10 < 16 > C30 D30 F10 ==> M50
- 11 < 16 > L30 ==> C10 M10 P30 T10 V10
- 12 < 16 > C10 D30 F10 M10 M50 P30 ==> R10
- 13 < 15 > D30 F10 F20 ==> M50
- 14 < 15 > C20 D30 F10 M50 ==> C30
- 15 < 15 > P13 ==> C10 M10 P30
- 16 < 15 > C10 C40 M10 P30 ==> T10
- 17 < 14 > P40 ==> C10 M10 P13 P30 V10
- 18 < 14 > C10 M10 P13 P30 V10 ==> P40
- 19 < 13 > C53 ==> C10 D30 F10 M10 M50 P30 R10
- 20 < 13 > D10 ==> C10 M10 P30
- 21 < 13 > P21 ==> C10 D30 F10 M10 M50 P30 R10 V10
- 22 < 12 > F40 ==> D30 F10 M50 V10
- 23 < 12 > R23 ==> D30 F10 M50 R10 R30
- 24 < 12 > D30 F10 M50 R10 R30 ==> R23
- 25 < 12 > C10 C53 D30 F10 M10 M50 P30 R10 T10 ==> C40
- 26 < 12 > C10 C40 D30 F10 M10 M50 P30 R10 T10 ==> C53
- 27 < 12 > C10 M10 P30 R30 T10 V10 ==> L30
- 28 < 11 > G20 ==> C10 D30 F10 M10 M50 P30 R10 V10
- 29 < 11 > C10 D10 D30 F10 M10 P30 ==> M50 R10
- 30 < 11 > C10 C30 D10 M10 P30 T10 ==> C40
- 31 < 11 > C10 C40 D10 M10 P30 T10 ==> C30
- 32 < 11 > D30 F10 F40 M50 R10 V10 ==> C10 M10 P30
- 33 < 11 > V30 ==> C10 M10 P13 P30
- 34 < 10 > F30 ==> C40
- 35 < 10 > F33 ==> D30 F10 F40 M50 V10
- 36 < 10 > G11 ==> C40 D30 F10 M50

37 < 10 > L20 $\implies$ C30 C40 V10
 38 < 10 > M40 $\implies$ R30 V10
 39 < 10 > C10 D10 M10 P13 P30 $\implies$ F20
 40 < 10 > C10 C53 D10 D30 F10 M10 M50 P30 R10 $\implies$ C20 C30
 41 < 10 > C10 C30 D10 D30 F10 M10 M50 P30 R10 $\implies$ C20 C53
 42 < 10 > C10 M10 P13 P30 R30 $\implies$ V30
 43 < 10 > C10 D30 F10 G20 M10 M50 P30 R10 T10 V10 $\implies$ L30
 44 < 10 > C10 D30 F10 F40 M10 M50 P30 R10 T10 V10 $\implies$ L30
 45 < 10 > C10 D30 F10 F40 M10 M50 P21 P30 R10 V10 $\implies$ C30
 46 < 10 > C10 C30 D30 F10 G20 M10 M50 P30 R10 V10 $\implies$ C20
 47 < 10 > C10 C30 D30 F10 F40 M10 M50 P30 R10 V10 $\implies$ P21
 48 < 9 > C30 D23 $\implies$ F20
 49 < 9 > F43 $\implies$ C40 F30
 50 < 9 > M20 $\implies$ R30 V10
 51 < 9 > C10 D23 M10 P30 $\implies$ P13
 52 < 9 > C10 C53 D30 F10 M10 M50 P13 P30 R10 $\implies$ F20
 53 < 9 > R20 $\implies$ C40 V10
 54 < 9 > C10 C30 C40 M10 P13 P30 T10 $\implies$ F20
 55 < 9 > T33 $\implies$ C10 D30 F10 M10 P30 V10
 56 < 9 > C10 C20 C30 C53 D10 D30 F10 M10 M50 P30 R10 V10 $\implies$ P21
 57 < 9 > D30 F10 M40 R30 V10 $\implies$ M50
 58 < 9 > D30 F10 F20 F40 M50 V10 $\implies$ F33
 59 < 9 > C10 D30 F10 G20 M10 M50 P21 P30 R10 V10 $\implies$ C20 C30 F40
 60 < 9 > C10 D10 D30 F10 M10 M50 P21 P30 R10 V10 $\implies$ C20 C30 C53
 61 < 9 > C10 C30 D30 F10 M10 M50 P13 P30 P40 R10 V10 $\implies$ P21
 62 < 9 > C10 C20 C30 D30 F10 F40 M10 M50 P21 P30 R10 V10 $\implies$ G20
 63 < 9 > C10 C20 C30 C53 D30 F10 M10 M50 P21 P30 R10 V10 $\implies$ D10
 64 < 9 > C10 M10 P13 P30 P40 T10 V10 V30 $\implies$ L30
 65 < 8 > C40 D23 F20 $\implies$ C30
 66 < 8 > C40 F20 F30 $\implies$ F43
 67 < 8 > G13 $\implies$ C10 D30 F10 G20 M10 M50 P30 R10 V10
 68 < 8 > L11 $\implies$ C10 D30 F10 L30 M10 P30 T10 V10
 69 < 8 > D23 D30 F10 M50 $\implies$ F20
 70 < 8 > C10 D23 F20 M10 P13 P30 $\implies$ D10
 71 < 8 > T22 $\implies$ C10 C20 C30 D30 F10 G20 M10 M50 P30 R10 V10
 72 < 8 > T32 $\implies$ C10 M10 P30 T10 V10
 73 < 8 > T40 $\implies$ C30 D30 F10 M50 R30 V10
 74 < 8 > C40 F20 R20 V10 $\implies$ C30
 75 < 8 > C40 D30 F10 F40 M50 V10 $\implies$ C30
 76 < 8 > C30 C40 R20 V10 $\implies$ F20

77 < 8 > C30 C40 D30 F10 L20 M50 V10 ==> C20
 78 < 8 > C10 M10 M40 P30 R30 V10 ==> L30 T10
 79 < 8 > C10 D30 F10 M10 P30 T10 T33 V10 ==> L30
 80 < 8 > C10 D30 F10 M10 M50 P30 R10 T33 V10 ==> C30 F40 P21
 81 < 8 > C10 D30 F10 F40 M10 M50 P13 P30 P40 R10 V10 ==> F33
 82 < 8 > C10 D10 D30 F10 G20 M10 M50 P30 R10 V10 ==> F40
 83 < 8 > C10 D10 D30 F10 F40 M10 M50 P30 R10 V10 ==> G20
 84 < 8 > C10 C53 D30 F10 G20 M10 M50 P30 R10 V10 ==> C20 C30
 85 < 8 > C10 C53 D30 F10 F40 M10 M50 P30 R10 V10 ==> C30 P21
 86 < 8 > C10 C53 D30 F10 F20 M10 M50 P13 P30 P40 R10 V10 ==> P21
 87 < 8 > C10 C40 M10 P30 R20 T10 V10 ==> P13 P40
 88 < 8 > C10 C30 C40 L20 M10 P30 T10 V10 ==> D10 L30
 89 < 8 > V23 ==> C10 M10 P30 R30 V10
 90 < 8 > V43 ==> D30 F10 F20 M50 V10
 91 < 7 > D21 ==> C10 C30 D30 F10 L30 M10 M50 P21 P30 R10 T10 V10
 92 < 7 > C40 D23 D30 F10 ==> C20
 93 < 7 > C30 C40 F30 F43 ==> F20
 94 < 7 > L13 ==> C10 C30 L30 M10 P13 P30 P40 T10 V10
 95 < 7 > L21 ==> C10 C30 D30 F10 M10 M50 P30 R10 T10
 96 < 7 > M23 ==> C10 D10 D30 F10 F40 G20 L30 M10 M50 P30 R10 T10 V10
 97 < 7 > M32 ==> C10 C20 C30 D30 F10 M10 M50 P30 R10 T10 V10
 98 < 7 > C40 D30 F10 F30 M50 ==> R10
 99 < 7 > C30 D23 D30 F10 F20 M50 ==> C20
 100 < 7 > C10 C20 C30 C53 D30 F10 F20 M10 M50 P13 P30 R10 ==> D10
 101 < 7 > C10 C20 C30 D30 F10 F20 M10 M50 P13 P30 R10 T10 ==> V21
 102 < 7 > C10 C40 F30 M10 P13 P30 T10 ==> F43
 103 < 7 > T21 ==> C10 M10 P30 T10 V10
 104 < 7 > C30 C40 D23 F20 V10 ==> L20
 105 < 7 > C10 C20 C30 D30 F10 M10 M50 P30 R10 R23 R30 V10 ==> G20
 106 < 7 > C10 C40 C53 D30 F10 G11 M10 M50 P30 R10 T10 V10 ==> P21
 107 < 7 > C40 M40 R30 V10 ==> C30
 108 < 7 > C40 D30 F10 R20 V10 ==> C20
 109 < 7 > C30 M20 R30 V10 ==> F20
 110 < 7 > C20 C30 C40 D30 F10 F40 M50 V10 ==> L20
 111 < 7 > C10 M10 M20 P30 R30 V10 ==> P13 P40 V30
 112 < 7 > C10 D30 F10 M10 P30 T10 T32 V10 ==> M50 R10
 113 < 7 > C10 D30 F10 L30 M10 M40 M50 P30 R10 R23 R30 T10 V10 ==> F40
 114 < 7 > C10 D30 F10 G20 M10 M50 P13 P30 P40 R10 V10 ==> F33 F40 G13
 115 < 7 > C10 D30 F10 F40 L30 M10 M50 P30 R10 R23 R30 T10 V10 ==> M40
 116 < 7 > C10 D30 F10 F40 G13 G20 M10 M50 P30 R10 V10 ==> F33 P13 P40

117 < 7 > C10 D23 M10 P13 P30 P40 T10 V10 ==> L30
 118 < 7 > C10 D10 D30 F10 F40 G20 L30 M10 M50 P30 R10 T10 V10 ==> M23
 119 < 7 > C10 C30 D30 F10 M10 M50 P21 P30 R10 R23 R30 V10 ==> F40
 120 < 7 > C10 C30 C40 F20 M10 P13 P30 P40 R20 T10 V10 ==> D10
 121 < 7 > C10 C30 C40 D10 F20 M10 P13 P30 P40 T10 V10 ==> R20
 122 < 7 > C10 C20 C30 D30 F10 F20 G20 M10 M50 P30 R10 V10 ==> T22
 123 < 7 > V21 ==> C10 C20 C30 D30 F10 F20 M10 M50 P13 P30 R10 T10
 124 < 7 > V31 ==> C20 C30 C40 D30 F10 M50 V10
 125 < 7 > C30 D30 F10 F20 M50 V10 V43 ==> C20
 126 < 6 > G22 ==> C10 C40 D30 F10 M10 P13 P30 T10
 127 < 6 > P20 ==> C10 M10 P30 R30 T10
 128 < 6 > P41 ==> C20 C30 C40 D30 F10 G11 M50 R30
 129 < 6 > C40 D30 F10 F30 F43 M50 R10 ==> F20
 130 < 6 > C30 C40 D30 F10 F30 M50 R10 ==> C20
 131 < 6 > C20 C30 C40 D30 F10 G11 M50 R30 ==> P41
 132 < 6 > C10 C20 C30 C40 C53 D30 F10 G11 M10 M50 P30 R10 T10 ==> D10
 133 < 6 > T20 ==> C40 F20 R30
 134 < 6 > C30 C40 D30 F10 G11 M50 R30 V10 ==> M40 T40
 135 < 6 > C20 C30 C40 D30 F10 M50 R30 V10 ==> V31
 136 < 6 > D30 F10 M20 M50 R30 V10 ==> F20
 137 < 6 > C30 C40 F20 L20 R20 V10 ==> D23
 138 < 6 > C30 C40 D30 F10 M50 R30 T40 V10 ==> G11 M40
 139 < 6 > C30 C40 D30 F10 M40 M50 R30 V10 ==> G11 T40
 140 < 6 > C30 C40 D30 F10 F33 F40 M50 V10 ==> F20
 141 < 6 > C10 M10 P13 P30 P40 T10 T32 V10 ==> F20
 142 < 6 > C10 D10 D30 F10 F33 F40 G20 M10 M50 P30 R10 V10 ==> F20
 143 < 6 > C10 C40 D23 L30 M10 P13 P30 P40 T10 V10 ==> R20
 144 < 6 > C10 C30 D30 F10 M10 M50 P30 R10 R23 R30 T40 V10 ==> F40 P21
 145 < 6 > C10 C30 D30 F10 F40 L30 M10 M40 M50 P21 P30 R10 R23 R30 T10 V10 ==> D21
 146 < 6 > C10 C30 D30 F10 F20 F33 F40 M10 M50 P21 P30 R10 T33 V10 ==> C53
 147 < 6 > C10 C30 D21 D30 F10 F40 L30 M10 M50 P21 P30 R10 T10 V10 ==> M40 R23 R30
 148 < 6 > C10 C30 C53 D30 F10 F33 F40 M10 M50 P21 P30 R10 V10 ==> F20 T33
 149 < 6 > C10 C20 D30 F10 L11 L30 M10 P30 T10 V10 ==> C40
 150 < 6 > C10 C20 C40 D30 F10 M10 P13 P30 P40 T10 V10 ==> R20
 151 < 6 > C10 C20 C30 D30 F10 M10 M32 M50 P21 P30 R10 T10 V10 ==> P13 P40
 152 < 6 > C10 C20 C30 D30 F10 G13 G20 M10 M50 P30 R10 T22 V10 ==> F20
 153 < 6 > C10 C20 C30 D30 F10 F33 F40 G20 M10 M50 P21 P30 R10 T22 V10 ==> F20
 154 < 6 > C10 C20 C30 D30 F10 F20 M10 M32 M50 P30 R10 T10 V10 ==> C40 C53
 155 < 6 > C10 C20 C30 C40 C53 D30 F10 M10 M32 M50 P30 R10 T10 V10 ==> F20
 156 < 6 > C10 C20 C30 C40 C53 D10 D30 F10 L20 L30 M10 M50 P21 P30 R10 T10 V10 ==>

F40 G20 M23

157 < 6 > V20 ==> D30 F10 M40 M50 R30 V10

158 < 6 > C10 D23 M10 P13 P30 V30 ==> R30

159 < 6 > V41 ==> C10 M10 P13 P30 P40 V10 V30

160 < 6 > C40 D30 F10 F20 M50 V10 V43 ==> C20 C30

161 < 6 > C10 D30 F10 F20 M10 M50 P30 R10 V10 V43 ==> G13 G20 L30 T10

162 < 6 > C10 D30 F10 F20 G13 G20 L30 M10 M50 P30 R10 T10 V10 ==> V43

163 < 5 > C50 ==> C40 D23 M20 R30 V10

164 < 5 > C40 D23 F30 ==> F43 R30

165 < 5 > L10 ==> C20 C30 D30 F10 F20 M50

166 < 5 > M30 ==> D23 F20 R30 V10

167 < 5 > M31 ==> C10 C20 D30 F10 M10 P30 R30 V10

168 < 5 > M41 ==> C10 C30 C40 F20 L30 M10 P30 T10 V10

169 < 5 > R33 ==> C10 C20 C30 C40 C53 D10 D30 F10 F20 M10 M32 M50 P13 P21 P30 P40
R10 R20 T10 V10 V21

170 < 5 > C10 C20 D23 D30 F10 M10 P13 P30 T10 ==> C40

171 < 5 > C10 C40 M10 P20 P30 R30 T10 ==> F30 F43

172 < 5 > C10 C40 D30 F10 F30 F43 M10 P13 P30 T10 ==> G22

173 < 5 > C10 C40 C53 D30 F10 F20 G22 M10 M50 P13 P30 R10 T10 ==> G11

174 < 5 > C10 C40 C53 D30 F10 F20 G11 M10 M50 P13 P30 R10 T10 ==> G22

175 < 5 > C10 C30 C40 F20 F30 F43 M10 P13 P30 T10 ==> D10

176 < 5 > C10 C30 C40 D10 F20 F30 F43 M10 P30 T10 ==> P13

177 < 5 > C10 C30 C40 C53 D30 F10 L21 M10 M50 P30 R10 T10 ==> F20

178 < 5 > C40 D30 F10 F20 M50 R30 T20 ==> G11

179 < 5 > T41 ==> C10 C20 D30 F10 L30 M10 P13 P30 P40 T10 T33 V10

180 < 5 > C40 D23 R30 V10 ==> C50 M20

181 < 5 > C30 D23 F20 R30 V10 ==> M20

182 < 5 > C10 D10 D30 F10 M10 M50 P30 R10 R23 R30 V10 ==> F40 G20

183 < 5 > C40 F20 R30 T20 V10 ==> M20

184 < 5 > M20 M40 R30 V10 ==> C30 F20

185 < 5 > D23 F20 M20 R30 V10 ==> C30

186 < 5 > C40 F30 R20 V10 ==> C10 F43 M10 P13 P30 P40 T10

187 < 5 > C40 F30 M20 R30 V10 ==> F43

188 < 5 > C30 D30 F10 F33 F40 M50 R30 T40 V10 ==> F20

189 < 5 > C30 D30 F10 F20 M40 M50 R30 V10 ==> T40

190 < 5 > C30 D30 F10 F20 M20 M50 R30 V10 ==> T40

191 < 5 > C30 D30 F10 F20 F33 F40 M50 R30 V10 ==> T40

192 < 5 > C30 C40 D30 F10 F40 M50 R30 V10 ==> G11 M40 T40

193 < 5 > C30 C40 D30 F10 F40 G11 M50 V10 ==> M40 R30 T40

194 < 5 > C20 C30 C40 D30 F10 G11 L20 M50 V10 ==> M40 P41 R30 T40 V20 V31

195 < 5 > C20 C30 C40 D30 F10 F20 M50 R10 R20 V10 ==> C10 C53 D10 M10 M32 P13
 P21 P30 P40 R33 T10 V21
 196 < 5 > C10 L30 M10 P13 P30 P40 T10 T21 V10 ==> V30
 197 < 5 > C10 F20 L30 M10 P13 P30 P40 T10 T32 V10 ==> V30
 198 < 5 > C10 D30 F10 L11 L30 M10 M50 P21 P30 R10 T10 V10 ==> C40 C53 G11
 199 < 5 > C10 D30 F10 G20 L11 L30 M10 M50 P30 R10 T10 V10 ==> R23 R30
 200 < 5 > C10 D23 D30 F10 M10 P13 P30 P40 T33 V10 ==> C20
 201 < 5 > C10 D10 D30 F10 F20 F33 F40 G13 G20 M10 M50 P13 P30 P40 R10 V10 ==> D23
 202 < 5 > C10 D10 D23 D30 F10 F20 M10 M50 P13 P30 P40 R10 V10 ==> F33 F40 G13 G20
 203 < 5 > C10 C40 M10 P30 T10 T21 V10 ==> F30
 204 < 5 > C10 C40 F30 M10 P30 T10 T32 V10 ==> F20 F43
 205 < 5 > C10 C40 F30 F43 L30 M10 P30 R30 T10 V10 ==> V23
 206 < 5 > C10 C40 F30 F43 L30 M10 P13 P30 P40 T10 V10 ==> V30
 207 < 5 > C10 C40 F20 M10 P30 T10 T32 V10 ==> F30 F43
 208 < 5 > C10 C40 D30 F10 F30 L30 M10 P30 T10 V10 ==> L11
 209 < 5 > C10 C40 C53 D30 F10 L30 M10 M50 P21 P30 R10 R23 R30 T10 V10 ==> G11
 210 < 5 > C10 C30 D30 F10 M10 M50 P30 R10 T10 T32 V10 ==> C20
 211 < 5 > C10 C30 D30 F10 L21 M10 M50 P21 P30 R10 T10 V10 ==> P13 P40
 212 < 5 > C10 C30 D30 F10 L13 L30 M10 M50 P13 P21 P30 P40 R10 T10 V10 ==> C20 F33
 F40 G13 G20
 213 < 5 > C10 C30 D30 F10 L11 L30 M10 M50 P30 R10 T10 V10 ==> C20 C40 C53
 214 < 5 > C10 C30 C53 D30 F10 F40 M10 M50 P21 P30 R10 R23 R30 V10 ==> T40
 215 < 5 > C10 C30 C40 F20 L13 L30 M10 P13 P30 P40 T10 V10 ==> D10 D23 L20 R20
 216 < 5 > C10 C30 C40 D10 L30 M10 P30 R30 T10 V10 ==> L20
 217 < 5 > C10 C30 C40 D10 D23 F20 L20 L30 M10 P13 P30 P40 R20 T10 V10 ==> L13
 218 < 5 > C10 C30 C40 C53 D30 F10 G11 L30 M10 M50 P21 P30 R10 T10 V10 ==> D21
 219 < 5 > C10 C30 C40 C53 D21 D30 F10 L30 M10 M50 P21 P30 R10 T10 V10 ==> G11
 220 < 5 > C10 C20 D30 F10 L30 M10 P13 P30 P40 T10 T33 V10 ==> T41
 221 < 5 > C10 C20 D23 D30 F10 M10 P13 P30 P40 V10 ==> T33
 222 < 5 > C10 C20 C40 D30 F10 L30 M10 P30 R30 T10 V10 ==> L11
 223 < 5 > C10 C20 C30 D30 F10 G20 L30 M10 M32 M50 P30 R10 T10 V10 ==> G13
 224 < 5 > C10 C20 C30 D30 F10 F40 G20 M10 M50 P21 P30 R10 T22 T33 V10 ==> C53 D10
 F20 F33
 225 < 5 > C10 C20 C30 D30 F10 F33 F40 G13 G20 L30 M10 M50 P13 P21 P30 P40 R10 T10
 V10 ==> L13
 226 < 5 > C10 C20 C30 C53 D30 F10 G13 G20 M10 M50 P30 R10 V10 ==> F20 T22
 227 < 5 > C30 D30 F10 M40 M50 R30 V10 V20 ==> C20 C40 G11 L20 P41 T40 V31
 228 < 5 > C10 F20 L30 M10 P30 R30 T10 V10 V23 ==> C40
 229 < 5 > C10 C40 F30 L30 M10 P30 R30 T10 V10 V23 ==> F43
 230 < 5 > C10 C30 D30 F10 M10 M50 P13 P30 R10 T10 V30 ==> L21

231 < 5 > C10 C40 D30 F10 M10 P13 P30 T10 V30 ==> G22
 232 < 5 > C10 D30 F10 F33 F40 M10 M50 P13 P30 P40 R10 V10 V30 ==> R23 R30
 233 < 5 > C10 M10 P13 P20 P30 R30 T10 V30 ==> D23
 234 < 5 > C10 L30 M10 M20 P13 P30 P40 R30 T10 V10 V23 V30 ==> C40
 235 < 5 > C10 F20 M10 P13 P30 P40 R30 V10 V23 V30 ==> M20
 236 < 5 > C10 D23 M10 P13 P30 R30 T10 V30 ==> P20
 237 < 5 > C10 D10 F20 M10 P13 P30 R30 V30 ==> D23
 238 < 5 > C10 C40 L30 M10 P13 P30 P40 R30 T10 V10 V30 ==> M20 V23
 239 < 5 > C10 C30 F20 M10 P13 P30 P40 R30 V10 V30 ==> M20
 240 < 5 > C20 C30 C40 D30 F10 L20 M50 V10 V31 ==> G11 M40 P41 R30 T40 V20
 241 < 5 > C20 C30 C40 D30 F10 G11 M40 M50 P41 R30 T40 V10 V31 ==> L20 V20
 242 < 5 > V40 ==> C10 C40 L30 M10 P30 R30 T10 V10
 243 < 5 > C10 D30 F10 M10 P13 P30 P40 V10 V30 V41 ==> M50 P21 R10
 244 < 5 > C10 M10 P13 P30 P40 R30 V10 V30 V41 ==> V23
 245 < 5 > C20 C30 D23 D30 F10 F20 M50 V10 V43 ==> C40 L20
 246 < 5 > C20 C30 C40 D30 F10 F20 L20 M50 V10 V43 ==> D23
 247 < 5 > C20 C30 C40 D23 D30 F10 F20 L20 M50 V10 ==> V43
 248 < 4 > D11 ==> C40 D30 F10 M20 R30 V10
 249 < 4 > G23 ==> C30 C40 C50 D23 F20 L20 M20 M30 R30 V10
 250 < 4 > L33 ==> C20 C30 C40 D30 F10 F20 G11 L10 M50
 251 < 4 > M33 ==> C10 D30 F10 F20 M10 M50 P13 P30 R10 R23 R30 T10 V30
 252 < 4 > M42 ==> C10 C40 C53 D30 F10 F20 F30 F43 G11 G22 M10 M50 P13 P30 R10 T10
 253 < 4 > C20 C30 D23 D30 F10 F20 L10 M50 ==> R30
 254 < 4 > C20 C30 C40 D30 F10 F20 L10 M50 ==> G11 L33
 255 < 4 > R13 ==> C40 C50 D23 M20 R20 R30 V10
 256 < 4 > C20 C30 D30 F10 F20 L10 M50 R30 ==> D23
 257 < 4 > C20 C30 D23 D30 F10 F20 M50 R30 ==> L10
 258 < 4 > C10 D10 M10 P20 P30 R30 T10 ==> D23 F20 P13 V30
 259 < 4 > C10 C40 F20 F30 F43 M10 P20 P30 R30 T10 ==> C30
 260 < 4 > C10 C40 C53 D30 F10 F20 F30 F43 G11 M10 M50 P30 R10 T10 ==> G22 M42 P13
 261 < 4 > C10 C30 M10 P20 P30 R30 T10 ==> C40 F20 F30 F43
 262 < 4 > C10 C30 C40 F20 F30 F43 M10 P30 R30 T10 ==> P20
 263 < 4 > C10 C30 C40 C53 D30 F10 F20 L21 M10 M50 P13 P30 R10 T10 ==> G11 G22
 264 < 4 > C10 C30 C40 C53 D30 F10 F20 G11 L21 M10 M50 P30 R10 T10 ==> G22 P13
 265 < 4 > C10 C30 C40 C53 D30 F10 F20 G11 G22 M10 M50 P13 P30 R10 T10 ==> L21
 266 < 4 > C10 C20 C40 D30 F10 G22 M10 P13 P30 T10 ==> F30 F43
 267 < 4 > C10 C20 C30 C40 C53 D30 F10 F20 L21 M10 M50 P30 R10 T10 ==> F30 F43
 268 < 4 > C10 C20 C30 C40 C53 D30 F10 F20 F30 F43 M10 M50 P30 R10 T10 ==> L21
 269 < 4 > C10 C20 C30 C40 C53 D10 D30 F10 M10 M50 P30 R10 R23 R30 T10 ==> G11 P41
 270 < 4 > C10 C20 C30 C40 C53 D10 D30 F10 F30 M10 M50 P30 R10 T10 ==> G11

271 < 4 > C10 C40 F20 M10 P30 R30 T10 T20 ==> P13 V30
 272 < 4 > T42 ==> C10 F20 L30 M10 P30 T10 T32 V10
 273 < 4 > T43 ==> C10 C40 F20 F30 F43 M10 P13 P30 T10
 274 < 4 > C20 D23 D30 F10 R30 V10 ==> M20
 275 < 4 > C10 C40 C53 D30 F10 F20 F30 F43 M10 M50 P30 R10 T10 V10 ==> T32
 276 < 4 > C10 C20 C30 D30 F10 F20 L21 M10 M50 P30 R10 T10 V10 ==> T32
 277 < 4 > D30 F10 F20 F33 F40 M20 M50 R30 V10 ==> C30 T40
 278 < 4 > D23 M40 R30 V10 ==> F20 M30
 279 < 4 > D23 D30 F10 M20 R30 V10 ==> C20
 280 < 4 > C40 R20 R30 V10 ==> C50 D23 M20 R13
 281 < 4 > C40 D30 F10 F20 M20 M50 R30 V10 ==> G11 T20
 282 < 4 > C30 D23 F20 M20 M30 R30 V10 ==> C40 C50 G23 L20
 283 < 4 > C30 C40 F30 L20 V10 ==> R30
 284 < 4 > C30 C40 F20 M20 R30 T20 V10 ==> M40
 285 < 4 > C30 C40 F20 M20 M40 R30 V10 ==> T20
 286 < 4 > C30 C40 F20 L20 M20 R30 V10 ==> C50 D23 G23 M30
 287 < 4 > C10 M10 P30 T10 T21 T32 V10 ==> D10 F20 P13 P40
 288 < 4 > C10 F20 M10 P30 T10 T21 V10 ==> D10 P13 P40 T32
 289 < 4 > C10 D30 F10 M10 M50 P21 P30 R10 T10 T21 V10 ==> C20 C30
 290 < 4 > C10 D30 F10 L11 L30 M10 P13 P30 P40 T10 V10 ==> V30
 291 < 4 > C10 D30 F10 G20 L30 M10 M50 P30 R10 T10 T32 V10 ==> R23 R30
 292 < 4 > C10 D30 F10 F40 L11 L30 M10 M50 P30 R10 T10 V10 ==> D10 G20 M23 M40
 R23 R30 V20
 293 < 4 > C10 D30 F10 F20 M10 M50 P21 P30 R10 T10 T32 V10 ==> P13 P40
 294 < 4 > C10 D10 F20 M10 P30 T10 T32 V10 ==> P13 P40 T21
 295 < 4 > C10 D10 D30 F10 F40 G20 L30 M10 M23 M40 M50 P30 R10 R23 R30 T10 V10
 ==> L11 V20
 296 < 4 > C10 C40 F30 F43 M10 P30 T10 T21 V10 ==> P13 P40 R20
 297 < 4 > C10 C40 D30 F10 G22 L30 M10 P13 P30 P40 T10 V10 ==> V30
 298 < 4 > C10 C40 D30 F10 F30 M10 P30 T10 T21 V10 ==> C20
 299 < 4 > C10 C40 C53 D30 F10 M10 M50 P21 P30 R10 T10 T32 V10 ==> G11
 300 < 4 > C10 C40 C53 D30 F10 L30 M10 M50 P30 R10 T10 T32 V10 ==> L11
 301 < 4 > C10 C40 C53 D30 F10 F30 M10 M50 P21 P30 R10 T10 V10 ==> G11
 302 < 4 > C10 C30 D30 F10 M10 M50 P30 R10 T10 T21 V10 ==> C20 P21
 303 < 4 > C10 C30 D30 F10 L21 L30 M10 M50 P13 P21 P30 P40 R10 T10 V10 ==> D21 V30
 304 < 4 > C10 C30 D30 F10 F40 M10 M50 P21 P30 R10 R23 R30 T33 T40 V10 ==> C53
 305 < 4 > C10 C30 D21 D30 F10 L30 M10 M50 P13 P21 P30 P40 R10 T10 V10 ==> L21 V30
 306 < 4 > C10 C30 C40 F30 M10 P30 T10 T21 V10 ==> D10
 307 < 4 > C10 C30 C40 F20 L30 M10 M41 P13 P30 P40 T10 V10 ==> D10 D23 L13 L20 R20
 308 < 4 > C10 C30 C40 D10 F20 L30 M10 M41 P30 T10 V10 ==> D23 L13 L20 P13 P40 R20

309 < 4 > C10 C20 D30 F10 L30 M10 M31 P30 R30 T10 V10 ==> C40 L11 V40
 310 < 4 > C10 C20 C40 D30 F10 L30 M10 P13 P30 P40 R20 T10 T33 T41 V10 ==> D23
 311 < 4 > C10 C20 C30 D30 F10 L21 M10 M50 P30 R10 T10 T32 V10 ==> F20
 312 < 4 > C10 C20 C30 D30 F10 G20 M10 M31 M50 P30 R10 R23 R30 V10 ==> C53 D10
 F40 P21 T40
 313 < 4 > C10 C20 C30 D30 F10 F40 G20 M10 M50 P21 P30 R10 R23 R30 T22 V10 ==> T40
 314 < 4 > C10 C20 C30 D30 F10 F33 F40 G13 G20 L13 L30 M10 M50 P13 P21 P30 P40 R10
 T10 T33 T41 V10 ==> M32
 315 < 4 > C10 C20 C30 D30 F10 F33 F40 G13 G20 L13 L30 M10 M32 M50 P13 P21 P30 P40
 R10 T10 V10 ==> T33 T41
 316 < 4 > C10 C20 C30 D30 F10 F20 M10 M50 P30 R10 T10 T32 V10 ==> L21
 317 < 4 > C10 C20 C30 C53 D10 D30 F10 F40 G20 M10 M50 P21 P30 R10 R23 R30 T40 V10
 ==> M31
 318 < 4 > C10 C20 C30 C40 C53 D10 D21 D30 F10 G11 L30 M10 M50 P21 P30 R10 T10 V10
 ==> L11
 319 < 4 > C10 D30 F10 F40 L30 M10 M40 M50 P30 R10 R23 R30 T10 V10 V20 ==> D10
 G20 L11 M23
 320 < 4 > C10 D30 F10 F20 M10 M50 P30 R10 R23 R30 V10 V23 ==> C53
 321 < 4 > C10 C53 D30 F10 M10 M50 P30 R10 R23 R30 V10 V23 ==> F20
 322 < 4 > C10 L30 M10 P20 P30 R30 T10 V10 V23 ==> C40 F30 F43
 323 < 4 > C10 D30 F10 M10 P30 R30 T33 V10 V23 ==> P13 P40 V30
 324 < 4 > C10 D30 F10 M10 M50 P21 P30 R10 R23 R30 V10 V23 ==> P13 P40 V30 V41
 325 < 4 > C10 C30 C40 L30 M10 P30 R30 T10 V10 V23 ==> F20
 326 < 4 > C10 C30 C40 D10 F20 M10 P13 P30 T10 V30 ==> F30 F43
 327 < 4 > C10 D30 F10 M10 P13 P30 P40 T33 V10 V30 ==> R30 V23
 328 < 4 > C10 C53 D30 F10 F20 M10 M50 P13 P21 P30 P40 R10 V10 V30 ==> V41
 329 < 4 > C10 C40 L30 M10 P13 P30 P40 R20 T10 V10 V30 ==> F30 F43
 330 < 4 > C10 C30 L13 L30 M10 P13 P30 P40 T10 V10 V30 ==> R30
 331 < 4 > C10 D30 F10 M10 M50 P13 P30 P40 R10 R23 R30 V10 V23 V30 ==> P21 V41
 332 < 4 > C10 D30 F10 F20 M10 M50 P13 P21 P30 P40 R10 R23 R30 V10 V30 ==> M20
 333 < 4 > C10 D30 F10 F20 M10 M20 M50 P13 P30 P40 R10 R23 R30 V10 V30 ==> P21
 334 < 4 > C10 D23 M10 P13 P30 P40 R30 V10 V23 V30 ==> M20
 335 < 4 > C10 D23 M10 M20 P13 P30 P40 R30 V10 V30 ==> V23
 336 < 4 > C10 C20 C30 C40 C53 D30 F10 L30 M10 M50 P30 R10 T10 V10 V31 ==> G20 L11
 R23 R30
 337 < 4 > C10 C20 C30 C40 C53 D10 D30 F10 M10 M50 P21 P30 R10 T10 V10 V31 ==> G11
 338 < 4 > C10 C40 D30 F10 L30 M10 P30 R30 T10 V10 V40 ==> C20 L11 M31
 339 < 4 > C10 C30 C40 L30 M10 P30 R30 T10 V10 V40 ==> D10 L20
 340 < 4 > C10 F20 L30 M10 P13 P30 P40 T10 V10 V30 V41 ==> C40
 341 < 4 > C10 D30 F10 F20 M10 M50 P13 P21 P30 P40 R10 V10 V30 V41 ==> C53

342 < 4 > C10 C40 L30 M10 P13 P30 P40 T10 V10 V30 V41 ==> F20
 343 < 4 > C10 M10 M20 P13 P30 P40 R30 V10 V23 V30 V41 ==> F20
 344 < 4 > C20 C30 C40 D30 F10 F20 M50 R20 V10 V43 ==> D23 F33 F40 L20
 345 < 4 > C10 C30 C40 C53 D30 F10 F20 L30 M10 M41 M50 P30 R10 T10 V10 ==> C20 G13
 G20 M32 T22 V43
 346 < 4 > C20 C30 C40 D23 D30 F10 F20 F33 F40 L20 M50 V10 V43 ==> R20
 347 < 4 > C10 C20 C30 C40 C53 D30 F10 F20 G13 G20 L30 M10 M50 P30 R10 T10 T22 V10
 V43 ==> M32 M41
 348 < 3 > C23 ==> C10 C40 F20 F30 F43 L30 M10 M20 P13 P30 P40 R30 T10 V10 V23 V30
 349 < 3 > C41 ==> C10 D30 F10 F20 F33 F40 G13 G20 M10 M50 P13 P30 P40 R10 R23 R30
 V10 V30
 350 < 3 > C52 ==> C10 D30 F10 F33 F40 G13 G20 L30 M10 M40 M50 P13 P30 P40 R10 R23
 R30 T10 V10 V30
 351 < 3 > D12 ==> C10 C30 D30 F10 F20 L21 L30 M10 M50 P30 R10 R23 R30 T10 V10
 352 < 3 > F23 ==> C10 C20 D30 F10 L30 M10 P30 R30 T10 T21 T33 V10
 353 < 3 > G10 ==> C10 C40 C50 D23 F30 F43 L30 M10 M20 P13 P20 P30 P40 R13 R20 R30
 T10 V10 V23 V30
 354 < 3 > M22 ==> C10 C20 C30 D30 F10 L21 M10 M50 P30 R10 R23 R30 T10
 355 < 3 > C20 C30 C40 D23 D30 F10 F20 G11 M50 ==> L10 L33 P41 R30 T20
 356 < 3 > P11 ==> C10 C20 C30 C40 C53 D10 D21 D30 F10 F40 G11 G20 L11 L20 L30 M10
 M23 M31 M40 M50 P21 P30 P41 R10 R23 R30 T10 T40 V10 V20 V31 V40
 357 < 3 > C20 C30 C40 D30 F10 F20 G11 L10 L33 M50 R10 ==> F30 F43
 358 < 3 > C10 C20 C30 D30 F10 F20 L10 M10 M50 P30 R10 ==> C53 D10 P13
 359 < 3 > C40 D30 F10 F20 F30 F43 G11 M50 R10 R23 R30 ==> T20
 360 < 3 > C10 D30 F10 M10 M50 P20 P30 R10 R23 R30 T10 ==> F20
 361 < 3 > C10 C40 D30 F10 F30 F43 M10 P20 P30 R30 T10 ==> C20
 362 < 3 > C10 C20 D30 F10 M10 P20 P30 R30 T10 ==> C40 F30 F43
 363 < 3 > C10 C20 C40 D30 F10 F30 F43 M10 P30 R30 T10 ==> P20
 364 < 3 > C30 C40 F20 F30 F43 R30 T20 ==> D23
 365 < 3 > C20 C30 C40 D30 F10 F20 G11 M50 P41 R30 T20 ==> D23 L10 L33
 366 < 3 > C20 C30 C40 D30 F10 F20 G11 L10 L33 M50 V10 ==> V31
 367 < 3 > C40 D30 F10 F30 F43 M20 R30 V10 ==> D11
 368 < 3 > C30 C40 F30 M40 R30 V10 ==> L20
 369 < 3 > C30 C40 F20 F30 F43 M20 R30 V10 ==> C50 D23 G23 L20 M30
 370 < 3 > C30 C40 F20 F30 F43 L20 R30 V10 ==> C50 D23 G23 M20 M30
 371 < 3 > C20 C40 D30 F10 M20 R30 V10 ==> C50 D11 D23
 372 < 3 > C10 L30 M10 P20 P30 R30 T10 T32 V10 ==> F20 T42
 373 < 3 > C10 L30 M10 P20 P30 R30 T10 T21 V10 ==> D23 P13 P40 V30
 374 < 3 > C10 L30 M10 M40 P30 R30 T10 T32 V10 ==> D30 F10 F40 G20 M50 R10 R23
 375 < 3 > C10 L30 M10 M40 P30 R30 T10 T21 V10 ==> D30 F10 F40 G20 M50 R10 R23

376 < 3 > C10 F20 L30 M10 P30 R30 T10 T32 T42 V10 ==> P20
 377 < 3 > C10 D30 F10 L30 M10 P30 T10 T21 T33 V10 ==> C20 F23 R30
 378 < 3 > C10 D30 F10 L30 M10 P20 P30 R30 T10 V10 ==> L11
 379 < 3 > C10 D30 F10 L30 M10 M50 P30 R10 R23 R30 T10 T21 V10 ==> F40 G20 M40
 380 < 3 > C10 D30 F10 L11 L30 M10 P30 T10 T33 V10 ==> C20 C40 M31 R30 V40
 381 < 3 > C10 D30 F10 L11 L30 M10 M50 P30 R10 T10 T21 V10 ==> D10
 382 < 3 > C10 D30 F10 G20 L30 M10 M50 P30 R10 T10 T21 V10 ==> F40 M40 R23 R30
 383 < 3 > C10 D30 F10 F40 L30 M10 M50 P30 R10 T10 T32 V10 ==> G20 M40 R23 R30
 384 < 3 > C10 D30 F10 F40 L30 M10 M50 P30 R10 T10 T21 V10 ==> G20 M40 R23 R30
 385 < 3 > C10 D30 F10 F20 L30 M10 M50 P30 R10 T10 T32 T42 V10 ==> L11
 386 < 3 > C10 D10 D30 F10 L30 M10 M50 P30 R10 T10 T32 V10 ==> L11
 387 < 3 > C10 D10 D30 F10 L30 M10 M50 P30 R10 T10 T21 V10 ==> L11
 388 < 3 > C10 C40 F30 L30 M10 P30 R30 T10 T21 V10 ==> V40
 389 < 3 > C10 C40 F20 F30 F43 L30 M10 P30 T10 T32 T42 V10 ==> C30
 390 < 3 > C10 C30 F20 L30 M10 P30 T10 T32 T42 V10 ==> C40 F30 F43
 391 < 3 > C10 C30 D30 F10 F20 L21 L30 M10 M50 P30 R10 R23 R30 T10 V10 ==> D12
 392 < 3 > C10 C30 C40 F20 F30 F43 L30 M10 P30 T10 T32 V10 ==> T42
 393 < 3 > C10 C20 D30 F10 L30 M10 P30 R30 T10 T21 V10 ==> F23 T33
 394 < 3 > C10 C20 C40 D30 F10 F30 F43 G22 M10 P13 P30 P40 R20 T10 V10 ==> T21
 395 < 3 > C10 C20 C30 D30 F10 M10 M50 P13 P21 P30 P40 R10 T10 T21 V10 ==> L21 M32
 396 < 3 > C10 C20 C30 D30 F10 M10 M32 M50 P30 R10 T10 T32 V10 ==> C40 C53 F20 F30
 F43 L21
 397 < 3 > C10 C20 C30 D30 F10 L30 M10 M50 P21 P30 R10 T10 T32 V10 ==> D21
 398 < 3 > C10 C20 C30 D30 F10 L30 M10 M50 P21 P30 R10 T10 T21 V10 ==> D21
 399 < 3 > C10 C20 C30 D30 F10 L21 M10 M32 M50 P13 P21 P30 P40 R10 T10 V10 ==> T21
 400 < 3 > C10 C20 C30 D30 F10 G20 L30 M10 M50 P30 R10 R23 R30 T10 T32 V10 ==> T22
 401 < 3 > C10 C20 C30 D30 F10 G20 L21 L30 M10 M50 P30 R10 T10 V10 ==> G13 R23 R30
 402 < 3 > C10 C20 C30 D30 F10 G13 G20 L30 M10 M50 P30 R10 R23 R30 T10 V10 ==> L21
 403 < 3 > C10 C20 C30 C40 C53 D30 F10 F20 F30 F43 L21 M10 M50 P30 R10 T10 T32 V10
 ==> M32
 404 < 3 > C10 C20 C30 C40 C53 D10 D30 F10 F30 G11 M10 M50 P21 P30 R10 T10 V10 ==>
 T21
 405 < 3 > D30 F10 F33 F40 M40 M50 R30 V10 V20 ==> F20
 406 < 3 > C10 D10 M10 P30 R30 V10 V23 ==> C30 D23 F20 M20 P13 P40 V30
 407 < 3 > C10 L30 M10 P30 R30 T10 T32 V10 V23 ==> C40 F20 F30 F43
 408 < 3 > C10 L30 M10 P30 R30 T10 T21 V10 V23 ==> P13 P40 V30
 409 < 3 > C10 L30 M10 M40 P30 R30 T10 V10 V23 ==> C30 P13 P40 V30 V41
 410 < 3 > C10 D30 F10 L11 L30 M10 P30 R30 T10 V10 V23 ==> C40 F30 F43
 411 < 3 > C10 D30 F10 G20 M10 M50 P30 R10 R23 R30 V10 V23 ==> C20 C30 G13
 412 < 3 > C10 D30 F10 F40 M10 M50 P30 R10 R23 R30 V10 V23 ==> C30 F33 P13 P21 P40

T33 V30 V41

413 < 3 > C10 C30 D30 F10 L30 M10 M50 P30 R10 R23 R30 T10 V10 V23 ==> L21

414 < 3 > C10 C40 F20 F30 F43 M10 P13 P30 T10 T43 V30 ==> R30 T20

415 < 3 > C10 F20 L30 M10 P13 P30 P40 T10 T32 T42 V10 V30 ==> D10 T21

416 < 3 > C10 D30 F10 L11 L30 M10 M50 P13 P30 P40 R10 T10 V10 V30 ==> F20 T32

417 < 3 > C10 D30 F10 F33 F40 G13 G20 L30 M10 M40 M50 P13 P30 P40 R10 R23 R30 T10 V10 V30 ==> C52

418 < 3 > C10 D30 F10 F20 L30 M10 M33 M50 P13 P30 P40 R10 R23 R30 T10 V10 V30 ==> M20 P21

419 < 3 > C10 D30 F10 F20 L30 M10 M20 M50 P13 P21 P30 P40 R10 R23 R30 T10 V10 V30 ==> M33

420 < 3 > C10 D30 F10 F20 F33 F40 G13 G20 M10 M50 P13 P30 P40 R10 R23 R30 V10 V30 ==> C41

421 < 3 > C10 D10 F20 L30 M10 P13 P30 P40 T10 T21 T32 V10 V30 ==> T42

422 < 3 > C10 D10 D23 F20 M10 M30 P13 P30 P40 R30 V10 V30 ==> L30 P20 T10

423 < 3 > C10 D10 D23 F20 L30 M10 P13 P20 P30 P40 R30 T10 V10 V30 ==> M30

424 < 3 > C10 C40 F20 L30 M10 M20 P13 P30 P40 R30 T10 T20 V10 V23 V30 ==> V41

425 < 3 > C10 C40 F20 F30 F43 M10 P13 P30 R30 T10 T20 V30 ==> T43

426 < 3 > C10 C40 F20 F30 F43 L30 M10 M20 P13 P30 P40 R30 T10 V10 V23 V30 ==> C23

427 < 3 > C10 C40 D30 F10 G22 L11 L30 M10 P13 P30 P40 T10 V10 V30 ==> F30 F43

428 < 3 > C10 C40 C53 D30 F10 F20 G11 G22 M10 M50 P13 P30 R10 R23 R30 T10 V30 ==> M33 T20

429 < 3 > C10 C40 C50 D23 F30 F43 L30 M10 M20 P13 P20 P30 P40 R13 R20 R30 T10 V10 V23 V30 ==> G10

430 < 3 > C10 C30 D30 F10 F20 L21 M10 M50 P13 P30 R10 R23 R30 T10 V30 ==> M33

431 < 3 > C10 C20 D30 F10 M10 P13 P30 P40 R30 V10 V23 V30 ==> T33

432 < 3 > C10 C40 F30 L30 M10 P30 R30 T10 V10 V40 ==> T21

433 < 3 > C10 C30 C40 D10 L20 L30 M10 M40 P30 R30 T10 V10 V40 ==> C20 C53 D21 D30 F10 F40 G11 G20 L11 M23 M31 M50 P11 P21 P41 R10 R23 T40 V20 V31

434 < 3 > C10 D10 F20 M10 P13 P30 P40 V10 V30 V41 ==> C30

435 < 3 > C10 C40 F20 L30 M10 M20 P13 P30 P40 R30 T10 V10 V23 V30 V41 ==> T20

436 < 3 > C10 C30 L30 M10 P13 P30 P40 R30 T10 V10 V23 V30 V41 ==> M40

437 < 3 > D23 D30 F10 F20 M50 R30 V10 V43 ==> M30 M40 V20

438 < 3 > D30 F10 F20 M20 M50 R30 V10 V43 ==> C20 C30 M40 T40

439 < 3 > D30 F10 F20 F33 F40 M50 R30 V10 V43 ==> M40

440 < 3 > C20 C30 D30 F10 F20 M50 R30 T40 V10 V43 ==> M20 M40

441 < 3 > D30 F10 F20 M40 M50 R30 V10 V20 V43 ==> D23 M30

442 < 3 > C20 C30 C40 D30 F10 F20 M50 V10 V31 V43 ==> P10 R30

443 < 3 > P10 ==> C20 C30 C40 D30 F10 F20 M50 R30 V10 V31 V43

444 < 3 > D23 D30 F10 F20 M30 M50 R30 V10 ==> M40 V20 V43

445 < 3 > C20 C30 D30 F10 F20 M20 M40 M50 R30 T40 V10 ==> V43
 446 < 3 > C10 D30 F10 G13 G20 L30 M10 M50 P30 R10 R23 R30 T10 T32 V10 ==> F20 V43
 447 < 3 > C10 D30 F10 F20 G20 L30 M10 M50 P30 R10 R23 R30 T10 T32 V10 ==> G13 V43
 448 < 3 > V32 ==> C10 C20 C30 C40 C53 D10 D23 D30 F10 F20 F33 F40 G13 G20 L13 L20
 L30 M10 M23 M32 M41 M50 P13 P21 P30 P40 R10 R20 R33 T10 T22 T33 T41 V10 V21 V43
 449 < 3 > C10 D30 F10 F20 G13 G20 L30 M10 M50 P30 R10 R23 R30 T10 V10 V43 ==> T32
 450 < 3 > C10 C20 C30 C40 C53 D10 D23 D30 F10 F20 F33 F40 G13 G20 L13 L20 L30 M10
 M23 M32 M41 M50 P13 P21 P30 P40 R10 R20 R33 T10 T22 T33 T41 V10 V21 V43 ==> V32
 451 < 2 > C33 ==> C10 C40 D11 D30 F10 F30 F43 G22 L11 L22 L30 M10 M20 P13 P30 P40
 R30 T10 V10 V23 V30
 452 < 2 > D22 ==> C10 C40 C53 D30 F10 F20 F30 F43 G11 G22 M10 M42 M50 P13 P21 P30
 P40 R10 T10 T32 T43 V10
 453 < 2 > D31 ==> C10 C23 C30 C40 C50 D10 D23 F13 F20 F30 F43 G10 G23 L13 L20 L30
 M10 M20 M30 M53 P13 P20 P30 P40 R13 R20 R30 T10 V10 V23 V30
 454 < 2 > F13 ==> C10 C23 C30 C40 C50 D10 D23 D31 F20 F30 F43 G10 G23 L13 L20 L30
 M10 M20 M30 M53 P13 P20 P30 P40 R13 R20 R30 T10 V10 V23 V30
 455 < 2 > L22 ==> C10 C33 C40 D11 D30 F10 F30 F43 G22 L11 L30 M10 M20 P13 P30 P40
 R30 T10 V10 V23 V30
 456 < 2 > L23 ==> C10 C41 D10 D23 D30 F10 F20 F33 F40 G13 G20 M10 M50 P13 P30 P40
 R10 R23 R30 V10 V30
 457 < 2 > M21 ==> C10 C20 C30 C40 C53 D10 D30 F10 F20 F30 F43 G11 G22 L21 M10 M32
 M42 M50 P13 P21 P30 P40 R10 R20 R33 T10 T21 T32 V10 V21
 458 < 2 > M43 ==> C10 C20 D23 D30 F10 M10 M20 M31 P13 P30 P40 R30 T33 V10 V23
 V30
 459 < 2 > M53 ==> C10 C23 C30 C40 C50 D10 D23 D31 F13 F20 F30 F43 G10 G23 L13 L20
 L30 M10 M20 M30 P13 P20 P30 P40 R13 R20 R30 T10 V10 V23 V30
 460 < 2 > C10 C20 C30 D30 F10 F20 L21 M10 M22 M50 P30 R10 R23 R30 T10 ==> C40 C53
 F30 F43 P20
 461 < 2 > C10 C20 C30 C40 C53 D30 F10 F20 F30 F43 L21 M10 M50 P20 P30 R10 R23 R30
 T10 ==> M22
 462 < 2 > T31 ==> C10 C30 C40 D10 D23 F20 F30 F43 M10 P13 P20 P30 R30 T10 T20 T43
 V30
 463 < 2 > C40 D11 D30 F10 F20 G11 M20 M50 R10 R23 R30 T20 V10 ==> F30 F43
 464 < 2 > C20 C30 D30 F10 F20 F33 F40 L10 M50 V10 ==> D23 M20 R30 T40
 465 < 2 > C10 L30 M10 M40 P20 P30 R30 T10 V10 ==> D10 D23 F20 M30 P13 P40 V30
 466 < 2 > C10 D30 F10 F33 F40 G20 L30 M10 M40 M50 P30 R10 R23 R30 T10 T21 V10 ==>
 C52 G13 P13 P40 V30
 467 < 2 > C10 C40 F20 F30 F43 M10 P13 P30 P40 T10 T32 T43 V10 ==> C53 D22 D30 F10
 G11 G22 M42 M50 P21 R10
 468 < 2 > C10 C30 C40 F20 L30 M10 M41 P30 R30 T10 V10 ==> F30 F43 P20 T32 T42 V23

469 < 2 > C10 C30 C40 F20 F30 F43 L30 M10 M41 P30 T10 V10 ==> P20 R30 T32 T42 V23
 470 < 2 > C10 C30 C40 D10 L20 L30 M10 P30 T10 T32 V10 ==> R30 V40
 471 < 2 > C10 C30 C40 C53 D12 D30 F10 F20 L21 L30 M10 M50 P30 R10 R23 R30 T10 V10
 ==> V23
 472 < 2 > C10 C20 C30 D30 F10 G13 G20 L21 L30 M10 M32 M50 P30 R10 R23 R30 T10
 V10 ==> M22 V23
 473 < 2 > C10 C20 C30 D30 F10 G13 G20 L21 L30 M10 M22 M50 P30 R10 R23 R30 T10
 V10 ==> M32 V23
 474 < 2 > C10 C20 C30 C53 D30 F10 F20 G13 G20 M10 M50 P30 R10 R23 R30 T22 V10 ==>
 V23
 475 < 2 > C10 C20 C30 C40 C53 D30 F10 F20 L11 L30 M10 M32 M50 P30 R10 T10 V10 ==>
 F30 F43 L21 T32 T42
 476 < 2 > C10 C20 C30 C40 C53 D10 D30 F10 F20 F30 F43 G11 G22 L21 M10 M42 M50 P13
 P30 R10 T10 T43 V21 ==> L10 L33 V42
 477 < 2 > C10 C20 C30 C40 C53 D10 D30 F10 F20 F30 F43 G11 G22 L21 M10 M32 M42
 M50 P13 P21 P30 P40 R10 R20 R33 T10 T21 T32 V10 V21 ==> M21
 478 < 2 > C10 C20 C30 C40 C53 D10 D30 F10 F20 F30 F43 G11 G22 L10 L21 L33 M10 M42
 M50 P13 P30 R10 T10 V21 ==> T43 V42
 479 < 2 > C10 C20 D30 F10 M10 M31 P30 R30 V10 V23 ==> D23 M20 M43 P13 P40 T33
 V30
 480 < 2 > C10 C20 C30 D30 F10 G13 G20 L21 L30 M10 M50 P30 R10 R23 R30 T10 V10 V23
 ==> M22 M32
 481 < 2 > C10 C30 C40 F20 F30 F43 L30 M10 P20 P30 R30 T10 T32 T42 V10 V23 ==> M41
 482 < 2 > C10 C20 C30 C53 D10 D30 F10 F20 L10 M10 M50 P13 P30 R10 V30 ==> D23 R23
 R30
 483 < 2 > C10 D10 D30 F10 F20 L30 M10 M50 P13 P30 P40 R10 T10 V10 V30 ==> L11 T21
 T32 T42
 484 < 2 > C10 L30 M10 M20 P13 P30 P40 R30 T10 T21 V10 V30 ==> C40 C50 D23 F30 F43
 G10 P20 R13 R20 V23 V40
 485 < 2 > C10 D30 F10 L30 M10 P13 P30 P40 R30 T10 T21 V10 V23 V30 ==> C20 F23 T33
 T41
 486 < 2 > C10 D30 F10 L11 L30 M10 P13 P30 P40 R30 T10 T21 V10 V30 ==> D23 P20
 487 < 2 > C10 D30 F10 L11 L30 M10 M20 P13 P30 P40 R30 T10 V10 V30 ==> C33 C40 D11
 F30 F43 G22 L22 V23
 488 < 2 > C10 D23 D30 F10 L11 L30 M10 P13 P20 P30 P40 R30 T10 V10 V30 ==> T21
 489 < 2 > C10 C41 D10 D23 D30 F10 F20 F33 F40 G13 G20 M10 M50 P13 P30 P40 R10 R23
 R30 V10 V30 ==> L23
 490 < 2 > C10 C40 D11 D30 F10 G22 L30 M10 M20 P13 P30 P40 R30 T10 V10 V23 V30 ==>
 C33 F30 F43 L11 L22
 491 < 2 > C10 C30 L30 M10 P13 P30 P40 R30 T10 T21 V10 V30 ==> L13 V23

492 < 2 > C10 C30 F20 L30 M10 M20 P13 P30 P40 R30 T10 T32 V10 V30 ==> L13
 493 < 2 > C10 C30 C40 D10 D23 F20 F30 F43 M10 P13 P20 P30 R30 T10 T20 T43 V30 ==> T31
 494 < 2 > C10 C23 C30 C40 C50 D10 D23 F20 F30 F43 G10 G23 L13 L20 L30 M10 M20 M30 P13 P20 P30 P40 R13 R20 R30 T10 V10 V23 V30 ==> D31 F13 M53
 495 < 2 > C10 C20 D30 F10 M10 M31 P13 P30 P40 R30 V10 V30 ==> D23 M20 M43 T33 V23
 496 < 2 > C10 C20 D30 F10 M10 M20 P13 P30 P40 R30 T33 V10 V23 V30 ==> D23 M31 M43
 497 < 2 > C10 C20 D30 F10 L30 M10 P13 P30 P40 R30 T10 T33 T41 V10 V23 V30 ==> F23 T21
 498 < 2 > C10 C20 C30 D21 D30 F10 L21 L30 M10 M32 M50 P13 P21 P30 P40 R10 T10 T21 V10 V30 ==> V41
 499 < 2 > C20 C30 C40 D30 F10 F20 F30 F43 G11 M50 R10 V10 V31 ==> L10 L33
 500 < 2 > C20 C30 C40 D30 F10 F20 M50 R20 V10 V31 ==> G11 L10 L33
 501 < 2 > C10 C20 C30 C40 C53 D30 F10 F20 M10 M32 M50 P30 R10 T10 V10 V31 ==> F30 F43 L21 T32
 502 < 2 > C10 C40 L30 M10 P30 R30 T10 T32 V10 V40 ==> C30 D10 L20
 503 < 2 > C10 C40 F20 L30 M10 P30 R30 T10 V10 V40 ==> C30 D10 L20
 504 < 2 > C10 C40 L30 M10 P30 R30 T10 V10 V23 V40 ==> C50 D23 F30 F43 G10 M20 P13 P20 P40 R13 R20 T21 V30
 505 < 2 > C10 L30 M10 P13 P30 P40 T10 T21 V10 V30 V41 ==> C20 C30 D21 D30 F10 L21 M32 M50 P21 R10
 506 < 2 > C10 C40 F20 F30 F43 L30 M10 P13 P30 P40 T10 T32 V10 V30 V41 ==> C53 D30 F10 G11 G22 L11 M42 M50 P21 R10
 507 < 2 > C10 C20 C30 D21 D30 F10 L21 L30 M10 M50 P13 P21 P30 P40 R10 T10 V10 V30 V41 ==> M32 T21
 508 < 2 > V42 ==> C10 C20 C30 C40 C53 D10 D30 F10 F20 F30 F43 G11 G22 L10 L21 L33 M10 M42 M50 P13 P30 R10 T10 T43 V21
 509 < 2 > C20 C30 D30 F10 F20 L10 M50 V10 V43 ==> C13 C40 C50 D11 D23 G11 G23 L20 L33 M13 M20 M30 M40 P10 P22 P31 P41 R30 T13 T20 T30 T40 V20 V31
 510 < 2 > C20 C30 C40 D30 F10 F20 G11 M50 V10 V43 ==> C13 C50 D11 D23 G23 L10 L20 L33 M13 M20 M30 M40 P10 P22 P31 P41 R30 T13 T20 T30 T40 V20 V31
 511 < 2 > C20 C30 C40 D30 F10 F20 F30 F43 M50 R10 V10 V43 ==> P10 R23 R30 V31
 512 < 2 > C13 ==> C20 C30 C40 C50 D11 D23 D30 F10 F20 G11 G23 L10 L20 L33 M13 M20 M30 M40 M50 P10 P22 P31 P41 R30 T13 T20 T30 T40 V10 V20 V31 V43
 513 < 2 > M13 ==> C13 C20 C30 C40 C50 D11 D23 D30 F10 F20 G11 G23 L10 L20 L33 M20 M30 M40 M50 P10 P22 P31 P41 R30 T13 T20 T30 T40 V10 V20 V31 V43
 514 < 2 > P22 ==> C13 C20 C30 C40 C50 D11 D23 D30 F10 F20 G11 G23 L10 L20 L33 M13 M20 M30 M40 M50 P10 P31 P41 R30 T13 T20 T30 T40 V10 V20 V31 V43

515 < 2 > P31 ==> C13 C20 C30 C40 C50 D11 D23 D30 F10 F20 G11 G23 L10 L20 L33 M13
 M20 M30 M40 M50 P10 P22 P41 R30 T13 T20 T30 T40 V10 V20 V31 V43
 516 < 2 > T13 ==> C13 C20 C30 C40 C50 D11 D23 D30 F10 F20 G11 G23 L10 L20 L33 M13
 M20 M30 M40 M50 P10 P22 P31 P41 R30 T20 T30 T40 V10 V20 V31 V43
 517 < 2 > T30 ==> C13 C20 C30 C40 C50 D11 D23 D30 F10 F20 G11 G23 L10 L20 L33 M13
 M20 M30 M40 M50 P10 P22 P31 P41 R30 T13 T20 T40 V10 V20 V31 V43
 518 < 2 > C30 C40 D11 D30 F10 F20 G11 M20 M40 M50 R30 T20 T40 V10 ==> C13 C20
 C50 D23 G23 L10 L20 L33 M13 M30 P10 P22 P31 P41 T13 T30 V20 V31 V43
 519 < 2 > C10 D30 F10 G13 G20 L11 L30 M10 M50 P30 R10 R23 R30 T10 V10 ==> F20 P20
 T32 T42 V43
 520 < 2 > C10 D30 F10 F33 F40 G20 L30 M10 M40 M50 P30 R10 R23 R30 T10 T32 V10 ==>
 C41 C52 F20 G13 P13 P40 V30 V43
 521 < 2 > C10 D30 F10 F20 L11 L30 M10 M50 P20 P30 R10 R23 R30 T10 V10 ==> G13 G20
 T32 T42 V43
 522 < 2 > C20 C30 C40 D30 F10 F20 F30 F43 M50 R10 R23 R30 V10 V31 ==> P10 V43
 523 < 2 > C20 C30 C40 D30 F10 F20 M50 P10 R10 R23 R30 V10 V31 V43 ==> F30 F43
 524 < 1 > C21 ==> C10 C30 C40 C53 D12 D21 D30 F10 F20 F33 F40 G11 G22 L12 L21 L30
 M10 M20 M33 M40 M50 P13 P21 P30 P40 R10 R23 R30 T10 T20 T33 T40 V10 V23 V30 V41
 525 < 1 > C42 ==> C10 C20 C30 C52 D13 D21 D30 F10 F23 F33 F40 G13 G20 L13 L21 L30
 M10 M22 M32 M40 M50 P13 P21 P30 P40 R10 R23 R30 T10 T21 T33 T41 V10 V23 V30 V41
 526 < 1 > D13 ==> C10 C20 C30 C42 C52 D21 D30 F10 F23 F33 F40 G13 G20 L13 L21 L30
 M10 M22 M32 M40 M50 P13 P21 P30 P40 R10 R23 R30 T10 T21 T33 T41 V10 V23 V30 V41
 527 < 1 > F22 ==> C10 C20 C30 C40 C53 D10 D21 D30 F10 F32 F40 G11 G20 L11 L20 L30
 M10 M23 M31 M40 M50 P11 P21 P30 P41 R10 R23 R30 T10 T22 T32 T40 V10 V20 V31 V40
 528 < 1 > F32 ==> C10 C20 C30 C40 C53 D10 D21 D30 F10 F22 F40 G11 G20 L11 L20 L30
 M10 M23 M31 M40 M50 P11 P21 P30 P41 R10 R23 R30 T10 T22 T32 T40 V10 V20 V31 V40
 529 < 1 > L12 ==> C10 C21 C30 C40 C53 D12 D21 D30 F10 F20 F33 F40 G11 G22 L21 L30
 M10 M20 M33 M40 M50 P13 P21 P30 P40 R10 R23 R30 T10 T20 T33 T40 V10 V23 V30 V41
 530 < 1 > L31 ==> C10 C20 C30 C41 C53 D10 D23 D30 F10 F20 F33 F40 G13 G20 L10 L23
 M10 M20 M31 M43 M50 P13 P21 P30 P40 R10 R23 R30 T11 T22 T33 T40 V10 V23 V30 V41
 531 < 1 > M51 ==> C10 C20 C33 C40 C50 D11 D23 D30 F10 F23 F30 F43 G10 G22 L11 L22
 L30 M10 M20 M31 M43 P13 P20 P30 P40 R13 R20 R30 T10 T21 T33 T41 V10 V23 V30 V40
 532 < 1 > T11 ==> C10 C20 C30 C41 C53 D10 D23 D30 F10 F20 F33 F40 G13 G20 L10 L23
 L31 M10 M20 M31 M43 M50 P13 P21 P30 P40 R10 R23 R30 T22 T33 T40 V10 V23 V30 V41
 533 < 1 > V13 ==> C10 C20 C30 C40 C53 D10 D23 D30 F10 F20 F30 F43 G11 G22 L10 L21
 L33 M10 M22 M33 M42 M50 P13 P20 P30 P41 R10 R23 R30 T10 T20 T31 T43 V21 V30 V42
 534 < 1 > C10 D10 D23 D30 F10 F20 M10 M33 M50 P13 P20 P30 R10 R23 R30 T10 V30 ==>
 C20 C30 C40 C53 F30 F43 G11 G22 L10 L21 L33 M22 M42 P41 T20 T31 T43 V13 V21 V42
 535 < 1 > C10 C20 C33 C40 C50 D11 D23 D30 F10 F23 F30 F43 G10 G22 L11 L22 L30 M10
 M20 M31 M43 P13 P20 P30 P40 R13 R20 R30 T10 T21 T33 T41 V10 V23 V30 V40 ==> M51

536 < 1 > C10 C20 C30 C40 C53 D10 D21 D30 F10 F40 G11 G20 L11 L20 L30 M10 M23 M31 M40 M50 P11 P21 P30 P41 R10 R23 R30 T10 T22 T32 T40 V10 V20 V31 V40 ==> F22 F32

537 < 1 > C10 C30 C40 C53 D12 D21 D30 F10 F20 F33 F40 G11 G22 L21 L30 M10 M20 M33 M40 M50 P13 P21 P30 P40 R10 R23 R30 T10 T20 T33 T40 V10 V23 V30 V41 ==> C21 L12

538 < 1 > C10 C20 C30 C52 D21 D30 F10 F23 F33 F40 G13 G20 L13 L21 L30 M10 M22 M32 M40 M50 P13 P21 P30 P40 R10 R23 R30 T10 T21 T33 T41 V10 V23 V30 V41 ==> C42 D13

539 < 1 > C10 C20 C30 C41 C53 D10 D23 D30 F10 F20 F33 F40 G13 G20 L10 L23 M10 M20 M31 M43 M50 P13 P21 P30 P40 R10 R23 R30 T22 T33 T40 V10 V23 V30 V41 ==> L31 T11

540 < 1 > C22 ==> C10 C31 C41 C52 D10 D23 D30 F10 F20 F33 F40 G13 G20 L11 L23 L30 M10 M23 M30 M40 M50 P13 P20 P30 P40 R10 R23 R30 T10 T21 T32 T42 V10 V20 V30 V43

541 < 1 > C31 ==> C10 C22 C41 C52 D10 D23 D30 F10 F20 F33 F40 G13 G20 L11 L23 L30 M10 M23 M30 M40 M50 P13 P20 P30 P40 R10 R23 R30 T10 T21 T32 T42 V10 V20 V30 V43

542 < 1 > D20 ==> C10 C20 C30 C40 C53 D12 D30 F10 F20 F30 F43 G13 G20 L11 L21 L30 M10 M22 M32 M41 M50 P10 P20 P30 P43 R10 R23 R30 T10 T22 T32 T42 V10 V23 V31 V43

543 < 1 > P43 ==> C10 C20 C30 C40 C53 D12 D20 D30 F10 F20 F30 F43 G13 G20 L11 L21 L30 M10 M22 M32 M41 M50 P10 P20 P30 R10 R23 R30 T10 T22 T32 T42 V10 V23 V31 V43

544 < 1 > C10 C20 C30 C40 C53 D30 F10 F30 G20 L11 L30 M10 M50 P30 R10 R23 R30 T10 T22 V10 V31 ==> D12 D20 F20 F43 G13 L21 M22 M32 M41 P10 P20 P43 T32 T42 V23 V43

545 < 1 > C10 C41 C52 D10 D23 D30 F10 F20 F33 F40 G13 G20 L11 L23 L30 M10 M23 M30 M40 M50 P13 P20 P30 P40 R10 R23 R30 T10 T21 T32 T42 V10 V20 V30 V43 ==> C22 C31

A.5 Algorithme NextClosure

Algorithm 8 NextClosure

```

1:  $A \leftarrow (0, 0, \dots 0)$ 
2: while  $A \neq (1, 1, \dots 1)$  do
3:   // tant que tous les objets n'ont pas été traités
4:    $i \leftarrow m$  //  $m$  est le nombre d'objets (i.e. la longueur du vecteur  $A$ )
5:   if  $a_i = 1$  then
6:      $i = i - 1$ 
7:     Aller ligne 5 // recherche de la première valeur nulle
8:   else
9:      $a_i = 1$ 
10:    for  $j = i + 1$  to  $m$  do
11:       $a_j = 0$ 
12:    end for
13:     $A' = h(A)$  // calcul de la fermeture de  $A$ 
14:    for  $j = 1$  to  $i - 1$  do
15:      if  $a_j < a'_j$  then
16:         $i = i - 1$ 
17:        Aller ligne 5 // Si  $a_j = 0$  et  $a'_j = 1$  signifie que ce n'est pas un fermé
18:      end if
19:    end for
20:     $A = A'$  ;
21:  end if
22: end while

```

A.6 Concepts issus du contexte histogramme pour la validation des algorithmes

extension	A_{min}	A_{max}	B_{min}	B_{max}
o1	A0-33-66	A0-33-66	B0-100	B0-100
o2	A66-33-0	A66-33-0	B50-50	B50-50
o3	A33-33-33	A33-33-33	B0-100	B0-100
o4	A33-33-33	A33-33-33	B66-33	B66-33
o5	A0-100-0	A0-100-0	B33-66	B33-66
o6	A0-100-0	A0-100-0	B50-50	B50-50
o1 o3	A0-33-33	A33-33-66	B0-100	B0-100
o1 o5	A0-33-0	A0-100-66	B0-66	B33-100
o2 o3	A33-33-0	A66-33-33	B0-50	B50-100
o2 o4	A33-33-0	A66-33-33	B50-33	B66-50
o2 o6	A0-33-0	A66-100-0	B50-50	B50-50
o3 o4	A33-33-33	A33-33-33	B0-33	B66-100
o3 o5	A0-33-0	A33-100-33	B0-66	B33-100
o4 o6	A0-33-0	A33-100-33	B50-33	B66-50
o5 o6	A0-100-0	A0-100-0	B33-50	B50-66
o1 o2 o3	A0-33-0	A66-33-66	B0-50	B50-100
o1 o3 o4	A0-33-33	A33-33-66	B0-33	B66-100
o1 o5 o6	A0-33-0	A0-100-66	B0-50	B50-100
o1 o3 o5	A0-33-0	A33-100-66	B0-66	B33-100
o2 o5 o6	A0-33-0	A66-100-0	B33-50	B50-66
o2 o3 o4	A33-33-0	A66-33-33	B0-33	B66-100
o2 o4 o6	A0-33-0	A66-100-33	B50-33	B66-50
o3 o5 o6	A0-33-0	A33-100-33	B0-50	B50-100
o4 o5 o6	A0-33-0	A33-100-33	B33-33	B66-66
o1 o2 o3 o4	A0-33-0	A66-33-66	B0-33	B66-100
o1 o3 o5 o6	A0-33-0	A33-100-66	B0-50	B50-100
o2 o3 o5 o6	A0-33-0	A66-100-33	B0-50	B50-100
o2 o4 o5 o6	A0-33-0	A66-100-33	B33-33	B66-66
o3 o4 o5 o6	A0-33-0	A33-100-33	B0-33	B66-100
o1 o2 o3 o5 o6	A0-33-0	A66-100-66	B0-50	B50-100
o1 o3 o4 o5 o6	A0-33-0	A33-100-66	B0-33	B66-100
o2 o3 o4 o5 o6	A0-33-0	A66-100-33	B0-33	B66-100
o1 o2 o3 o4 o5 o6	A0-33-0	A66-100-66	B0-33	B66-100